REINFORCEMENT LEARNING FOR QUANTAL STACKELBERG EQUILIBRIUM IN MEAN-FIELD GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online reinforcement learning in episodic Stackelberg mean-field games, where a leader interacts with an infinite population of boundedly rational, myopic followers governed by a quantal-response model. The leader knows neither the followers' reward function nor their realized rewards and must infer their response model from behavioral observations. We analyze the complete observation (CO) setting, where the leader observes the exact joint state-action distribution of the population, and then extend the analysis to finite-sample observation (FSO), where only N followers are observed per step and the true population distributions remain unobserved. We propose an algorithm combining kernelized optimistic least-squares value iteration with a confidence set based on maximum likelihood estimation for the followers' reward parameter. Under our stated assumptions, for a leader kernel of rank d_L and a follower feature dimension d_F, the algorithm achieves high-probability regret $\widetilde{O}(d_L^{3/2}H^2\sqrt{T})$ under FSO relative to the optimal CO policy, where T is the number of episodes and H is the horizon length. With fixed response and regularization parameters, this rate holds when $d_F \lesssim \sqrt{d_L}$, $T \gtrsim d_F^2$, and $N = \widetilde{\Omega}(|\mathcal{Z}|T)$, where $\mathcal{Z}$ is the followers' state-action space. We further establish a CO leader-side lower bound $\Omega(d_L H\sqrt{T})$, showing that the $\sqrt{T}$ dependence is unavoidable. For the FSO commitment-rule class in this paper, we prove a fixed-dimensional lower bound $\Omega(T/\sqrt{N})$; hence $N = \widetilde{\Omega}(T)$ is necessary for uniform expected $\widetilde{O}(\sqrt{T})$ regret within this class.

1 INTRODUCTION

A leader acting on a large, strategically responsive population must learn both how individuals respond and how their aggregate behavior changes the environment. As a stylized example motivated by ride-sharing pricing models (Bimpikis et al., 2019; Chen, 2016), consider a platform choosing regional prices while observing sampled drivers rather than the entire driver population. Drivers respond locally to prices, but their price sensitivities and realized rewards are unavailable to the platform. The platform therefore learns incentives from behavior while optimizing revenue in a changing population. Mean-field games (MFGs) describe large homogeneous populations through distributions (Caines et al., 2017; Lasry & Lions, 2007); combining this description with Stackelberg commitment captures the leader's influence on that population (Aurell et al., 2022; Bensoussan et al., 2017).

We study finite-horizon Stackelberg mean-field games with myopic quantal-response followers and unknown rewards and dynamics. Quantal response models bounded rationality through a softmax response to expected immediate rewards (McFadden, 1973; McKelvey & Palfrey, 1995). Our objective is online leader regret relative to a quantal Stackelberg equilibrium (QSE) (Chen et al., 2023a). Complete observation (CO) reveals the exact follower state-action distribution and provides a benchmark. Our main problem, finite-sample observation (FSO), reveals only N paired follower states and actions per step, not any true population distribution, including the initial one. Before each episode, the leader fixes and announces a tuple of commitment rules using completed episodes; current state batches supply arguments at which those fixed rules are executed.

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.

Challenges and contributions. We propose Algorithm 1 (OVI-FSO), combining kernel optimistic value iteration, softmax maximum-likelihood confidence sets, and empirical population estimates. The response-estimation component builds on Chen et al. (2023a); extending online quantal Stackelberg learning to an infinite population creates three linked obstacles.

- *Mean-field-coupled leader environment and scalable function approximation.* In single-follower linear games, the leader's reward function and the system transition kernel admit finite-dimensional linear representations indexed by a state-action-follower-action tuple (s, a, b). In our setting, the leader's reward function and the system transition kernel depend on the followers' joint state-action distribution $\mathcal{P} = \mu \otimes \nu$, where μ is the follower-state distribution and ν is the state-conditioned follower response policy, so a finite-dimensional linear-MDP analysis no longer suffices for the general population-dependent models considered here. We use separate leader and follower feature representations and propose a kernelized optimistic value-iteration method to handle this population dependence while accommodating nonlinear leader-side function approximation. We also establish a CO leader-side lower bound $\Omega(d_L H\sqrt{T})$ on a finite-action linear-MFG family, showing that the square-root dependence on T and the linear dependence on d_L are unavoidable.
- *Empirical-mean-field plug-in error propagating through the Bellman recursion.* In Chen et al. (2023a), the leader observes the follower's actions exactly, so Bellman regression uses observed state-action tuples and incurs no empirical plug-in error. In our FSO setting, the leader substitutes an empirical population distribution where the true mean field appears. This incurs a plug-in bias that a naive quadratic reduction mixes with the exploration term under a square root, yielding a $T^{3/4}$-type upper-bound term at the threshold $N = \widetilde{\Theta}(T)$ for fixed problem scales; see Remark C.32. We resolve this with a deterministic total-variation (TV) perturbation bound that performs the substitution before quadratic reduction, converting the bias into an additive residual that is dominated once the per-step sample size N is sufficiently large.
- *State-indexed response policy and distributional planning without policy regularity.* At a fixed leader state and local policy, the population response is a policy over followers' local states rather than a single response distribution. Analyzing each local state separately introduces a spurious $|\mathcal{X}|$ factor, and the leader's optimistic policy selector is not necessarily Lipschitz in the population distribution, preventing plug-in analyses that require a Lipschitz policy selector. We resolve both by applying the elliptical-potential lemma once via a population-weighted Fisher matrix, and by establishing Lipschitz stability of the optimistic value envelope directly, thereby avoiding policy-Lipschitz assumptions.

Results and scope. Our general reproducing kernel Hilbert space (RKHS) bound depends on both information gain and the complexity of the full optimistic value class. Its finite-rank specialization separates leader learning, follower-response estimation, and finite population observation. In the data-abundant regime stated in Section 4.1, the rate is $\widetilde{O}(d_L^{3/2} H^2 \sqrt{T})$. The two lower bounds isolate leader exploration and the information lost under FSO: the former already holds under CO, while the latter persists with known model primitives for the specified commitment-rule class. Together they establish the unavoidable $\sqrt{T}$ temporal dependence and the sharp T exponent of the class-specific sample requirement. We first introduce the model and observation protocols, then develop the CO baseline and the FSO algorithm, establish their regret guarantees, and present the information-theoretic lower bounds.

1.1 Related work

Symmetric mean-field RL addresses equilibrium and control objectives. Subramanian & Mahajan (2019) study stationary mean-field equilibria and social-welfare optima; Guo et al. (2023) analyze Q-learning and policy-gradient methods under contraction assumptions; and Elie et al. (2020) study fictitious play. Our setting adds a strategic leader who learns the population's response while the followers' reward parameters and realized rewards are unobserved. Finite population observation enters both the leader's planning and the execution of its commitment rules.

Stackelberg mean-field computation and learning provide direct neighboring approaches. Dayanıklı & Lauriere (2025) use a penalized single-level mean-field-control reformulation and a neural numerical method. Zeng et al. (2025) establish finite-time stationary-point convergence for single-loop actor-critic learning with Markovian samples under gradient alignment. He et al. (2025) extend Stackelberg

Markov games to a continuum of agents and use softmax approximation within policy iteration. Our target is high-probability episodic online regret to a quantal Stackelberg benchmark with latent follower rewards, unknown leader dynamics, and adaptive finite population samples.

For finite-agent Stackelberg RL, Zhong et al. (2023) study Stackelberg–Nash learning with rational followers and revealed rewards in their main formulations and an additional exploration-and-estimation procedure for unknown rewards; Lauffer et al. (2024) study a perfectly rational responder with unknown linear utility. Our closest methodological precursor is Chen et al. (2023a), whose single-myopic-follower framework learns latent rewards through softmax maximum likelihood estimation (MLE) and uses optimistic confidence-set planning. Our extension introduces a distributional state, population-indexed leader dynamics, and adaptive cross-sectional observation. Section B.1 details the methodological inheritance and the mean-field-specific mechanisms.

Our leader-side optimism follows coverage-free linear and kernelized RL (Jin et al., 2020; Yang et al., 2020). The analysis additionally handles latent follower responses and empirical distributional inputs: predictive likelihood geometry controls the former, while the population perturbation argument controls the latter. An extended discussion of related work is deferred to Section B.

2 Problem Formulation and Observation Protocols

The episodic game has leader state and action spaces $\mathcal{S}, \mathcal{A}$ that are finite or countably infinite, finite follower spaces $\mathcal{X}, \mathcal{B}$, and horizon length $H \in \mathbb{N}$. Each episode starts from the fixed deterministic state $(s_1, \mu_1) = (s_0, \mu_0)$. At step h, the follower population has state distribution $\mu_h \in \Delta(\mathcal{X})$ and response policy $\nu_h : \mathcal{X} \to \Delta(\mathcal{B})$. Writing $\mathcal{Z} = \mathcal{X} \times \mathcal{B}$, its joint distribution is $\mathcal{P}_h = \mu_h \otimes \nu_h$, meaning $\mathcal{P}_h(x, b) = \mu_h(x)\nu_h(b \mid x)$. The leader receives $u_h(s, a, \mathcal{P}) \in [0, 1]$, and the next system state is drawn from the Borel kernel $\mathbb{P}_h(\cdot \mid s, a, \mathcal{P})$ on $\mathcal{S} \times \Delta(\mathcal{X})$.

Commitment and quantal response. Fix a compact metrizable local-policy class $\Pi_{\mathrm{loc}} \subseteq \Delta(\mathcal{A})$, equal to $\Delta(\mathcal{A})$ for finite $\mathcal{A}$. An exact-population commitment policy is a tuple of Borel maps $\pi_h : \mathcal{S} \times \Delta(\mathcal{X}) \to \Pi_{\mathrm{loc}}$; their class Π is closed under measurable stepwise and statewise pasting. The local policy realized at (s, μ) is $p = \pi_h(\cdot \mid s, \mu)$. The follower reward is $r_h(x, b; s, a) = \langle \varphi(s, a, x, b), \theta_h^* \rangle \in [0, 1]$, with known feature map φ and bound B_Θ, whereas $\theta_h^* \in \Theta = \{\theta \in \mathbb{R}^{d_F} : \|\theta\|_2 \le B_\Theta\}$ is unknown and may depend on h. For known inverse temperature $\eta > 0$, define $\varphi^p(s, x, b) = \sum_a p(a)\varphi(s, a, x, b)$ and

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

The myopic follower responds to the committed mixture, not to the sampled leader action. The policy-level response is $\nu_h^\pi = \nu^{\pi_h(\cdot \mid s, \mu), \theta_h^*}$. The reward parameter and realized follower rewards are unobserved to the leader.

Assumption 2.1. *The followers' feature map φ satisfies $\|\varphi(s, a, x, b)\|_2 \le 1$ for all (s, a, x, b).*

Assumption 2.2 (Causal MFG transition factorization)**.** *For every step $h \in [H]$, there exist a Borel probability mass kernel $\mathcal{T}_{L,h}(\cdot \mid s, a, \mu, \nu)$ on the discrete leader-state space $\mathcal{S}$ and a single-valued Borel population evolution map $\Phi_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{X}) \times \Delta(\mathcal{B})^{\mathcal{X}} \to \Delta(\mathcal{X})$ such that, for every response policy $\nu : \mathcal{X} \to \Delta(\mathcal{B})$ and $\mathcal{P} = \mu \otimes \nu$, the system's transition kernel factors as*

$$\mathbb{P}_h(s', \mathrm{d}\mu' \mid s, a, \mathcal{P}) = \mathcal{T}_{L,h}(s' \mid s, a, \mu, \nu)\, \delta_{\Phi_h(s,a,\mu,\nu)}(\mathrm{d}\mu'), \quad s' \in \mathcal{S}. \tag{2.1}$$

Thus, conditional on (s, a, μ, ν), the leader state may remain stochastic, whereas the infinite-population mean field evolves deterministically through Φ_h.

In the representative-agent specialization, the step-h population evolution map Φ_h is induced by a microscopic follower-state transition kernel $P_h^X(\cdot \mid s, a, \mu, x, b) \in \Delta(\mathcal{X})$ through $[\Phi_h(s, a, \mu, \nu)](x') = \sum_{x \in \mathcal{X}, b \in \mathcal{B}} \mu(x)\nu(b \mid x) P_h^X(x' \mid s, a, \mu, x, b)$, which is the standard deterministic mean-field-limit dynamics in MFGs (Lasry & Lions, 2007).

CO benchmark and regret. Under CO, the leader commits to π before rollout, observes the exact population variables, and subsequently observes the joint distribution induced by the realized local

policy. Define

$$J(\pi) = \mathbb{E}^{\pi}\left[\sum_{h=1}^{H} u_h(s_h, a_h, \mu_h \otimes \nu_h^{\pi})\right], \qquad \mathrm{Reg}_{\mathrm{CO}}(T) = \sum_{t=1}^{T}[J(\pi^*) - J(\pi^t)].$$

A *quantal Stackelberg equilibrium* (QSE) is a policy pair (π^*, ν^{π^*}) satisfying $\pi^* \in \mathrm{argmax}_{\pi\in\Pi} J(\pi)$. We assume the benchmark-attainment conditions of Proposition C.3, so that a maximizing policy $\pi^* \in \Pi$ exists. Both the learner and comparator use the same local-policy class Π_{loc}.

For any bounded Borel measurable function $W : \mathcal{S} \times \Delta(\mathcal{X}) \to \mathbb{R}$, write

$$(P_h W)(s, a, \mathcal{P}) := \sum_{s'\in\mathcal{S}} \int_{\Delta(\mathcal{X})} W(s', \mu')\, \mathbb{P}_h(s', \mathrm{d}\mu' \mid s, a, \mathcal{P}).$$

For every admissible nonstationary policy $\pi \in \Pi$, define the leader's action- and state-value functions (U_h^{π}, W_h^{π}) under policy π recursively by $W_{H+1}^{\pi} \equiv 0$ and

$$U_h^{\pi} = u_h + P_h W_{h+1}^{\pi}, \qquad W_h^{\pi}(s, \mu) = \left\langle U_h^{\pi}(s, \cdot, \mu \otimes \nu_h^{\pi}), \pi_h(\cdot \mid s, \mu)\right\rangle_{\mathcal{A}}. \tag{2.2}$$

The rectangular pointwise Bellman envelope over Π is defined by $W_{H+1}^{\mathrm{env}} \equiv 0$ and

$$U_h^{\mathrm{env}} := u_h + P_h W_{h+1}^{\mathrm{env}}, \quad W_h^{\mathrm{env}}(s, \mu) := \sup_{p\in\Pi_{\mathrm{loc}}} \left\langle U_h^{\mathrm{env}}\left(s, \cdot, \mu \otimes \nu^{p,\theta_h^*}(\cdot \mid \cdot; s)\right), p\right\rangle_{\mathcal{A}}. \tag{2.3}$$

As Π is rectangular, the pointwise envelope agrees with the optimal benchmark defined through $J(\pi)$ over the same class, under the attainment and measurable-selection conditions in Proposition C.3.

FSO commitment and paired observations. Let $\mathcal{F}_{\mathrm{obs}}^{t-1}$ contain only the leader-observed records of completed episodes. An FSO learner is a model-uniform function of these records and known model-class inputs, without any true population distribution as an input. Before episode t, it selects and publicly commits to g^t in

$$\mathfrak{G}_{\mathrm{FSO}} = \prod_{h=1}^{H}\{g_h : \mathcal{S} \times \Delta(\mathcal{X}) \to \Pi_{\mathrm{loc}} :\ g_h \text{ is Borel measurable}\}.$$

The tuple remains fixed throughout the episode. Its second argument is the current empirical population, even though its mathematical domain is the full simplex.

Assumption 2.3 (One-batch paired adaptive sampling)**.** *At each step (h, t), conditional on the past trajectory and on (s_h^t, μ_h^t), the leader observes N followers whose states $\{x_h^{t,n}\}_{n=1}^N$ are drawn i.i.d. from μ_h^t. The rule tuple g^t has already been selected using only $\mathcal{F}_{\mathrm{obs}}^{t-1}$ and fixed before the episode rollout. The realized local policy is obtained by evaluating g_h^t at $(s_h^t, \widehat{\mu}_h^t)$:*

$$p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t), \qquad \widehat{\mu}_h^t = \tfrac{1}{N}\sum_{n=1}^{N} \delta_{x_h^{t,n}}. \tag{2.4}$$

Conditional on the past trajectory, the current system state (s_h^t, μ_h^t), the state batch $\{x_h^{t,n}\}_{n=1}^N$, and the resulting p_h^t, the actions $\{b_h^{t,n}\}_{n=1}^N$ of these same N followers are mutually independent, with each $b_h^{t,n} \sim \nu^{p_h^t,\theta_h^*}(\cdot \mid x_h^{t,n}; s_h^t)$. The empirical joint distribution obtained from the observed N state-action pairs $\{(x_h^{t,n}, b_h^{t,n})\}_{n=1}^N$ is

$$\widehat{\mathcal{P}}_h^t := \tfrac{1}{N}\sum_{n=1}^{N} \delta_{(x_h^{t,n}, b_h^{t,n})}. \tag{2.5}$$

After the state batch, the leader discloses p_h^t, samples $a_h^t \sim p_h^t$, and observes her reward u_h^t and the same followers' paired responses before the environment advances. Pairing preserves only the within-step state–action correspondence; follower identities need not be tracked across steps or episodes. The same cohort is unnecessary: one cohort may form $\widehat{\mu}_h^t$ and another collect paired state–action observations under the resulting fixed p_h^t; see Section C.13.3. Theorem 4.1 analyzes the one-batch paired protocol, while observing only separate state and action marginals can preclude a uniform $\widetilde{O}(\sqrt{T})$ guarantee; see Proposition F.3. For a fixed rule $g = \{g_h\}_{h=1}^H \in \mathfrak{G}_{\mathrm{FSO}}$, define, at system state (s, μ) and step h, an independent state batch $\mathbf{x} \sim \mu^{\otimes N}$, the empirical marginal $\widehat{\mu}_{\mathbf{x}} := N^{-1}\sum_{n=1}^{N} \delta_{x^n}$, and the realized local policy $p_{\mathbf{x}} := g_h(s, \widehat{\mu}_{\mathbf{x}})$. Let $\nu_{h,\mathbf{x}}^*(\cdot \mid x; s) := \nu^{p_{\mathbf{x}},\theta_h^*}(\cdot \mid x; s)$ and $\mathcal{P}_{h,\mathbf{x}}^* := \mu \otimes \nu_{h,\mathbf{x}}^*$. The value of g under the true FSO protocol is defined recursively by

$$\begin{aligned} Q_h^g(s, a, \mathcal{P}) &:= u_h(s, a, \mathcal{P}) + (P_h V_{h+1}^g)(s, a, \mathcal{P}), \\ V_h^g(s, \mu) &:= \mathbb{E}_{\mathbf{x}\sim\mu^{\otimes N}}\left[\left\langle Q_h^g(s, \cdot, \mathcal{P}_{h,\mathbf{x}}^*), p_{\mathbf{x}}\right\rangle_{\mathcal{A}}\right], \qquad V_{H+1}^g \equiv 0. \end{aligned} \tag{2.6}$$

We write

$$J_{\mathrm{FSO}}(g) := V_1^g(s_0, \mu_0). \tag{2.7}$$

The occurrence of μ_0 in this evaluator-side value functional specifies the true initial population distribution and does not reveal μ_0 to the FSO leader. At the first step, the executable rule is evaluated at the observed current leader state and state batch as $p_1 = g_1(s_1, \widehat{\mu}_1)$. This definition keeps the current state batch inside the commitment rule; in particular, it does not replace $p_{\mathbf{x}}$ by its conditional mean.

For an episode-start sequence $g^t \in \mathfrak{G}_{\mathrm{FSO}}$, the regret against the exact-population CO benchmark is

$$\mathrm{Reg}_{\mathrm{FSO}}(T) = \textstyle\sum_{t=1}^{T} [J(\pi^*) - J_{\mathrm{FSO}}(g^t)]. \tag{2.8}$$

Both regrets compare with the exact-population CO value. Remark C.41 separates learning within the FSO commitment-rule class from its benchmark gap to CO.

2.1 Function approximation structure

The general model combines the finite-dimensional follower reward representation with RKHS structure for the leader's Bellman backups. We use an RKHS $\mathbb{H}$ on $\mathcal{Z}_{lf} = \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z})$ with known kernel K, canonical feature $\phi(z) = K(z, \cdot)$, and $\sup_z K(z, z) \le 1$.

Assumption 2.4. *The leader's feature map ϕ is Lipschitz continuous in $\mathcal{P}$ with constant $L_\phi > 0$:*

$$\|\phi(s, a, \mathcal{P}_1) - \phi(s, a, \mathcal{P}_2)\|_{\mathbb{H}} \;\le\; L_\phi \, \|\mathcal{P}_1 - \mathcal{P}_2\|_1 , \quad \forall\, s, a, \mathcal{P}_1, \mathcal{P}_2. \tag{2.9}$$

Let $\mathbb{B}_h W := u_h + P_h W$ be the population Bellman backup of any value function W. The leader's regressions require these backups to have a uniform RKHS norm bound over the optimistic continuation functions and follower feasible sets used during learning. We impose this structural condition at the model level.

Assumption 2.5 (RKHS Bellman optimality completeness)**.** *For some $R_U > 0$, the following structural closure property holds: for any bounded Borel measurable $U : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to [0, H]$ and any nonempty compact feasible set $\Theta' \subseteq \Theta$ for which the envelope $W_{\max}^{U,\Theta'}$ below is Borel measurable, the Bellman backup satisfies*

$$\mathbb{B}_h W_{\max}^{U,\Theta'} \in \mathcal{U}^* := \{f \in \mathbb{H} : \|f\|_{\mathbb{H}} \le R_U H\}, \quad \textit{for all } h \in [H], \tag{2.10}$$
$$\textit{where} \quad W_{\max}^{U,\Theta'}(s, \mu) \;:=\; \sup_{p \in \Pi_{\mathrm{loc}},\, \theta \in \Theta'} \langle U(s, \cdot, \mu \otimes \nu^{p,\theta}(\cdot|\cdot; s)), p \rangle_{\mathcal{A}}.$$

The closure constant R_U is required to be independent of *the specific compact subset Θ'.*

Uniformity over Θ' covers the data-dependent follower feasible sets generated during learning. Bellman closure is stronger than realizability of an optimal action-value function; the distinction and its relation to kernel RL are discussed in Section C.3.

The following finite-dimensional subclass supplies the linear specialization of this framework. Here B_d denotes the Euclidean unit ball in $\mathbb{R}^d$.

Definition 2.6 (Linear leader-follower MFG)**.** For $d_L, d_F \in \mathbb{N}$, the MFG is (d_L, d_F)-linear if there exist feature maps $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to B_{d_L}$, $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{Z} \to B_{d_F}$, parameters $\xi_h^* \in B_{d_L} \subset \mathbb{R}^{d_L}$ and $\theta_h^* \in \Theta := \overline{\mathbb{B}(0, B_\Theta)} = \{\theta \in \mathbb{R}^{d_F} : \|\theta\|_2 \le B_\Theta\}$, and an $\mathbb{R}^{d_L}$-valued signed measure ω_h^* on the Borel space $\mathcal{S} \times \Delta(\mathcal{X})$, such that

$$u_h(s, a, \mathcal{P}) \;=\; \langle \phi(s, a, \mathcal{P}), \xi_h^* \rangle, \quad r_h(x, b; s, a) \;=\; \langle \varphi(s, a, x, b), \theta_h^* \rangle,$$
$$\mathbb{P}_h(s', \mathrm{d}\mu' \,|\, s, a, \mathcal{P}) \;=\; \big\langle \phi(s, a, \mathcal{P}),\, \omega_h^*(s', \mathrm{d}\mu') \big\rangle, \tag{2.11}$$

where the feature-measure representation in (2.11) defines a valid Borel probability measure for every $(s, a, \mathcal{P})$. The corresponding integrated transition probability is specified in Remark C.2. The finite-rank dimension d_L measures leader-side complexity, while d_F is the follower-feature dimension throughout. In the general RKHS setting, the maximal information gain $G_K(T, \lambda)$ provides a corresponding measure of leader-side complexity, with $G_K(T, \lambda) = \widetilde{O}(d_L)$ in the finite-rank specialization.

For this subclass, the Euclidean feature map induces $K(z, z') = \langle \phi(z), \phi(z') \rangle$. We identify its linear span with the canonical RKHS feature span by the isometry preserving these inner products, and use ϕ for either representation. For the linear subclass, the integrated transition-measure bound in (H.9) gives $R_U = 1 + \sqrt{d_L}$ uniformly over these feasible sets by Proposition H.2. Example H.3 gives a nonlinear, infinite-rank model with no exact finite-dimensional transition factorization, satisfying the same uniform closure.

3 Optimistic learning under CO and FSO

We first describe optimistic learning under complete observation and then adapt it to finite-sample population observations. Both algorithms estimate the leader's Bellman backups and optimize jointly over a local policy and a follower parameter consistent with the available observations. All estimation and planning use completed episodes, before the next rule tuple is announced. We assume access to an exact oracle for joint optimistic maximization over the local policy and follower parameter. Its maximizing selectors are jointly Borel measurable in the completed-episode observable history and the queried state and population distribution, as specified in Assumption C.14.

3.1 CO baseline

Under CO, observing $\mathcal{P}_h^t = \mu_h^t \otimes \nu^{p_h^t, \theta_h^*}$ reveals the executed follower response through $\nu^{p_h^t, \theta_h^*}(b \mid x; s_h^t) = \mathcal{P}_h^t(x, b)/\mu_h^t(x)$ whenever $\mu_h^t(x) > 0$. Write $\nu_h^t = \nu^{p_h^t, \theta_h^*}(\cdot \mid \cdot; s_h^t)$ for this observed response. At each step $h \in [H]$, Optimistic Value Iteration for Complete Observation (OVICO) maintains the response-consistent version space

$$\Theta_h^{t-1} := \big\{\theta \in \Theta : \nu^{p_h^\tau, \theta}(\cdot \mid \cdot; s_h^\tau) = \nu_h^\tau \text{ on } \operatorname{supp}(\mu_h^\tau) \text{ for every } \tau \in [t-1]\big\}. \tag{3.1}$$

These sets contain θ_h^* and shrink as observations accumulate.

OVICO uses the kernel ridge-regression (KRR) estimator (C.8) and bonus (C.9) with exact joint-distribution inputs, exact next-population targets, and radius $\beta = B_T^{\mathrm{CO}}$, forming $\widehat{U}_h^t$ by (C.10). It selects the optimistic commitment policy through

$$\pi_h^t(\cdot \mid s, \mu) \in \operatorname{argmax}_{p \in \Pi_{\mathrm{loc}}} \max_{\theta \in \Theta_h^{t-1}} \mathbb{E}_{a \sim p}\big[\widehat{U}_h^t(s, a, \mu \otimes \nu^{p,\theta}(\cdot \mid \cdot; s))\big]. \tag{3.2}$$

Thus CO learning combines estimation of the leader's Bellman backups with extrapolation of follower responses to unobserved queries. Section C.6 gives the full algorithm and regret guarantee, and Section 4.1 compares the finite-rank CO and FSO rates.

3.2 Follower confidence sets

Under FSO, the exact responses used in (3.1) are replaced by paired state-action samples, so we construct a likelihood-based confidence set for each step. The negative log-likelihood of θ is

$$\mathcal{L}_h^t(\theta) \;=\; -\textstyle\sum_{\tau=1}^{t-1} \sum_{n=1}^{N} \log \nu^{p_h^\tau, \theta}(b_h^{\tau,n} | x_h^{\tau,n}; s_h^\tau). \tag{3.3}$$

Following the MLE framework of Chen et al. (2023a), let $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$ be the closed ball parameter domain. For the empty history at $t = 1$, set $\widehat{\theta}_h^1 = 0 \in \Theta$. For every $t \geq 2$, define $\widehat{\theta}_h^t \in \operatorname{argmin}_{\theta \in \Theta} \mathcal{L}_h^t(\theta)$ using a Borel measurable tie-breaking rule fixed once and for all. The empirical Fisher matrix, regularized by $\lambda_\theta I_{d_F}$ with $\lambda_\theta > 0$ for the construction of the confidence ellipsoid, is

$$\widehat{F}_h^t \;:=\; \lambda_\theta I_{d_F} + \eta^2 \textstyle\sum_{\tau=1}^{t-1} \sum_{n=1}^{N} \Sigma_{x_h^{\tau,n}, s_h^\tau}^{p_h^\tau, \widehat{\theta}_h^t}, \quad \text{where} \quad \Sigma_{x,s}^{p,\theta} := \mathbb{C}_{b \sim \nu^{p,\theta}(\cdot|x;s)}\big[\varphi^p(s, x, b)\big]. \tag{3.4}$$

The confidence ellipsoid for parameter θ, with squared radius β_{F}^t from (G.13), is

$$\Theta_h^t \;=\; \big\{\theta \in \Theta :\, \|\theta - \widehat{\theta}_h^t\|_{\widehat{F}_h^t}^2 \leq \beta_{\mathrm{F}}^t\big\}. \tag{3.5}$$

The quantity β_{F}^t is a *squared* radius; the calibration gives $\beta_{\mathrm{F}}^t = \beta_{\mathrm{F}}^T$ for every $t \leq T$. For a confidence parameter $\zeta \in (0, 1/5]$, the radius specified in Lemma G.5 gives simultaneous coverage $\theta_h^* \in \Theta_h^t$ with probability at least $1 - \zeta$ and has order $\widetilde{O}(d_F)$ for fixed response and regularization parameters. The MLE centers the confidence set, and the planner jointly optimizes over the local policy and a parameter in that set.

3.3 EMPIRICAL-DESIGN OPTIMISTIC VALUE ITERATION

At the backward step h of episode t, use $\widehat{z}_h^\tau = (s_h^\tau, a_h^\tau, \widehat{\mathcal{P}}_h^\tau)$ and targets $[y_h^t]_\tau = u_h^\tau + \widehat{W}_{h+1}^t(s_{h+1}^\tau, \widehat{\mu}_{h+1}^\tau)$ for $\tau < t$, starting with $\widehat{W}_{H+1}^t = 0$. The Gram matrix and query vector have entries $[K_h^t]_{\tau,\tau'} = K(\widehat{z}_h^\tau, \widehat{z}_h^{\tau'})$ and $[k_h^t(z)]_\tau = K(\widehat{z}_h^\tau, z)$. The KRR estimate, uncertainty bonus, and optimistic action value are

$$\begin{aligned} \widehat{f}_h^t(z) &= k_h^t(z)^\top (K_h^t + \lambda I)^{-1} y_h^t, \\ \Gamma_h^t(z) &= B_T^{\mathrm{FSO}} \lambda^{-1/2} \sqrt{K(z,z) - k_h^t(z)^\top (K_h^t + \lambda I)^{-1} k_h^t(z)}, \\ \widehat{U}_h^t &= \mathrm{Trun}_{[0,H-h+1]}(\widehat{f}_h^t + \Gamma_h^t + L_U^{\mathrm{FSO}} \bar{\epsilon}_N). \end{aligned} \tag{3.6}$$

Here Trun clips pointwise to the indicated interval, and the paired joint-distribution error bound is $\bar{\epsilon}_N = \min\{\epsilon_N, 2\}$, with $\epsilon_N = \widetilde{O}(\sqrt{|\mathcal{Z}|/N})$ defined in (G.11). Let B_T^{base} be the deterministic oracle radius selected by the finite dyadic calibration in Section C.4. The actual radius B_T^{FSO} defined in (G.34) absorbs *historical Bellman-target* perturbations under the radius-absorption conditions stated with the regret guarantee below. The separate correction $L_U^{\mathrm{FSO}} \bar{\epsilon}_N$, where $L_U^{\mathrm{FSO}} = 2L_\phi B_T^{\mathrm{FSO}} \lambda^{-1/2}$, handles the *current empirical query*.

We fix a maximizing selector supplied by Assumption C.14 once and for all for deterministic tie-breaking.

Using the optimistic action value in (3.6), define the objective and maximizing selector for the FSO commitment rule g_h^t and its associated optimistic follower parameter $\bar{\theta}_h^t$ by

$$\big(g_h^t(s,\widehat{\mu}), \bar{\theta}_h^t(s,\widehat{\mu})\big) \in \operatorname*{argmax}_{p \in \Pi_{\mathrm{loc}}, \theta \in \Theta_h^t} G_h^t(s,\widehat{\mu}; p, \theta) := \mathbb{E}_{a\sim p}\big[\widehat{U}_h^t(s, a, \widehat{\mu} \otimes \nu^{p,\theta}(\cdot|\cdot; s))\big] \tag{3.7}$$

The leader's optimistic state-value estimator is then defined as the resulting optimum:

$$\widehat{W}_h^t(s,\widehat{\mu}) := \max_{p \in \Pi_{\mathrm{loc}}, \theta \in \Theta_h^t} G_h^t(s,\widehat{\mu}; p, \theta) = G_h^t\big(s, \widehat{\mu}; g_h^t(s,\widehat{\mu}), \bar{\theta}_h^t(s,\widehat{\mu})\big). \tag{3.8}$$

Thus the backward sweep defines the entire rule, including its values at hypothetical empirical populations, before execution. The measurable construction for arbitrary hypothetical batches is given in (G.4) and (G.5). The generally non-convex joint optimization is a statistical-analysis oracle; operation counts and computational limitations are separated in Remark C.15 and section C.13.5.

Algorithm 1: OVI-FSO: episode-start planning and fixed-rule execution.

Input: T, H, N; calibrated $B_T^{\mathrm{FSO}}, \beta_{\mathrm{F}}^T$; λ, λ_θ.

Initialize $\mathcal{D}^0 = \varnothing$, $\widehat{\theta}_h^1 = 0$, $\widehat{F}_h^1 = \lambda_\theta I_{d_F}$, and $\Theta_h^1 = \Theta$ for every h;
for $t = 1, \ldots, T$ **do**
  Set $\widehat{W}_{H+1}^t = 0$ and use the step-specific confidence sets from completed episodes;
  for $h = H, \ldots, 1$ **do**
    Fit empirical-design KRR and form $\widehat{U}_h^t$ by (3.6); Obtain $g_h^t, \bar{\theta}_h^t$ by (3.7) and $\widehat{W}_h^t$ by (3.8);
  Commit to and publicly announce the full tuple g^t;
  for $h = 1, \ldots, H$ **do**
    Observe s_h^t and the state batch; form $\widehat{\mu}_h^t$ and disclose $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$; Sample $a_h^t \sim p_h^t$ and observe u_h^t; observe the same followers' actions and form $\widehat{\mathcal{P}}_h^t$;
  Set $\mathcal{D}^t$ to $\mathcal{D}^{t-1}$ augmented by the completed observable episode; update each step's MLE, Fisher matrix, and confidence set for episode $t+1$.

We summarize our algorithm OVI-FSO in Algorithm 1, which is the compact presentation of the full version Algorithm 3; the rollout evaluates the announced tuple without replanning it.

4 MAIN RESULTS

The FSO guarantee concerns (2.8) and the same algorithm in its full or compact presentation. We first state the complete RKHS guarantee in Theorem 4.1, then specialize its complexity to finite rank.

Complexity and calibration. Define the maximal information gain G_K and G_K^+ by

$$G_K(T,\lambda) = \sup_{m\in\{0,\dots,T\}} \sup_{z_{1:m}\in\mathcal{Z}_{lf}^m} \tfrac{1}{2}\log\det(I_m + K_{z_{1:m}}/\lambda)$$

and $G_K^+ = \max\{G_K, 1\}$, where $K_{z_{1:m}} = [K(z_i, z_j)]_{i,j}$, repeated points are allowed, and the empty-design value is zero. The complete deterministic complexities $\mathfrak{D}_T$ for FSO and $\mathfrak{D}_{T,\mathrm{CO}}$ for CO combine the information gain, $1 + \lambda R_U^2 + \log(HT/\zeta)$, and a supremum-norm cover of the optimistic value class; their exact definitions are in (C.21) and (C.23). This cover is fixed before observing data and controls the adaptive continuation functions used as regression targets. The dyadic calibration must be finite; its construction and sufficient spectral conditions are given in Section C.4.1. The FSO radius-absorption conditions (G.36) are $2\sqrt{T}\,\bar{\epsilon}_N L_\phi \lambda^{-1/2} \le \frac{1}{2}$ and $\sqrt{T}\,\bar{\epsilon}_N R_U H L_\phi \le B_T^{\mathrm{base}}$. The first bounds the feedback coefficient of historical target error; the second bounds its radius-independent part. Together they give $B_T^{\mathrm{FSO}} \le 4B_T^{\mathrm{base}} = \widetilde{O}(H\sqrt{\mathfrak{D}_T})$.

Let $B_\phi = \sup_z \|\phi(z)\|_{\mathbb{H}} \le 1$; the softmax variance and Fisher comparison constants C_{JDE}, R_Σ, and C_{sc} are defined in (C.18). The exact trace factor $L_{T,N}$, increment envelope R_Y, and finite-sample coefficient C_{FS} are defined in Remark C.35.

Theorem 4.1 (Regret of OVI-FSO)**.** *Let $\zeta \in (0, 1/5]$. Suppose the follower-side linear quantal-response model and the parameter domain $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$ are as specified in Section 2. Assume that Assumptions 2.1 to 2.5 and C.14 hold. Assume the conditions of Proposition C.3 and use its attained benchmark $\pi^* \in \Pi$. Choose the follower regularizer and radii as required by Lemma G.5. Assume the minimal dyadic calibration (C.25) is finite and the deterministic conditions (G.36) hold. Assume $\lambda \ge (T+1)/T$ as required by the setting-indexed oracle RKHS uncertainty lemma, Lemma H.1. Set the algorithmic radius to B_T^{FSO} from (G.34). Then, with probability at least $1 - 5\zeta$,*

$$\begin{aligned}\mathrm{Reg}_{\mathrm{FSO}}(T) \lesssim\ & 4L_U^{\mathrm{FSO}} C_{\mathrm{sc}} \eta H \sqrt{\beta_{\mathrm{F}}^T d_F T L_{T,N}} + 4L_U^{\mathrm{FSO}} C_{\mathrm{sc}}^2 C_{\mathrm{JDE}} \beta_{\mathrm{F}}^T d_F H L_{T,N} \\ & + 4B_T^{\mathrm{FSO}} H \sqrt{T G_K(T,\lambda)} + 2B_T^{\mathrm{FSO}} B_\phi H \sqrt{T}\log(2eH/\zeta) \\ & + C_{\mathrm{FS}}(B_T^{\mathrm{FSO}}) HT\bar{\epsilon}_N + 4L_U^{\mathrm{FSO}} R_Y H \log(2H/\zeta)\,. \end{aligned} \tag{4.1}$$

Remark 4.2 (Constants and interpretation for Theorem 4.1). The leader radius is B_T^{FSO} from (G.34), and the squared follower confidence radius is β_{F}^T. The first two terms in (4.1) are follower-response terms, the next two are leader Bellman-estimation terms, the fifth collects residuals from current queries, the uniform bonus, batch mismatch, and the TV perturbation bound, and the final term is the Bernstein residual from the follower joint-distribution-error martingale. Further details are given in Remarks C.35 and C.36.

For the rate interpretation, treat η, B_Θ, $|\mathcal{B}|$, λ_θ, and L_ϕ as fixed constants. Under the additional conditions that either $R_\Sigma = 0$ or $N\eta^2 R_\Sigma \ge \lambda_\theta$, and that $0 < c_1 \le \lambda \le c_2 < \infty$, (4.1) implies

$$\mathrm{Reg}_{\mathrm{FSO}}(T) = \widetilde{O}\left(H^2\sqrt{\mathfrak{D}_T}\left(\sqrt{G_K^+(T,\lambda)T} + d_F\sqrt{T} + d_F^2 + T\bar{\epsilon}_N\right)\right). \tag{4.2}$$

The four terms in (4.2) account for leader estimation, first- and second-order response error, and finite population observation, respectively. Without these additional conditions, the exact factor $L_{T,N}$ and the explicit $\lambda^{-1/2}L_\phi$ dependence in (4.1) must be retained.

4.1 Finite-rank interpretation and sample requirement

For a leader kernel of rank d_L, define $D_L = 1 + d_L + d_L^2 + d_F^2$. Under bounded positive λ and $\lambda R_U^2 \lesssim D_L$, the information gain is $\widetilde{O}(d_L)$ and the complete CO and FSO complexities are $\widetilde{O}(D_L)$. The d_L^2 term covers posterior shapes, while the follower feasible-set geometry contributes d_F^2.

For the CO baseline, assume the CO hypotheses of Theorem C.10, its finite calibration (C.28), constant-order λ, and $\lambda R_U^2 \lesssim D_L$, with the remaining primitive constants fixed. Then the finite-rank CO guarantee is, with probability at least $1 - 3\zeta$,

$$\mathrm{Reg}_{\mathrm{CO}}(T) = \widetilde{O}\big(H^2\sqrt{d_L D_L T} + \sqrt{d_F}H^2\sqrt{D_L T} + d_F H^2\sqrt{D_L}\big). \tag{4.3}$$

CO isolates leader learning and response extrapolation, with a separate calibration that does not depend on N. The full conditions and constants are in Theorem C.10.

Corollary 4.3 (Finite-rank linear MFG specialization). *Under the conditions of Theorem 4.1, suppose either $R_\Sigma = 0$ or $N\eta^2 R_\Sigma \geq \lambda_\theta$. Suppose also that $0 < c_1 \leq \lambda \leq c_2 < \infty$ and $\lambda R_U^2 \lesssim D_L$, with the remaining primitive constants fixed. Then, with the same probability,*

$$\mathrm{Reg}_{\mathrm{FSO}}(T) = \widetilde{O}\big(H^2\sqrt{d_L D_L T} + d_F H^2\sqrt{D_L T} + d_F^2 H^2\sqrt{D_L} + H^2 T\sqrt{D_L|\mathcal{Z}|/N}\big). \quad (4.4)$$

Data-abundant regime. Retain all conditions of Corollary 4.3 and suppose $d_F \lesssim \sqrt{d_L}$, $T \gtrsim d_F^2$, and $N = \widetilde{\Omega}(T|\mathcal{Z}|)$. Then $D_L = \Theta(d_L^2)$, and the regret is $\widetilde{O}(d_L^{3/2} H^2\sqrt{T})$. The explicit sufficient sample-size condition (G.37) ensures radius absorption and has this linear T order when its primitive constants are fixed. The per-step sample size N therefore scales with the episode budget T in this regime. Where both CO bounds apply, the upper bound (4.3) matches the $\sqrt{T}$ dependence of Theorem C.44 up to logarithmic factors; Remark C.48 gives the parameter-wise comparison. The lower bounds below address the necessary temporal orders; the full result table and other dimension regimes are in Section C.11. For comparison with expected lower bounds, choose $\zeta = (5HT)^{-1}$: the failure contribution is at most one because regret is at most HT.

4.2 Proof Overview

The FSO analysis separates response uncertainty from the population error introduced by sampling. Lipschitz stability of the optimistic value envelope and of the objective with the empirical maximizer held fixed controls batch-dependent planning error, without requiring the maximizing policy itself to be Lipschitz. Historical Bellman-target perturbations are absorbed into the predictor radius, while current-query errors are handled by the explicit correction in (3.6). For follower-response error, a deterministic TV perturbation bound extracts population-estimation error before quadratic aggregation. A population-weighted Fisher argument then controls the remaining response error without a separate statewise multiplier. Together these steps yield the additive finite-sample contribution in (4.2); the full proof architecture is given in Section G.1.

5 Information-Theoretic Lower Bounds

The lower bounds isolate leader learning and finite population observation, both measured against the optimal CO value. Let $\mathfrak{M}_{\mathrm{j}}(d_L, H)$ be the common class of $(d_L, 1)$-linear MFGs defined in Section C.12, with ambient leader-feature dimension d_L, $|\mathcal{S}| = 4$, $|\mathcal{A}| = 2^{d_L-3}$, and $|\mathcal{X}| = |\mathcal{B}| = 2$. Let $\mathfrak{A}_{\mathrm{FSO}}$ consist of possibly randomized, model-uniform learners that use known model-class inputs and completed observable episodes to select $g^t \in \mathfrak{G}_{\mathrm{FSO}}$ before episode t.

Corollary 5.1 (Lower bound for FSO). *For integers $d_L \geq 19$, $H \geq 3$, $T \geq (d_L-3)^2$, and $N \geq 1$,*

$$\mathfrak{R}_{\mathrm{FSO}}(d_L, H, T, N) := \inf_{\mathfrak{A}\in\mathfrak{A}_{\mathrm{FSO}}} \sup_{M\in\mathfrak{M}_{\mathrm{j}}(d_L,H)} \mathbb{E}_M\big[\mathrm{Reg}^M_{\mathfrak{A},\mathrm{FSO}}(T)\big] \geq c_{\mathrm{j}}\left(d_L H\sqrt{T} + \frac{T}{\sqrt{N}}\right), \quad (5.1)$$

where $\mathrm{Reg}^M_{\mathfrak{A},\mathrm{FSO}}(T)$ denotes (2.8) evaluated in model M against its optimal CO value, and $c_{\mathrm{j}} > 0$ is an absolute constant.

The leader-learning component follows from Theorem C.44 and already holds under CO. The sampling component follows from Theorem C.46 and persists even when the reward, transition, and response primitives are revealed. Embedding the two constructions into the common model class and taking the worst case gives (5.1). The complete individual statements are in Section C.12, and the common-class embedding is proved in Section F.3. Consequently, uniform expected $\widetilde{O}(\sqrt{T})$ regret within $\mathfrak{G}_{\mathrm{FSO}}$ requires $N = \widetilde{\Omega}(T)$.

6 Conclusion

We developed optimistic RL algorithms for quantal Stackelberg mean-field games under complete population observation and finite-sample observation. Our regret guarantees separate leader learning, follower-response estimation, and population sampling error. The lower bounds establish the unavoidable $\sqrt{T}$ leader-learning dependence and the linear-in-T sample requirement for uniform expected $\widetilde{O}(\sqrt{T})$ regret within the FSO commitment-rule class. A numerical illustration in Section D examines temporal and training-batch trends under a fixed evaluation batch size.

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. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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

Daron Acemoglu and Asuman Ozdaglar. Competition and efficiency in congested markets. *Mathematics of operations research*, 32(1):1–31, 2007.

Charalambos D Aliprantis and Kim C Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, Heidelberg, 3rd edition, 2006.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

Emile Anand and Ishani Karmarkar. Learning approximate Nash equilibria in cooperative multi-agent reinforcement learning via mean-field subsampling. *arXiv preprint arXiv:2603.03759*, 2026.

Emile Anand, Richard Hoffmann, Sarah Liaw, and Adam Wierman. Graphon mean-field subsampling for cooperative heterogeneous multi-agent reinforcement learning, 2026. URL https://arxiv.org/abs/2602.16196.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2): 217–271, 2022.

Alexander Aurell, Rene Carmona, Gokce Dayanikli, and Mathieu Lauriere. Optimal incentives to mitigate epidemics: a Stackelberg mean field game approach. *SIAM Journal on Control and Optimization*, 60(2):S294–S322, 2022.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. The sample complexity of Stackelberg games. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2053–2061. PMLR, 2025.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of Stackelberg equilibria in general-sum games. *Advances in Neural Information Processing Systems*, 34:25799–25811, 2021.

Maria-Florina Balcan, Kiriaki Fragkia, and Keegan Harris. Learning in structured Stackelberg games, 2025. URL https://arxiv.org/abs/2504.09006.

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

A. Bensoussan, M. H. M. Chau, Y. Lai, and S. C. P. Yam. Linear-quadratic mean field Stackelberg games with state and control delays. *SIAM Journal on Control and Optimization*, 55(4):2748–2781, 2017.

Kostas Bimpikis, Ozan Candogan, and Daniela Saban. Spatial pricing in ride-sharing networks. *Operations Research*, 67(3):744–769, 2019.

Peter E. Caines, Minyi Huang, and Roland P. Malhamé. Mean field games. In *Handbook of Dynamic Game Theory*, pp. 1–28. Springer International Publishing, 2017.

Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*, 2020.

M Keith Chen. Dynamic pricing in a labor market: Surge pricing and flexible work on the Uber platform. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 455–455, 2016.

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*, 2023a.

Yang Chen, Libo Zhang, Jiamou Liu, and Shuyue Hu. Individual-level inverse reinforcement learning for mean field games. In *Proceedings of the 21st International Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2022)*, pp. 253–262, 2022.

Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. Adversarial inverse reinforcement learning for mean field games. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, pp. 1088–1096, 2023b.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pp. 355–366, 2008.

Gökçe Dayanıklı and Mathieu Lauriere. A machine learning method for Stackelberg mean field games. *Mathematics of Operations Research*, 50(4):3055–3093, 2025.

Joseph Diestel and J. Jerry Uhl. *Vector Measures*, volume 15 of *Mathematical Surveys*. American Mathematical Society, 1977.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Romuald Elie, Thibaut Mastrolia, and Dylan Possamaï. A tale of a principal and many, many agents. *Mathematics of Operations Research*, 44(2):440–467, 2019.

Romuald Elie, Julien Pérolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7143–7150, 2020.

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

Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. Implicit learning dynamics in Stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3133–3144. PMLR, 2020.

Eric Frankel, Kshitij Kulkarni, Dmitriy Drusvyatskiy, Sewoong Oh, and Lillian J. Ratliff. Finite-time convergence rates in stochastic Stackelberg games with smooth algorithmic agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17575–17656. PMLR, 2025.

Matthias Gerstgrasser and David C Parkes. Oracles & followers: Stackelberg equilibria in deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 11213–11236. PMLR, 2023.

Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Stochastic game theory for social science: A primer on quantal response equilibrium. In *Handbook of Experimental Game Theory*, pp. 8–47. Edward Elgar Publishing, 2020.

Denizalp Goktas, Arjun Prakash, and Amy Greenwald. Convex-concave zero-sum Markov Stackelberg games. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66818–66832, 2023. doi: 10.52202/075280-2917.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field multiagent reinforcement learning: A decentralized network approach. *Mathematics of Operations Research*, 50(1):506–536, 2025.

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.

Jun He, Andrew L. Liu, and Yihsu Chen. Learning in Stackelberg Markov games. *arXiv preprint arXiv:2509.16296*, 2025.

Yuanquan Hu, Xiaoli Wei, Junji Yan, and Hengxi Zhang. Graphon mean-field control for cooperative multi-agent reinforcement learning. *Journal of the Franklin Institute*, 360(18):14783–14805, 2023.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021. URL https://arxiv.org/abs/2012.15085v3. Extended version: arXiv:2012.15085v3, 2022.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2 (1):229–260, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Niklas Lauffer, Mahsa Ghasemi, Abolfazl Hashemi, Yagiz Savas, and Ufuk Topcu. No-regret learning in dynamic Stackelberg games. *IEEE Transactions on Automatic Control*, 69(3):1418–1431, 2024.

Sabita Maharjan, Quanyan Zhu, Yan Zhang, Stein Gjessing, and Tamer Basar. Dependable demand response management in the smart grid: A Stackelberg game approach. *IEEE Transactions on Smart Grid*, 4(1):120–132, 2013.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, 1973.

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

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for extensive form games. *Experimental economics*, 1:9–41, 1998.

Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33, 2020.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pp. 99–118, 1955.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1015–1022, 2010.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*, pp. 251–259, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Laboratories, 2003.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33:13903–13916, 2020.

Muhammad Aneeq uz Zaman, Alec Koppel, Sujay Bhatt, and Tamer Başar. Oracle-free reinforcement learning in mean-field games along a single sample path. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10178–10206. PMLR, 2023.

Sihan Zeng, Benjamin Patrick Evans, Sujay Bhatt, Leo Ardon, Sumitra Ganesh, and Alec Koppel. Learning in Stackelberg mean field games: A non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 38, 2025.

Geng Zhao, Banghua Zhu, Jiantao Jiao, and Michael I Jordan. Online learning in Stackelberg games with an omniscient follower. *arXiv preprint arXiv:2301.11518*, 2023.

Han Zhong, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Can reinforcement learning find Stackelberg-Nash equilibria in general-sum Markov games with myopically rational followers? *Journal of Machine Learning Research*, 24(35):1–52, 2023.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.

APPENDICES

A Notation **15**

B Additional related work **19**

B.1 Comparison with the framework of Chen et al. (2023a) 20

C Supplementary material for the main analysis **22**

C.1 Supplementary material for the problem formulation 22

C.2 Existence of the quantal Stackelberg equilibrium 24

C.3 Detailed RKHS formulations and two function-space choices 24

C.4 Constants and confidence-radius calibration 28

C.5 A concrete Lipschitz-feature example 31

C.6 Supplementary material for the complete-observation setting 32

C.7 Measurable planning, approximation conventions, and response regularity 33

C.8 Supplementary discussion for the FSO algorithm 36

C.9 Lipschitz, stability, and optimism lemmas for the FSO regret analysis 39

C.10 Supplementary material for the FSO LBE and JDE 40

C.11 Supplementary discussion for the FSO main theorem 40

C.12 Lower-bound proof details and open questions 43

C.13 Modeling rationale and auxiliary discussions 46

D Experiment **49**

D.1 Experimental setup 49

D.2 Results 51

D.3 Discussion 52

E Auxiliary lemmas **53**

F Value-function properties and proofs of main results **61**

F.1 Value-function Lipschitz corollary 61

F.2 Complete-observation main regret theorem 62

F.3 Information-theoretic lower bounds 62

G Proofs for the finite-sample observation setting **75**

G.1 Proof architecture for Theorem 4.1 75

G.2 Lipschitz, stability, and continuity lemmas 77

G.3 FSO filtrations and empirical concentration 78

G.4 MLE concentration and Fisher–trajectory equivalence 81

G.5 Auxiliary covariance bounds for the FSO analysis 86

G.6 Regret decomposition, optimism, and Bellman/joint-distribution error bounds . . . 87

G.7 Main FSO regret theorem 98

H Further auxiliary results and proofs **100**

H.1 Setting-indexed RKHS uncertainty and function-class containment 100

H.2 Proof of Proposition H.2 102

H.3 An infinite-rank certificate for Bellman completeness 104

H.4 Proof of Lemma H.5 107

H.5 Proof of Lemma H.6 110

H.6 Softmax comparison tools for follower-response analysis 111

H.7 Proof of Lemma H.9 113

A NOTATION

Unless explicitly stated otherwise, all logarithms are natural logarithms.

Table 1: General mathematical notation

Symbol		Meaning
$\mathbb{S}_+^d$	–	the class of d-dimensional symmetric positive semidefinite matrices
$\mathcal{F}(\mathcal{X})$	–	the class of measurable functions on space $\mathcal{X}$
$[N]$	–	$[N] := \{1, \dots, N\}$
$\lvert X \rvert$	–	the cardinality of a set X
$\Delta(\cdot)$	–	the set of probability distributions on $\cdot$
$\mathbb{C}[x]$	–	the covariance matrix of a random vector x
$\lVert\cdot\rVert_P$	–	the vector seminorm induced by $P \in \mathbb{S}_+^d$
$\lesssim$	–	inequality up to some constants and logarithmic factors
B_d	–	the unit ball in $\mathbb{R}^d$: $B_d := \{x \in \mathbb{R}^d : \lVert x \rVert_2 \leq 1\}$

Table 2: Notation specific to our Stackelberg mean-field game setting

Symbol		Meaning
$\mathcal{S}$	–	leader's state space
$\mathcal{A}$	–	leader's action space
$s_0 \in \mathcal{S}$	–	leader's initial state
$\pi = \{\pi_h\}_{h\in[H]}$	–	exact-population CO/QSE commitment policy, with $\pi_h : \mathcal{S} \times \Delta(\mathcal{X}) \to \Pi_{\text{loc}}$
Π	–	class of exact-population commitment policies with local policies in Π_{loc}
$g = \{g_h\}_{h\in[H]}$	–	FSO commitment rule, with $g_h : \mathcal{S} \times \Delta(\mathcal{X}) \to \Pi_{\text{loc}}$ evaluated at $(s_h, \widehat{\mu}_h)$
p_h	–	realized step-h local policy, a simplex point in Π_{loc}, obtained by evaluating π_h or g_h at its current input
$\mathcal{X}$	–	individual followers' state space
$\mathcal{B}$	–	individual followers' action space
$\mathcal{Z}$	–	product space of individual followers' state–action pair $\mathcal{Z} = \mathcal{X} \times \mathcal{B}$
$\mu_0 \in \Delta(\mathcal{X})$	–	fixed environment-level followers' initial state distribution
ν	–	individual followers' response policy: at the policy level, $\nu^{\pi} : \mathcal{S} \times \Delta(\mathcal{X}) \times \mathcal{X} \to \Delta(\mathcal{B})$ given the leader's commitment policy π; at local level, $\nu^{p,\theta}(\cdot\|x; s) \in \Delta(\mathcal{B})$ for each realized local policy $p \in \Delta(\mathcal{A})$ and leader state s (cf. (C.2))
$\mu_{\mathcal{P}}$	–	the followers' marginal state distribution for a given $\mathcal{P} \in \Delta(\mathcal{X} \times \mathcal{B})$: $\mu_{\mathcal{P}}(\cdot) = \sum_{b\in\mathcal{B}} \mathcal{P}(\cdot, b)$
$H \in \mathbb{N}$	–	horizon length
$\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$	–	system transition kernels, with $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to \Delta(\mathcal{S} \times \Delta(\mathcal{X}))$
$\mathcal{T}_{L,h}$	–	leader-state transition kernel in the factorization $\mathbb{P}_h = \mathcal{T}_{L,h}\, \delta_{\Phi_h}$ (Assumption 2.2)
Φ_h	–	deterministic population evolution map in (2.1)
P_h^X	–	microscopic follower-state transition kernel in the representative-agent specialization (C.1)
u_h	–	leader's reward at step h; $u_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{X} \times \mathcal{B}) \to [0, 1]$
u	–	collection of all leader's rewards: $u = \{u_h\}_{h\in[H]}$
r_h	–	followers' reward at step h; $r_h : \mathcal{X} \times \mathcal{B} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$
r	–	collection of all followers' rewards: $\{r_h\}_{h\in[H]}$

Table 3: Notation specific to our linear leader-follower MFG setting

Symbol		Meaning
ϕ	–	leader feature map from $(s, a, \mathcal{P})$ to the Euclidean feature ball: $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to B_{d_L}$
ω^*	–	vector-valued signed transition measures on the system state space at the next step: $\omega_h^*(s', \mathrm{d}\mu')$, with d_L coordinates in $\mathbb{R}$ and $h \in [H]$

Notation for follower parameters. In the CO analysis, θ_h^t denotes the response-consistent parameter selected by the optimistic planner from the version space Θ_h^{t-1} through (3.2). No MLE estimator is defined or used in the CO setting. In the FSO analysis, $\widehat{\theta}_h^t$ denotes the MLE estimator from Lemma G.5, whereas $\bar{\theta}_h^t$ denotes the optimistic follower parameter selected from the confidence ellipsoid Θ_h^t. For a fixed comparator commitment in the CO proof, $\tilde{\theta}_h^t$ denotes the corresponding optimistic follower parameter.

Leader and follower second-order operators. The leader KRR design operator $\Lambda_{L,h}^t$ acts on $\mathbb{H}$, whereas the follower empirical Fisher matrix $\widehat{F}_h^t$ acts on $\mathbb{R}^{d_F}$.

- $\Lambda^t_{L,h} := \lambda \mathrm{Id} + \sum_{\tau<t} \phi(z^\tau_h)\phi(z^\tau_h)^\top$ is the *leader-side KRR design operator*, an operator on the RKHS (or a $d_L \times d_L$ matrix in the finite-rank case). Used in Lemma H.1 and Lemma H.6.
- $\widehat{F}^t_h := \lambda_\theta I_{d_F} + \eta^2 \sum_{\tau<t} \sum_{n=1}^N \Sigma^{p^\tau_h, \widehat{\theta}^t_h}_{x^{\tau,n}_h, s^\tau_h}$ is the *follower-side empirical Fisher information matrix*, a $d_F \times d_F$ matrix acting on $\mathbb{R}^{d_F}$. Used in Lemma G.5 and Lemma G.17.

Quick reference: estimators, constants, and high-probability events

For ease of navigation in the long technical proofs, we collect below the estimators and algorithmic quantities (Table 4), the RKHS / MLE / dynamics / softmax constants (Table 5), and the high-probability events (Table 6) that recur throughout the paper. Each entry lists the symbol, its meaning, and the primary location where it is introduced or used.

Table 4: Estimators, planning quantities, and Gram/Fisher matrices used by Algorithm 3 and its analysis.

Symbol		Meaning
$\widehat{\theta}^t_h$	–	follower MLE estimator at episode t, step h (minimizer of $\mathcal{L}^t_h$ in (3.3); cf. Lemma G.5)
Θ^t_h	–	confidence ellipsoid for θ^*_h, $\{\theta \in \Theta : \left\|\theta - \widehat{\theta}^t_h\right\|^2_{\widehat{F}^t_h} \le \beta^t_{\mathrm{F}}\}$ (cf. (3.5))
$\widehat{\mu}^t_h$	–	empirical mean field at episode t, step h from the N-sample paired batch (cf. Assumption 2.3)
$\widehat{\mathcal{P}}^t_h$	–	paired empirical joint state-action distribution from the N-sample batch (cf. Assumption 2.3)
$\widehat{f}^t_h$	–	KRR estimator of the leader-side Bellman backup $\mathbb{B}_h \widehat{W}^t_{h+1}$ (cf. Step 1.1 of Algorithm 3)
$\widehat{U}^t_h$	–	optimistic action-value estimator, $\widehat{U}^t_h = \mathrm{Trun}_{[0,H-h+1]}(\widehat{f}^t_h + \Gamma^t_h + \Gamma_{\mathrm{FS}})$
$\widehat{W}^t_h$	–	optimistic value envelope, $\widehat{W}^t_h(s,\mu) = \max_{p\in\Pi_{\mathrm{loc}},\, \theta\in\Theta^t_h} G^t_h(s,\mu;p,\theta)$
$\Gamma^t_h(z)$	–	z-dependent RKHS UCB bonus, $B^{\mathrm{FSO}}_T \|\phi(z)\|_{(\widehat{\Lambda}^t_{L,h})^{-1}}$
Γ_{FS}	–	step-independent z-uniform FSO bonus, $L^{\mathrm{FSO}}_U \bar{\epsilon}_N$
g^t_h	–	step-h component of the FSO commitment-rule tuple g^t, mapping $(s,m) \in \mathcal{S} \times \Delta(\mathcal{X})$ to $p \in \Pi_{\mathrm{loc}}$, with an associated unannounced optimistic follower parameter $\bar{\theta}^t_h(s,m)$ from (3.7)
p^t_h	–	realized local policy at episode t, step h, obtained by evaluating the fixed rule as $p^t_h = g^t_h(s^t_h, \widehat{\mu}^t_h) \in \Pi_{\mathrm{loc}}$
$\widehat{\nu}^t_h$	–	algorithm's predicted follower response, induced by the optimistic follower parameter (CO: θ^t_h; FSO: $\bar{\theta}^t_h \in \Theta^t_h$), not necessarily by the MLE center $\widehat{\theta}^t_h$
$\Lambda^t_{L,h}$	–	CO KRR design operator on $\mathbb{H}$; $\lambda I_{\mathbb{H}} + \sum_{\tau<t} \phi(z^\tau_h) \otimes \phi(z^\tau_h)$, where $z^\tau_h = (s^\tau_h, a^\tau_h, \mathcal{P}^\tau_h)$
$\widehat{\Lambda}^t_{L,h}$	–	FSO empirical KRR design operator on $\mathbb{H}$; $\lambda I_{\mathbb{H}} + \sum_{\tau<t} \phi(\widehat{z}^\tau_h) \otimes \phi(\widehat{z}^\tau_h)$, where $\widehat{z}^\tau_h = (s^\tau_h, a^\tau_h, \widehat{\mathcal{P}}^\tau_h)$
$\widehat{F}^t_h$	–	follower-side empirical Fisher information matrix in $\mathbb{R}^{d_F \times d_F}$; $\lambda_\theta I_{d_F} + \eta^2 \sum_{\tau=1}^{t-1} \sum_{n=1}^N \Sigma^{p^\tau_h, \widehat{\theta}^t_h}_{x^{\tau,n}_h, s^\tau_h}$ (cf. (3.4))
V^t_h	–	auxiliary additive Fisher matrix for the JDE analysis, with each historical contribution evaluated at its contemporaneous MLE (cf. (G.15) and Lemma G.17)
$\Omega^{\theta^*_h}_{h,t}$	–	cumulative population Fisher, $\sum_{i<t,x} \mu^i_h(x) \Sigma^{p^i_h, \theta^*_h}_{x, s^i_h}$ (cf. (H.31), Lemma H.9)
$\Sigma^{p,\theta}_{x,s}$	–	response-feature covariance matrix (the corresponding per-observation Fisher matrix is $\eta^2\Sigma$); $\bar{\Sigma}^{p,\theta}_{s,\mu} = \sum_x \mu(x) \Sigma^{p,\theta}_{x,s}$ is the μ-weighted aggregate

Table 5: Constants grouped by analytical role (RKHS / MLE / softmax / dynamics / regret).

Symbol		Meaning
RKHS / leader-side constants		
$G_K(T,\lambda)$	–	maximal information gain of the leader-side kernel after T episodes (Srinivas et al., 2010; cf. (C.6)); $G_K = O(d_L \log T)$ in finite-rank
d_L	–	leader-feature dimension in the finite-rank linear MFG specialization (Definition 2.6)
L_ϕ	–	kernel-feature Lipschitz constant, $\|\phi(z)-\phi(z')\|_{\mathbb{H}} \le L_\phi \|\mathcal{P}-\mathcal{P}'\|_1$ (Assumption 2.4)
R_U	–	closure RKHS-norm bound: $\|f_h^*\|_{\mathbb{H}} \le R_U H$ for the true leader-Bellman backup (Assumption 2.5)
$R^{\mathrm{base}}_{T,\iota}(\mathsf{r})$	–	eventwise oracle-KRR norm bound $\mathsf{r}\lambda^{-1/2}$ under Lemma H.1, equation (H.4)
B_ϕ	–	kernel diagonal bound, $B_\phi = \sup_z \|\phi(z)\|_{\mathbb{H}} = \sup_z \sqrt{K(z,z)} \le 1$
B_T^{CO}	–	deterministic CO radius defined by (C.28)
B_T^{base}	–	deterministic FSO oracle radius defined by (C.25)
B_T^{cap}	–	deterministic FSO covering radius, $B_T^{\mathrm{cap}} = 4B_T^{\mathrm{base}}$
B_T^{FSO}	–	actual inflated OVI-FSO radius defined in (G.34)
$\beta^{\max}_{\mathrm{tar}}$	–	same-design target perturbation, $B_T^{\mathrm{FSO}} - B_T^{\mathrm{base}} = \sqrt{T}\,\bar{\epsilon}_N C^{\mathrm{env}}_{\mathrm{tar}}(B_T^{\mathrm{FSO}})$
L_U^{CO}	–	CO joint-distribution Lipschitz constant defined in (C.22)
L_U^{FSO}	–	FSO joint-distribution Lipschitz constant defined in (G.35)
$C^{(0)}_{\mathrm{tar}}$	–	intercept of the target-shift envelope in (G.39)
c_q	–	slope of the target-shift envelope in (G.39)
$C^{\mathrm{env}}_{\mathrm{tar}}(\mathsf{r})$	–	affine target-shift envelope $C^{(0)}_{\mathrm{tar}} + c_q\mathsf{r}$
λ	–	KRR regularization parameter (leader side); $\lambda \ge 1$, with $\lambda \ge (T+1)/T$ required by the main regret theorems
MLE / follower-side constants		
d_F	–	follower-feature dimension (Definition 2.6, Assumption 2.1)
$\|\mathcal{Z}\|$	–	joint follower state-action cardinality, $\|\mathcal{X}\| \cdot \|\mathcal{B}\|$
η	–	known inverse-temperature parameter, $\eta > 0$; any additional normalization is stated where it is used
$B_\Theta, B_A^*, B_A^\Theta, B_\varphi$	–	parameter norm bound $\theta \in \Theta = \overline{\mathbb{B}(0,B_\Theta)}$; advantage bound at the true parameter (H.23); advantage bound uniform over candidate parameters (C.17); follower-feature norm bound $\|\varphi\|_2 \le B_\varphi$
λ_θ	–	regularization parameter in the follower-side empirical Fisher matrix $\widehat{F}_h^t$ defined in (3.4)
β_{F}^t	–	squared empirical-Fisher confidence radius at episode t; $\beta_{\mathrm{F}}^t = \widetilde{O}(d_F)$ with fixed response, feature, parameter, and regularization constants (Remark C.36)
β_{F}^T	–	worst-case (final-episode) MLE radius used in main theorem statement
C_{sc}	–	two-sided softmax Fisher comparison constant $C_{\mathrm{sc}} = e^{4\eta B_\Theta}$, defined in (C.18)
Softmax / JDE / TV constants		
R_Σ	–	pointwise operator-norm bound on $\Sigma^{p,\theta}_{x,s}$, B_φ^2 (tight via Popoviciu)
C_{JDE}	–	second-order softmax-Lipschitz constant in centered L_1 form, $\eta^2 \exp(\eta B_A^\Theta)$ (Lemma G.17; cf. Chen et al., 2023a, Lemma G.1); used throughout main analysis (CO: Theorem C.10; FSO: Theorem 4.1)

Continued on next page

Table 5 (continued)

Symbol		Meaning
$\epsilon_N, \bar{\epsilon}_N$	–	combined untruncated empirical FSO scale $\epsilon_N = \widetilde{O}(\sqrt{\lvert\mathcal{Z}\rvert/N})$ and truncated finite-sample residual scale $\bar{\epsilon}_N = \min\{\epsilon_N, 2\}$ in ℓ_1 (Lemma G.3)
Dynamics constants		
$L_\Phi, L_\mathcal{T}$	–	step-uniform Lipschitz constants of the population evolution maps $\{\Phi_h\}_{h\in[H]}$ and the leader-state transition kernels $\{\mathcal{T}_{L,h}\}_{h\in[H]}$ in (μ, ν) (Assumption C.43)
$\mathcal{T}^\tau$	–	complete trajectory of episode τ (collection of observed tuples $\{(s_h^\tau, a_h^\tau, \ldots)\}_{h=1}^H$); distinct from the leader transition family $\{\mathcal{T}_{L,h}\}_{h\in[H]}$ in the preceding entry

Table 6: High-probability events used in the main FSO analysis. The main theorem (Theorem 4.1) holds on the intersection of these events with probability $\geq 1 - 5\zeta$.

Event		Meaning & source
$\mathcal{E}_{\text{emp}}$	–	paired adaptive empirical concentration: $\sup_{t,h} \lVert\widehat{\mu}_h^t - \mu_h^t\rVert_1 \leq \epsilon_{\mu,N} \leq \epsilon_N$ and $\sup_{t,h} \left\lVert\widehat{\mathcal{P}}_h^t - \mathcal{P}_h^t\right\rVert_1 \leq \epsilon_N$ (Lemma G.3)
$\mathcal{E}_{\text{or,FSO}}$	–	oracle-target RKHS uncertainty event for the predictable empirical design (Lemma G.10); together with the predictable historical event and the stated radius conditions, it yields the actual-predictor bound (G.46) via Corollary G.15
$\mathcal{E}_{\text{MLE,main}}$	–	well-specified MLE concentration event created by Lemma G.5; on this event, $\theta_h^* \in \Theta_h^t$ and (G.14)–(G.16) hold for all (t, h)
$\mathcal{E}_{\text{LBE-mart}}$	–	leader-side bonus-martingale concentration for $\{\Gamma_h^t\}$ (Step 4 of the proof of Lemma G.16)
$\mathcal{E}_{\text{JDE-mart}}$	–	JDE-side Bernstein-form martingale event (Stage 5.5 of Lemma G.17, eq. (G.71))

B Additional Related Work

Mean-field RL. Subramanian & Mahajan (2019) develop stochastic-gradient methods for local stationary mean-field equilibria and social-welfare optima, while Guo et al. (2023) analyze Q-learning and policy-gradient algorithms under contraction assumptions. Elie et al. (2020); Perrin et al. (2020) study fictitious-play schemes, and Angiuli et al. (2022) unify mean-field control and mean-field games through two-timescale stochastic approximation. Anahtarci et al. (2023) establish finite-sample fitted-Q-learning guarantees for regularized MFGs, Zaman et al. (2023) develop oracle-free learning from a single sample path, and Gu et al. (2025) study localized neural value-function training for MARL with mean-field structure. These symmetric mean-field formulations differ from the leader's online-regret problem with latent follower rewards and quantal response studied here. Anand & Karmarkar (2026) consider cooperative global/local-agent learning with mean-field subsampling and approximate best responses toward approximate Nash equilibria. For heterogeneous interactions, Hu et al. (2023) study cooperative graphon mean-field control and block approximations, while Anand et al. (2026) develop interaction-weighted graphon subsampling; these models go beyond the finite observed-type reward blocks discussed in Section C.13.5.

These symmetric, non-Stackelberg formulations typically give agents access to their own reward functions or reward feedback. Our leader instead infers the followers' unknown reward parameters from their behavior without observing their realized rewards.

Stackelberg mean-field learning. Dayanıklı & Lauriere (2025) reformulate a Stackelberg MFG bilevel problem as penalized single-level mean-field control and develop a neural numerical method.

Zeng et al. (2025) give finite-time and finite-sample convergence to a stationary point for a single-loop actor-critic algorithm with Markovian samples under a gradient-alignment condition. He et al. (2025) extend Stackelberg Markov games to a continuum of agents through mean-field approximation and use a softmax approximation within policy iteration. Our analysis concerns episodic online regret with latent reward inference and adaptive finite-sample population observation.

Stackelberg RL and related learning models. Learning in finite-agent Stackelberg games has been studied through equilibrium-search frameworks (Gerstgrasser & Parkes, 2023) and online learning methods (Zhao et al., 2023). The unknown-utility model of Lauffer et al. (2024) uses a perfectly rational best responder. Zhong et al. (2023) study online and offline RL for Stackelberg–Nash equilibria with myopically rational followers, treating known rewards in their main formulations and unknown rewards through an additional exploration-and-estimation procedure. Lauffer et al. (2024) analyze no-regret learning in dynamic Stackelberg games with unknown, linearly parameterized follower utility. Chen et al. (2023a) provide the closest precursor to our softmax-MLE analysis: they study a single boundedly rational follower with an unknown reward parameter and prove online regret guarantees. The mean-field setting introduces a distributional leader-state argument, adaptive population observations, and empirical plug-in errors in value estimation and planning; Section B.1 gives the detailed comparison. Related work studies side information about follower types and structured learnability (Balcan et al., 2025), finite-precision commitment learning (Bacchiocchi et al., 2025), and noisy reward feedback (Bai et al., 2021). Computational analyses cover implicit-gradient dynamics (Fiez et al., 2020), finite-time stochastic learning with smooth algorithmic agents (Frankel et al., 2025), and polynomial oracle complexity under convex-concave zero-sum Markov Stackelberg structure (Goktas et al., 2023). Our regret analysis uses a planning oracle for the generally non-convex joint optimization over the local policy and follower parameter.

Function approximation and inverse RL. The leader-side analysis builds on linear-MDP and kernelized RL methods (Jin et al., 2020; Yang et al., 2020). The joint-distribution Lipschitz assumption on ϕ and the RKHS value-iteration framework control complexity through information gain $G_K(T, \lambda)$ and a deterministic cover of the full optimistic value class, rather than directly through $|\mathcal{S}|$. Chen et al. (2022; 2023b) recover rewards from demonstrations in dynamic MFGs through individual-level inverse RL and adversarial IRL, respectively. Here, follower-reward estimates enter sequential leader-policy optimization, and performance is measured by online regret.

Our analysis follows the coverage-free tradition in linear and kernelized RL (Jin et al., 2020; Yang et al., 2020). In the Stackelberg mean-field setting, this principle must operate simultaneously on the leader-side RKHS value estimates and on the followers' softmax-MLE response model; the latter is handled by the Hellinger-affinity argument in Lemma G.5.

Lower bounds. Classical lower bounds include $\Omega(H\sqrt{SAK})$ for finite-horizon tabular MDPs (Azar et al., 2017), $\Omega(d\sqrt{K})$ for linear bandits (Dani et al., 2008), and $\Omega(dH\sqrt{KH})$ for linear-mixture MDPs (Zhou et al., 2021), with K denoting episodes or bandit rounds and within the respective parameter regimes. Theorem C.44 embeds a hypercube bandit testing construction (Lattimore & Szepesvári, 2020, Chapter 24) into the leader-side transition to obtain $\Omega(d_L H\sqrt{T})$. It establishes the unavoidable $\sqrt{T}$ dependence, while the complete-cover upper bound retains dimension and horizon gaps described in Remarks C.45 and C.48. Stronger lower bounds for other MDP classes require a separate verification of the present model assumptions before they can be transferred.

B.1 Comparison with the framework of Chen et al. (2023a)

Our analysis follows the performance-difference, response-estimation, Bellman-uncertainty, and elliptical-potential architecture used in prior work on quantal-response Stackelberg learning. The mean-field setting requires additional arguments for batch-adaptive commitments, empirical-design regression, true-query uncertainty, and finite-sample target perturbations.

B.1.1 Population-specific difficulties and technical contributions

The three challenges and contributions are stated in Section 1.

Compared with the single-follower QSE learning framework of Chen et al. (2023a), the mean-field setting introduces a distributional state μ, empirical population observations, and a joint distribution $\mathcal{P} = \mu \otimes \nu$. Our analysis addresses these differences through three mechanisms: a value-envelope planner that avoids policy-selector Lipschitzness, a TV perturbation bound that extracts finite-sample observation error outside the Cauchy–Schwarz step, and a μ-weighted Fisher aggregation that avoids statewise $|\mathcal{X}|$ penalties.

B.1.2 Planning-scheme relationship and mean-field-specific differences

Chen et al. (2023a, Algorithm 3) present two planning branches. Their Scheme 4 performs joint confidence-set optimization over the leader policy and the follower parameter, whereas their Scheme 5 uses the MLE point estimator with an explicit follower-side bonus. Our Scheme 1 follows the confidence-set principle of their Scheme 4.

Planning-scheme relationship. At each planner's own evaluation query, the joint optimization selects the optimistic follower parameter used to define $\widehat{W}_h^t$, so the corresponding value-mismatch term in Lemma E.2 vanishes by construction. In FSO, the conversion from the empirical planning query to the true mean field creates a separate value mismatch controlled by (G.29). The absence of a Scheme 5-type bonus in the main regret decomposition is therefore a consequence of selecting the confidence-set branch. The additional technical work is the adaptation of this branch to the mean-field joint distribution, leader-side RKHS uncertainty, paired batch-adaptive FSO observations, and population-weighted Fisher aggregation. See Remark C.13 for the corresponding algorithmic details.

Mean-field-specific difference: Batch-adaptive mean-field plug-in effects. The finite-sample protocol creates distinct historical and current-batch perturbations that have no analogue in the single-follower setting of Chen et al. (2023a). Historical empirical target perturbations are absorbed into the actual predictor radius B_T^{FSO} through (G.34). At the current query, the step-independent correction $\Gamma_{\mathrm{FS}} = L_U^{\mathrm{FSO}} \bar{\epsilon}_N$ converts the empirical joint-distribution input to the true input. The batch-adaptive performance-difference identity keeps the current random realized local policy inside the conditional expectation. Its value-mismatch term is controlled by two separate shifts: the optimal value envelope moves from μ to $\widehat{\mu}_{\mathbf{x}}$, and the fixed empirical-optimizer objective moves from $\widehat{\mu}_{\mathbf{x}}$ back to μ. This yields the exact factor of 2 in (G.29) without assuming Lipschitz continuity of the optimizer selector. Together with the true-query bonus conversion in Lemma G.16, these steps produce the finite-sample residual displayed in Theorem 4.1.

B.1.3 Mean-field aggregation: avoiding the spurious $|\mathcal{X}|$ factor

A smaller but still technically nontrivial improvement over a naive extension of Chen et al. (2023a) is the use of the μ-weighted aggregated covariance

$$\Omega_{h,t}^{\theta_h^*} := \sum_{i<t} \sum_{x \in \mathcal{X}} \mu_h^i(x)\, \Sigma_{x, s_h^i, \mu_h^i}^{\pi_h^i, \theta_h^*}$$

(see (H.31) and Lemma G.1) in place of a per-x covariance. This choice exchanges the $\sum_x \mu(x)$ with the trace operation *before* applying the elliptical-potential lemma, ensuring that the lemma is applied *once* (not $|\mathcal{X}|$ times). Without this aggregation, a spurious $|\mathcal{X}|$ factor would appear in the final regret bound; see Remark E.1 for the underlying μ-weighted ℓ_1-norm convention that enables the aggregation.

Summary. Our proof shares the high-level decomposition used in the proof of Chen et al. (2023a, Theorem 6.1) in Appendix D.2: performance difference, follower-response estimation, leader Bellman uncertainty, and elliptical-potential aggregation. The confidence-set planning architecture parallels Scheme 4 of their Algorithm 3. The additional work required here is the leader-side RKHS treatment of the distributional mean-field state, the batch-adaptive control of empirical mean-field plug-in errors, and the population-weighted Fisher aggregation that avoids a spurious $|\mathcal{X}|$ factor. These ingredients make the framework applicable to Stackelberg mean-field games under finite-sample observation.

C SUPPLEMENTARY MATERIAL FOR THE MAIN ANALYSIS

This appendix collects technical lemmas, propositions, corollaries, and detailed remarks used by the main analysis.

C.1 SUPPLEMENTARY MATERIAL FOR THE PROBLEM FORMULATION

C.1.1 POLICIES, REWARDS, TRANSITIONS, AND THE EXACT-POPULATION BENCHMARK

Consider a leader-follower MFG $\mathcal{M} = (\mathcal{S}, \mathcal{X}, \mathcal{A}, \mathcal{B}, H, s_0, \mu_0, \mathbb{P}, u, r)$ with leader's finite or countably infinite state/action spaces $\mathcal{S}, \mathcal{A}$, followers' finite state/action spaces $\mathcal{X}, \mathcal{B}$, horizon length $H \in \mathbb{N}$, and initial system state (s_0, μ_0). Let $\mathcal{Z} := \mathcal{X} \times \mathcal{B}$ and $\mathcal{P} \in \Delta(\mathcal{Z})$ be the followers' joint state-action distribution with state marginal distribution μ. For a response policy $\nu : \mathcal{X} \to \Delta(\mathcal{B})$, the notation $\mathcal{P} = \mu \otimes \nu$ means $\mathcal{P}(x, b) = \mu(x)\nu(b \mid x)$. Thus, $\nu(\cdot \mid x)$ specifies the followers' response at local state x. The system dynamics are specified by a collection of Borel transition kernels $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$, with $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to \Delta\big(\mathcal{S} \times \Delta(\mathcal{X})\big)$. At step h, the next system state (s_{h+1}, μ_{h+1}) is drawn from $\mathbb{P}_h(\cdot \mid s_h, a_h, \mathcal{P}_h)$. The leader and follower reward functions are $u = \{u_h\}_{h\in[H]}$ and $r = \{r_h\}_{h\in[H]}$, respectively, with $u_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to [0, 1]$ and $r_h : \mathcal{X} \times \mathcal{B} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$.

Fix a compact metrizable admissible local-policy class $\Pi_{\mathrm{loc}} \subseteq \Delta(\mathcal{A})$. When $\mathcal{A}$ is finite, set $\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$. For any $C \subseteq \Delta(\mathcal{A})$, let $\Pi(C)$ denote the class of finite-horizon exact-population *commitment policies* of the leader, consisting of nonstationary tuples $\pi = \{\pi_h\}_{h=1}^H$ such that every map $\pi_h : \mathcal{S} \times \Delta(\mathcal{X}) \to C$ is Borel measurable. Write $\Pi := \Pi(\Pi_{\mathrm{loc}})$ for the admissible commitment-policy class. The class Π is rectangular: it is closed under measurable stepwise and statewise pasting. That is, measurable local policy choices selected separately at different steps and states can be combined into a single policy in Π. This exact-population class is used for CO execution and for the QSE comparator.

Throughout the paper, the initial system state is fixed and deterministic; equivalently, the environment uses the degenerate initial distribution $\rho_0 := \delta_{(s_0,\mu_0)}$. This is an environment-level statement and does not mean that the FSO leader is given the numerical value of μ_0. The FSO algorithm is a model-uniform function of the leader's observable data and known model-class inputs and does not take any true population distribution, including μ_0, as an input.

Under CO, before the episode rollout begins, the leader announces and commits to the full tuple $\pi = \{\pi_h\}_{h=1}^H$; at step h, the *realized local policy* is $\pi_h(\cdot \mid s_h, \mu_h)$ and the leader samples a_h from it. Under FSO, the executable episode-level commitment is instead the full commitment-rule tuple $g = \{g_h\}_{h=1}^H$ defined formally in (C.33). It is selected from completed-episode observations and fixed before the next rollout. After the current state batch is observed at step h, the leader evaluates this already committed rule at $(s_h, \widehat{\mu}_h)$, discloses the resulting simplex point $p_h = g_h(s_h, \widehat{\mu}_h)$ as the realized local policy, and samples $a_h \sim p_h$. Thus π or g is the full episode-level commitment object, whereas p_h is its realized step-level local policy. The mean field drives the population evolution and the leader's reward.

The linear model is specified in Definition 2.6.

The deterministic mean-field factorization is stated in Assumption 2.2.

Mean-field dynamics and factorization. In the representative-agent specialization, the step-h population evolution map Φ_h is induced by a microscopic follower-state transition kernel $P_h^X(\cdot \mid s, a, \mu, x, b) \in \Delta(\mathcal{X})$ through

$$[\Phi_h(s, a, \mu, \nu)](x') = \textstyle\sum_{x\in\mathcal{X}, b\in\mathcal{B}} \mu(x)\nu(b \mid x)P_h^X(x' \mid s, a, \mu, x, b). \tag{C.1}$$

This is the standard deterministic mean-field-limit dynamics in MFGs (Caines et al., 2017; Lasry & Lions, 2007).

We distinguish the nonstationary commitment policy $\pi = \{\pi_h\}_{h=1}^H$ from the realized step-h local policy $p_h^\pi(s, \mu) := \pi_h(\cdot \mid s, \mu) \in \Delta(\mathcal{A})$ for given (s, μ). Define $\varphi^p(s, x, b) := \sum_a p(a)\varphi(s, a, x, b)$. As a result of *entropy-regularized* optimization with inverse temperature $\eta > 0$ (see (C.5) for the variational form) (Goeree et al., 2020; McKelvey & Palfrey, 1995), the followers' *quantal response*

(QR) under p is

$$\nu^{p,\theta_h^*}(b|x;s) = \frac{\exp\{\eta\langle\varphi^p(s,x,b),\theta_h^*\rangle\}}{\sum_{b'\in\mathcal{B}}\exp\{\eta\langle\varphi^p(s,x,b'),\theta_h^*\rangle\}}, \quad \nu_h^\pi(b|x;s,\mu) := \nu^{p_h^\pi(s,\mu),\theta_h^*}(b|x;s). \tag{C.2}$$

The policy-level expression $\nu_h^\pi(b|x;s,\mu)$ is convenient when discussing the commitment policy π. For fixed (h,s,x), distinct local policies can induce the same quantal response if their follower logits differ only by an action-independent additive constant.

We write $\|\cdot\|_{\mathbb{H}}$ for the norm of the Hilbert space $\mathbb{H}$ containing the leader feature maps. For the finite-dimensional representation in Definition 2.6, this is the Euclidean norm on $\mathbb{R}^{d_L}$; for the canonical RKHS representation introduced in Section 2.1, it is the RKHS norm.

The leader aims to maximize her expected cumulative reward

$$J(\pi) := \mathbb{E}^\pi\Big[\sum_{h\in[H]} u_h\big(s_h,a_h,\mu_h\otimes\nu_h^\pi(s_h,\mu_h)\big)\Big]. \tag{C.3}$$

The policy-specific and optimal Bellman recursions are given in (2.2) and (2.3).

Definition C.1 (Quantal Stackelberg equilibrium (QSE) relative to Π_{loc})**.** A quantal Stackelberg equilibrium (QSE) is a policy pair (π^*,ν^*) satisfying

$$\pi^* \in \operatorname*{argmax}_{\pi\in\Pi} J(\pi), \qquad \nu^* = \nu^{\pi^*},$$

where $\nu^\pi = \{\nu_h^\pi\}_{h\in[H]}$ is the quantal-response policy defined in (C.2). In other words, π^* maximizes the leader's expected return subject to the followers' QR model. When $\mathcal{A}$ is finite, this definition uses the unrestricted local simplex. When Π_{loc} is a compact restriction, comparison with the unrestricted local simplex incurs the approximation gap in (C.32).

For the CO protocol, the leader's online learning performance over T episodes is measured by the regret

$$\mathrm{Reg}_{\mathrm{CO}}(T) := \sum_{t=1}^{T}\left[J(\pi^*) - J(\pi^t)\right]. \tag{C.4}$$

Remark C.2 (Connection to System Transition Kernel). The system transition kernel in (2.11) is interpreted in the vector-measure sense. The differential $\omega_h^*(s',\mathrm{d}\mu')$ in (2.11) denotes evaluation of the vector-valued measure ω_h^* on the product space at the section $\{s'\}\times\mathrm{d}\mu'$; the integrated transition probability for any Borel $A\subseteq\mathcal{S}\times\Delta(\mathcal{X})$ is given by $\mathbb{P}_h(A\mid s,a,\mathcal{P}) = \sum_{s'\in\mathcal{S}}\int_{\{\mu':(s',\mu')\in A\}}\langle\phi(s,a,\mathcal{P}),\omega_h^*(s',\mathrm{d}\mu')\rangle$. This is the natural mean-field generalization of the linear-MDP notation $\mathbb{P}_h(s'\mid s,a) = \langle\phi(s,a),\mu_h^*(s')\rangle$ (Jin et al., 2020), with the codomain now a vector-valued measure on $\mathcal{S}\times\Delta(\mathcal{X})$ rather than on $\mathcal{S}$ alone.

With a continuum of followers, this transition kernel factorizes under Assumption 2.2 into a stochastic leader-state transition and a population evolution that is deterministic conditional on (s,a,μ,ν). Formally, the underlying physical dynamics can be expressed as a hybrid probability measure on $\mathcal{S}\times\Delta(\mathcal{X})$ given by:

$$\mathbb{P}_h(s',\mathrm{d}\mu'\mid s,a,\mu,\nu) = \underbrace{\mathcal{T}_{L,h}(s'\mid s,a,\mu,\nu)}_{\text{Leader's Dynamics}} \cdot \underbrace{\delta_{\Phi_h(s,a,\mu,\nu)}(\mathrm{d}\mu')}_{\text{Population's Dynamics}}.$$

Here $\mathcal{T}_{L,h}(s'\mid s,a,\mu,\nu)$ is the probability mass of the next leader state, and δ_y denotes the Dirac measure centered at $y\in\Delta(\mathcal{X})$. Hence the leader state may remain stochastic, while the infinite-population state evolves deterministically to $\Phi_h(s,a,\mu,\nu)$.

In the representative-agent specialization, the step-h population evolution map is induced by a microscopic follower-state transition kernel

$$P_h^X(\cdot\mid s,a,\mu,x,b)\in\Delta(\mathcal{X})$$

through

$$[\Phi_h(s,a,\mu,\nu)](x') = \sum_{x\in\mathcal{X}}\sum_{b\in\mathcal{B}}\mu(x)\nu(b\mid x)P_h^X(x'\mid s,a,\mu,x,b).$$

Here $P_h^X(\cdot \mid s, a, \mu, x, b)$ is the conditional distribution of the next local follower state given the current leader state-action, population distribution, and representative follower state-action. Thus, Definition 2.6 assumes that the aggregate system transition induced by the microscopic follower dynamics P_h^X, the response policy ν, and the leader-state transition $\mathcal{T}_{L,h}$ admits the stated low-rank representation. The step-uniform Lipschitz regularity of Φ_h and $\mathcal{T}_{L,h}$ in (μ, ν) is formalized in Assumption C.43. The condition is retained as a model-level regularity property and is satisfied by the lower-bound construction, but it is not invoked by the tight CO or FSO upper-bound proof.

C.2 Existence of the quantal Stackelberg equilibrium

The followers' quantal response (C.2) admits the equivalent entropy-regularized argmax characterization

$$\nu^{p,\theta_h^*}(\cdot|x; s) \in \operatorname*{argmax}_{\widetilde{\nu}\in\Delta(\mathcal{B})} \Big[\textstyle\sum_{b\in\mathcal{B}} \widetilde{\nu}(b) \sum_{a\in\mathcal{A}} p(a)\, r_h(x, b; s, a) + \eta^{-1}\mathcal{H}(\widetilde{\nu})\Big], \tag{C.5}$$

where $r_h(x, b; s, a) = \langle \varphi(s, a, x, b), \theta_h^* \rangle$ is the followers' linear reward and $\mathcal{H}(\widetilde{\nu}) := -\sum_b \widetilde{\nu}(b) \log \widetilde{\nu}(b)$ is the Shannon entropy on $\Delta(\mathcal{B})$. This variational form makes the uniqueness of the entropy-regularized follower response transparent.

The benchmark-attainment conditions require the recursively generated exact Bellman objectives to be Borel measurable in (s, μ) and continuous in p on the compact class Π_{loc}. These benchmark-attainment conditions concern the exact value recursion, while Assumption C.14 supplies maximizing selectors for the algorithms' optimistic objectives.

Proposition C.3 (Attainment of the rectangular QSE benchmark)**.** *Suppose Π_{loc} is compact and metrizable. Let $W_{H+1}^{\mathrm{env}} \equiv 0$, and define the Bellman envelopes recursively by (2.3). For every $h \in [H]$, assume that the recursively generated objective*

$$\Psi_h^{\mathrm{env}}(s, \mu, p) := \Big\langle (u_h + P_h W_{h+1}^{\mathrm{env}}) \Big(s, \cdot, \mu \otimes \nu^{p,\theta_h^*}(\cdot \mid \cdot; s)\Big), p \Big\rangle_{\mathcal{A}}$$

is Borel measurable in (s, μ) for every fixed p and continuous in p for every fixed (s, μ). Then the pointwise maximizing correspondence is nonempty and compact valued. It admits a Borel measurable selector. Backward induction yields a policy $\pi^* \in \Pi$ satisfying

$$J(\pi^*) = \sup_{\pi\in\Pi} J(\pi).$$

Consequently, (π^*, ν^{π^*}) is a QSE relative to Π_{loc}.

Proof of Proposition C.3. Compactness and continuity in p give pointwise attainment. The measurable maximum theorem gives a Borel value function and a Borel selector at each backward step. Rectangularity allows the stepwise selectors to be pasted into an admissible nonstationary policy. For any $\pi, \widetilde{\pi} \in \Pi$ and any collection of measurable sets $\{E_h\}_{h=1}^H$ with $E_h \in \mathcal{B}(\mathcal{S} \times \Delta(\mathcal{X}))$, the policy π' defined by $\pi'_h(s, \mu) = \pi_h(s, \mu)\mathbf{1}\{(s, \mu) \in E_h\} + \widetilde{\pi}_h(s, \mu)\mathbf{1}\{(s, \mu) \notin E_h\}$ also belongs to Π. The entropy-regularized follower objective is strictly concave, so the follower response is unique for every local policy. □

C.3 Detailed RKHS formulations and two function-space choices

This subsection supplements Section 2.1 with self-contained RKHS preliminaries, two function-space choices, and the complete deterministic value-class construction.

C.3.1 Canonical kernel representation and closure discussion

We develop the RKHS representation, information gain, and Bellman-closure interpretation before giving the kernel estimators and complete optimistic value class used by both algorithms.

RKHS lift and information gain. When $|\mathcal{S}|$ is large or infinite, a finite-dimensional leader feature map can be restrictive. Therefore, we work in an RKHS $\mathbb{H}$ on $\mathcal{Z}_{lf} := \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z})$ and use the canonical feature map

$$\phi(z) := K(z, \cdot) \in \mathbb{H}, \qquad \langle \phi(z), \phi(z') \rangle_{\mathbb{H}} = K(z, z'), \qquad \sup_z K(z, z) \le 1.$$

The main RKHS arguments use only this canonical representation. Additional compactness, continuity, reference-measure, and eigenfunction assumptions are imposed separately only when deriving the spectral examples and covering rates in Remark C.8 and lemma E.13.

One component of the statistical complexity of learning in $\mathbb{H}$ is the *maximal information gain* (Srinivas et al., 2010; Yang et al., 2020):

$$G_K(T,\lambda) := \sup_{m\in\{0,\dots,T\}} \sup_{(z_1,\dots,z_m)\in\mathcal{Z}_{lf}^m} \frac{1}{2}\log\det(I_m + K_{z_{1:m}}/\lambda)\,, \quad K_{z_{1:m}} = \big[K(z_i,z_j)\big]_{i,j=1}^m, \tag{C.6}$$

where repeated query points are permitted, and the value for $m=0$ is zero. For γ-finite-rank kernels (including the finite-dimensional linear kernel as the case $\gamma = d_L$), $G_K(T,\lambda) = \widetilde{O}(\gamma)$, recovering finite-dimensional linear MFGs (Definition 2.6) as a special case.

Let $\mathbb{B}_h W := u_h + P_h W$ be the population Bellman backup of any value function W. The leader-side kernel-LSVI analysis requires uniform RKHS realizability and norm control of its adaptive Bellman regression targets. Concretely, for every bounded action-value input U and every algorithmically admissible compact follower-parameter set Θ', the Bellman backup $\mathbb{B}_h W^{U,\Theta'}_{\max}$ must remain in a controlled RKHS-norm ball whenever the envelope is Borel measurable. This Bellman-compatible structure is substantive: Du et al. (2019) establish exponential sample-complexity lower bounds for certain uniformly approximate linear value-function representations at the approximation levels in their construction, whereas Weisz et al. (2021) establish exponential sample-complexity lower bounds for planning even under exact linear realizability of the optimal action-value function. Therefore, we impose the model-level closure condition in Assumption 2.5.

Uniformity over Θ' allows Assumption 2.5 to cover the data-dependent feasible sets generated by both algorithms. For infinite-dimensional kernels, this is a structural condition on the environment-kernel pair. The comparison with kernelized RL is developed below, and Proposition H.2 verifies the linear-MFG case under (H.9).

Two function-space choices when $\mathcal{S}\times\mathcal{A}$ is finite. When $\mathcal{Z}_l := \mathcal{S}\times\mathcal{A}$ is finite, one may specify an RKHS directly on $\mathcal{Z}_l \times \Delta(\mathcal{X}\times\mathcal{B})$ or assign a separate function $f_z \in \mathbb{H}_0$ to each $z\in\mathcal{Z}_l$, where $\mathbb{H}_0$ is an RKHS on $\Delta(\mathcal{X}\times\mathcal{B})$ with kernel K_0. The separate-function formulation is itself a joint-domain RKHS with the direct-sum kernel

$$K_\oplus\big((z,\mathcal{P}),(z',\mathcal{P}')\big) := \mathbf{1}_{\{z=z'\}} K_0(\mathcal{P},\mathcal{P}').$$

Its functions $U(z,\mathcal{P}) = f_z(\mathcal{P})$ satisfy

$$\|U\|^2_{\mathbb{H}_\oplus} = \sum_{z\in\mathcal{Z}_l} \|f_z\|^2_{\mathbb{H}_0}.$$

In particular, componentwise bounds $\|f_z\|_{\mathbb{H}_0} \le R$ imply $\|U\|_{\mathbb{H}_\oplus} \le \sqrt{|\mathcal{Z}_l|}\,R$. This construction imposes no sharing across distinct z, whereas other joint-domain kernels can encode such sharing. Thus a comparison of expressiveness requires specifying the kernels and norm budgets; the separate-function formulation is not unconditionally more general than a joint-domain RKHS formulation.

Embedding-based reduction (used in this paper). For RKHS analysis with general (possibly infinite) $\mathcal{S}\times\mathcal{A}$, we adopt the first formulation. For the spectral representation and covering-rate examples below, we use an embedding-based reduction: assume there is a preprocessing embedding $\psi_{\text{embed}}\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}^{d'}$ (Yang et al., 2020), so that $\mathcal{Z}_{lf} = \mathcal{S}\times\mathcal{A}\times\Delta(\mathcal{X}\times\mathcal{B}) \subset \mathbb{R}^{d_{\text{emb}}}$ for finite d_{emb}. We treat $\mathcal{Z}_{lf}$ as a compact subset of $\mathbb{R}^{d_{\text{emb}}}$ in these examples.

Explicit RKHS construction. For this spectral representation only, suppose $\mathcal{Z}_{lf}$ is a compact metric space, K is continuous, and ϖ_K is a Borel probability measure with full support. Define the integral operator $T_{K,\varpi_K} : L^2(\varpi_K) \to L^2(\varpi_K)$ by

$$(T_{K,\varpi_K} f)(z) := \int K(z,z')f(z')\,\varpi_K(\mathrm{d}z').$$

Let $\{(\sigma_i, \psi_i)\}_{i\in\mathbb{N}}$ be the nonzero Mercer eigenpairs of T_{K,ϖ_K}, with $\{\psi_i\}$ orthonormal in $L^2(\varpi_K)$. The resulting coordinate map is an isometric ℓ_2 representation of the canonical feature map. With this notation, the RKHS $\mathbb{H}$ admits the explicit construction

$$\mathbb{H} = \left\{ f = \sum_{i\in\mathbb{N}} f_i\psi_i : \sum_{i\in\mathbb{N}} \frac{f_i^2}{\sigma_i} < \infty \right\}, \qquad \langle f, g\rangle_{\mathbb{H}} = \sum_{i\in\mathbb{N}} \frac{f_i g_i}{\sigma_i},$$

with the convention that zero-eigenvalue coordinates are omitted. Equivalently, for an orthonormal basis $\{e_i\}_{i\in\mathbb{N}}$ of ℓ_2, the canonical feature map is

$$\phi(z) = \sum_{i\in\mathbb{N}} \sqrt{\sigma_i}\,\psi_i(z)e_i, \qquad K(z, z') = \langle\phi(z), \phi(z')\rangle_{\ell_2}, \qquad \text{for } z \in \mathcal{Z}_{lf}.$$

For the $\mathbb{H}$-valued signed transition measure ω on $\mathcal{S} \times \Delta(\mathcal{X})$, define the integrated variation norm induced by bounded measurable test functions as

$$\|\omega\|_{\mathbb{H},\mathrm{TV}} := \sup_{\|f\|_\infty\le 1} \left\| \sum_{s'\in\mathcal{S}} \int f(s', \mu')\,\omega(s', \mathrm{d}\mu') \right\|_{\mathbb{H}}. \tag{C.7}$$

Why the integrated TV-norm bound rather than per-s' bounds. The norm in (C.7) measures the integrated action of an $\mathbb{H}$-valued signed transition measure against bounded test functions on $\mathcal{S} \times \Delta(\mathcal{X})$. In the finite-rank linear-MFG specialization, the sufficient condition (H.9) imposes $\|\omega_h^*\|_{\mathbb{H},\mathrm{TV}} \le \sqrt{d_L}$, matching the standard linear-MDP scaling (Jin et al., 2020). This finite-rank bound is not part of the general RKHS closure assumption. The integrated condition controls the whole operator directly, whereas separate unit bounds on the integrated contributions of individual next leader states yield the bound $R_U = 1 + |\mathcal{S}|$ by the triangle inequality. These are different sufficient conditions, with no general implication in either direction.

Relation to single-agent literature and motivation for the Θ'-uniformity. Assumption 2.5 is the natural Stackelberg-MFG analogue of the kernelized-PEVI closure condition of Jin et al. (2021, Assumption 4.8): in the single-agent case, the inner supremum is over $a \in \mathcal{A}$ (a fixed action set) and the closure target is $V = \max_a Q$; here, the inner supremum is over $(p, \theta) \in \Pi_{\mathrm{loc}} \times \Theta'$ for a possibly data-dependent compact feasible set Θ', reflecting the bilevel maximization structure of the leader's planning step (cf. the joint argmax (3.2) in Algorithm 2 and (3.7) in Algorithm 3). The uniformity-in-Θ' requirement is precisely what allows the closure property to apply to the algorithmic iterate $\widehat{W}_{h+1}^t = W_{\max}^{\widehat{U}_{h+1}^t, \Theta'_{h+1}}$, where Θ'_{h+1} is the data-dependent follower-parameter feasible set: $\Theta'_{h+1} = \Theta_{h+1}^{t-1}$ (the CO step-wise version space (3.1)) or $\Theta'_{h+1} = \Theta_{h+1}^t$ (the FSO MLE confidence ellipsoid (3.5)). In both cases, Θ'_{h+1} is a nonempty closed subset of the compact set Θ, hence compact: in CO, Θ_{h+1}^{t-1} is the intersection of closed level sets of continuous softmax constraints with $\theta_{h+1}^* \in \Theta_{h+1}^{t-1}$ trivially; in FSO, $\Theta_{h+1}^t = \{\theta \in \Theta : \|\theta - \widehat{\theta}_{h+1}^t\|^2_{\widehat{F}_{h+1}^t} \le \beta_{\mathrm{F}}^t\}$ is closed with $\widehat{\theta}_{h+1}^t \in \Theta_{h+1}^t$ trivially. In linear MFGs (Definition 2.6), the independence of R_U from Θ' is automatic: $\|\mathbb{B}_h W\|_{\mathbb{H}}$ depends on W only through $\|W\|_\infty \le H$ (see Proposition H.2). For genuinely infinite-dimensional kernels, Assumption 2.5 should be read as a structural realizability condition on the environment-kernel pair rather than an automatic consequence of smoothness of the kernel alone.

In the linear MFG of Definition 2.6 satisfying (H.9), this Θ'-uniformity is not an extra assumption but a consequence of Proposition H.2: the closure constant $R_U = 1 + \sqrt{d_L}$ depends on the inner iterate only through its sup-norm ($\le H$) and measurability, both of which hold uniformly over the compact feasible sets generated by Algorithms 2 and 3. A self-contained verification is given in Proposition H.2 and its proof.

C.3.2 Kernel estimators, equivalent bonus formulas, and observable history

We present the following algorithmic framework, which will be instantiated differently in the CO setting (Section 3.1) and the FSO setting (Section 3).

Kernel optimistic LSVI. Kernel ridge regression (KRR) fits the leader's Bellman regression targets in the RKHS, using kernel evaluations to represent nonlinear functions. Least-squares value iteration (LSVI) repeats this regression backward over the horizon. For the CO formulation of KRR, let $z_h^\tau = (s_h^\tau, a_h^\tau, \mathcal{P}_h^\tau)$; the FSO algorithm (Algorithm 3) uses the empirical design and empirical next-population arguments. An upper-confidence-bound (UCB) bonus makes the resulting action-value estimates optimistic and encourages exploration (Jin et al., 2020; Sutton & Barto, 2018; Yang et al., 2020):

$$\widehat{f}_h^t \in \operatorname*{argmin}_{f\in\mathbb{H}} \sum_{\tau=1}^{t-1} \big[[y_h^t]_\tau - f(z_h^\tau)\big]^2 + \lambda \|f\|_{\mathbb{H}}^2 \implies \widehat{f}_h^t(z) = k_h^t(z)^\top (K_h^t + \lambda I)^{-1} y_h^t, \quad \text{(C.8)}$$

where the Gram matrix and response quantities are

$$K_h^t = [K(z_h^\tau, z_h^{\tau'})]_{\tau,\tau'=1}^{t-1}, \quad k_h^t(z) = [K(z_h^\tau, z)]_{\tau=1}^{t-1}, \quad [y_h^t]_\tau = u_h^\tau + \widehat{W}_{h+1}^t(s_{h+1}^\tau, \mu_{h+1}^\tau).$$

For $u, v \in \mathbb{H}$, write $u \otimes v$ for the rank-one operator $f \mapsto \langle v, f\rangle_{\mathbb{H}} u$ on $\mathbb{H}$. With $\Lambda_{L,h}^t := \lambda I_{\mathbb{H}} + \sum_{\tau=1}^{t-1} \phi(z_h^\tau) \otimes \phi(z_h^\tau)$, the UCB bonus function is set by

$$\Gamma_h^t(z) = \beta \, \|\phi(z)\|_{(\Lambda_{L,h}^t)^{-1}} = \beta\lambda^{-1/2}\big[K(z,z) - k_h^t(z)^\top (K_h^t + \lambda I)^{-1} k_h^t(z)\big]^{1/2}. \quad \text{(C.9)}$$

Adding Γ_h^t to the KRR estimator $\widehat{f}_h^t$ and performing a truncation into $[0, H-h+1]$ yields the optimistic action-value estimator

$$\widehat{U}_h^t = \mathrm{Trun}_{[0,H-h+1]}\big(\widehat{f}_h^t + \Gamma_h^t\big). \quad \text{(C.10)}$$

Here, $\widehat{W}_h^t(s,\mu) = \max_{p\in\Pi_{\mathrm{loc}},\, \theta\in\Theta'} \mathbb{E}_{a\sim p}[\widehat{U}_h^t(s, a, \mu \otimes \nu^{p,\theta}(\cdot|\cdot; s))]$ is the leader's optimistic value function, where the step-h follower-parameter feasible set $\Theta' \subseteq \Theta$ is setting-dependent: $\Theta' = \Theta_h^{t-1}$ (CO version space, (3.1)) and $\Theta' = \Theta_h^t$ (FSO MLE confidence ellipsoid, (3.5)). The admissible local-policy class $\Pi_{\mathrm{loc}} \subseteq \Delta(\mathcal{A})$ used as the maximization domain is formalized in Assumption C.14 ($\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$ when $\mathcal{A}$ is finite). The exact joint maximization over (p, θ) is given by (3.2) for CO and (3.7) for FSO.

Let $\mathcal{F}_{\mathrm{obs}}^{t-1}$ be the σ-field generated by the complete leader-observed records from episodes $1, \ldots, t-1$, as specified in (G.2). The FSO algorithm is required to be a model-uniform function of this observable history and the known model-class inputs; in particular, it is not parameterized by the numerical value of any true population distribution. The enlarged analysis history and associated conditional expectation are defined in (G.3). In the CO analysis, the same notation is used with the natural CO observable record, which includes the exact population variables supplied by the CO protocol.

The complete deterministic UCB function class. To control the adaptive value estimates uniformly, we introduce the realizable posterior-standard-deviation family, the four-parameter optimistic action-value class $\mathcal{U}_{\mathrm{UCB}}(h, R, B, C_{\mathrm{off}})$, and the corresponding value-envelope classes. The covers used below are fixed before the trajectory is observed.

Define the realizable posterior-standard-deviation class

$$\mathfrak{F}_{\lambda,T}^{\mathrm{post}} := \big\{ z \mapsto \|\phi(z)\|_{\Lambda(z_{1:m})^{-1}} : m \in \{0, \ldots, T\},\ z_{1:m} \in \mathcal{Z}_{lf}^m \big\}, \quad \text{(C.11)}$$

where

$$\Lambda(z_{1:m}) := \lambda I_{\mathbb{H}} + \sum_{i=1}^{m} \phi(z_i) \otimes \phi(z_i).$$

Every element of $\mathfrak{F}_{\lambda,T}^{\mathrm{post}}$ has the representation

$$z \mapsto \sqrt{\langle \phi(z), \Upsilon_{z_{1:m}} \phi(z)\rangle_{\mathbb{H}}}, \qquad \Upsilon_{z_{1:m}} = \Lambda(z_{1:m})^{-1},$$

with $0 \preceq \Upsilon_{z_{1:m}} \preceq \lambda^{-1} I_{\mathbb{H}}$.

For deterministic radii $R > 0$, $B \geq 0$, and $C_{\mathrm{off}} \geq 0$, define

$$\mathcal{U}_{\mathrm{UCB}}(h, R, B, C_{\mathrm{off}}) := \big\{ z \mapsto \mathrm{Trun}_{[0,H-h+1]}\left(U_0(z) + \beta f_{z_{1:m}}(z) + c\right) \big\}, \quad \text{(C.12)}$$

where $\|U_0\|_{\mathbb{H}} \le R$, $\beta \in [0, B]$, $f_{z_{1:m}} \in \mathfrak{F}^{\text{post}}_{\lambda,T}$, and $c \in [0, C_{\text{off}}]$.

Define the deterministic Fisher-matrix envelope

$$M_F := \lambda_\theta + TN\eta^2 R_\Sigma.$$

Define the following deterministic feasible-set families, common to all steps:

$$\mathfrak{A}_{\text{alg,CO}} := \left\{\Theta \cap A : A \subseteq \mathbb{R}^{d_F} \text{ is a closed affine subspace}, \quad \Theta \cap A \neq \varnothing\right\},$$
$$\mathfrak{A}_{\text{alg,FSO}} := \left\{\Theta \cap \{\theta : \|\theta - \bar\theta\|^2_\Lambda \le \beta\} : \begin{array}{c} \bar\theta \in \overline{\mathbb{B}(0, B_\Theta)}, \quad \lambda_\theta I_{d_F} \preceq \Lambda \preceq M_F I_{d_F}, \\ 0 \le \beta \le \beta^T_{\text{F}}, \quad \Theta \cap \{\theta : \|\theta - \bar\theta\|^2_\Lambda \le \beta\} \neq \varnothing \end{array}\right\}. \tag{C.13}$$

For $\iota \in \{\text{CO}, \text{FSO}\}$, let $\mathfrak{A}_{\text{alg},\iota}$ denote the corresponding family in (C.13).

Every realized CO version space belongs to $\mathfrak{A}_{\text{alg,CO}}$. Indeed, the observed softmax log ratios impose the affine constraints

$$\eta\left\langle \varphi^{p^\tau_h}(s^\tau_h, x, b) - \varphi^{p^\tau_h}(s^\tau_h, x, b'), \theta\right\rangle = \log\frac{\nu^\tau_h(b \mid x)}{\nu^\tau_h(b' \mid x)}$$

for every historical τ, every $x \in \text{supp}(\mu^\tau_h)$, and every $b, b' \in \mathcal{B}$. Every realized FSO confidence ellipsoid belongs to $\mathfrak{A}_{\text{alg,FSO}}$ because the constrained MLE lies in Θ, the empirical Fisher matrix lies in $[\lambda_\theta I_{d_F}, M_F I_{d_F}]$, and $\beta^t_{\text{F}} \le \beta^T_{\text{F}}$.

For $U \in \mathcal{U}_{\text{UCB}}(h, R, B, C_{\text{off}})$ and $\Theta' \in \mathfrak{A}_{\text{alg},\iota}$, define

$$V^\iota_{\Theta',U}(s, \mu) := \sup_{\substack{p \in \Pi_{\text{loc}} \\ \theta \in \Theta'}} \left\langle U\left(s, \cdot, \mu \otimes \nu^{p,\theta}(\cdot \mid \cdot; s)\right), p\right\rangle_{\mathcal{A}}.$$

Define

$$\mathcal{V}_{\text{UCB},\iota}(h, R, B, C_{\text{off}}) := \left\{V^\iota_{\Theta',U} : U \in \mathcal{U}_{\text{UCB}}(h, R, B, C_{\text{off}}),\ \Theta' \in \mathfrak{A}_{\text{alg},\iota}\right\}, \tag{C.14}$$

and let $N_{\infty,\iota}(\epsilon; h, R, B, C_{\text{off}})$ denote its supremum-norm covering number. For these value-envelope covering numbers, centers may lie in the ambient space of bounded Borel functions on $\mathcal{S} \times \Delta(\mathcal{X})$.

Let $N^{\mathfrak{A}}_{\text{Haus},\iota}(\delta)$ denote the Hausdorff covering number of $\mathfrak{A}_{\text{alg},\iota}$ in the Euclidean metric on Θ. If $B_\Theta = 0$, each nonempty clipped feasible set equals $\{0\}$, so define the corresponding Hausdorff covering contribution to be zero. If $B_\Theta > 0$, the geometric argument in Lemma E.12 gives, for every $\delta \in (0, B_\Theta]$,

$$\log N^{\mathfrak{A}}_{\text{Haus},\iota}(\delta) \le c_{\mathfrak{A}} d_F^2 \log\left(\frac{c'_{\mathfrak{A},\iota}}{\delta}\right), \tag{C.15}$$

where $c'_{\mathfrak{A},\iota}$ contains only the displayed deterministic bounds for setting ι. The constant $c'_{\mathfrak{A},\text{CO}}$ depends on the compact parameter ball and the affine-subspace geometry but does not depend on N, M_F, or β^T_{F}. The constant $c'_{\mathfrak{A},\text{FSO}}$ may depend on M_F and β^T_{F} inside a logarithm. We use the explicit specializations

$$N_{\infty,\text{CO}} := N_{\infty,\iota}\big|_{\iota=\text{CO}}, \qquad N_{\infty,\text{FSO}} := N_{\infty,\iota}\big|_{\iota=\text{FSO}}.$$

Let

$$L_g(R, B) := 2\eta B_\varphi L_\phi\left(R + B\lambda^{-1/2}\right). \tag{C.16}$$

The CO and capped FSO specializations are recorded below; the realized FSO class is specified in Corollary G.15.

C.4 Constants and confidence-radius calibration

This subsection collects the constants and deterministic radius conventions used in the main CO and FSO analyses.

For $\iota \in \{\text{CO}, \text{FSO}\}$, let $N_{\infty,\iota}(\epsilon; h, R, B, C_{\text{off}})$ denote the supremum-norm covering number of the complete deterministic value-envelope class defined in (C.14). The cover ranges over RKHS base

functions with norm at most R, posterior-standard-deviation shapes with bonus coefficient at most B, offsets at most C_{off}, and the setting-specific feasible sets in (C.13). The terminal value class is $\{0\}$, with log covering number zero. This covering contribution is distinct from the information gain $G_K(T,\lambda)$. The full class definitions and geometric bounds are given in Section C.3.

Throughout the analysis, we take the follower-feature norm bound to be $B_\varphi = 1$, as permitted by Assumption 2.1. For later softmax comparisons, define the advantage bound uniform over candidate parameters by

$$B_A^\Theta := 1 \vee \left(2B_\varphi B_\Theta + \eta^{-1}\log|\mathcal{B}|\right). \tag{C.17}$$

For later use, define the softmax variance and Fisher comparison constants by

$$C_{\mathrm{JDE}} := \eta^2 \exp(\eta B_A^\Theta), \qquad R_\Sigma := B_\varphi^2, \qquad C_{\mathrm{sc}} := e^{4\eta B_\Theta}. \tag{C.18}$$

Fix $\zeta \in (0, 1/5]$, let $B_\phi := \sup_z \|\phi(z)\|_{\mathbb{H}} = \sup_z \sqrt{K(z,z)} \le 1$, and use the operational cover scale $\varepsilon_{\mathrm{cov},T} := \min\{H/T, 1/2\}$ throughout the setting-specific uncertainty analysis.

For every candidate oracle radius $\mathsf{r} > 0$, define the setting-specific oracle base-function radii and the FSO capped base-function radius by

$$R_{T,\iota}^{\mathrm{base}}(\mathsf{r}) := \mathsf{r}\lambda^{-1/2} \quad \text{for} \quad \iota \in \{\mathrm{CO}, \mathrm{FSO}\}, \qquad R_{T,\mathrm{FSO}}^{\mathrm{cap}}(\mathsf{r}) := 4\mathsf{r}\lambda^{-1/2}. \tag{C.19}$$

We use the deterministic CO radius B_T^{CO} and FSO oracle radius B_T^{base} selected by (C.28) and (C.25), respectively. The corresponding finite calibrations are required by the main regret theorems; sufficient conditions and proofs are given in Section C.4.1. The CO calibration uses only the CO function class and has no dependence on the FSO sample size N.

$$B_T^{\mathrm{cap}} := 4B_T^{\mathrm{base}}, \qquad L_U^{\mathrm{cap}} := 2L_\phi B_T^{\mathrm{cap}}\lambda^{-1/2}, \qquad C_{\mathrm{off}}^{\mathrm{cap}} := 2L_U^{\mathrm{cap}}. \tag{C.20}$$

$$\begin{aligned}\mathfrak{D}_T := {} & 1 + \lambda R_U^2 + G_K(T,\lambda) + \log(HT/\zeta) \\ & + \max_{j\in[H+1]} \log N_{\infty,\mathrm{FSO}}\left(\varepsilon_{\mathrm{cov},T}; j, R_{T,\mathrm{FSO}}^{\mathrm{cap}}(B_T^{\mathrm{base}}), B_T^{\mathrm{cap}}, C_{\mathrm{off}}^{\mathrm{cap}}\right).\end{aligned} \tag{C.21}$$

Define the CO action-value Lipschitz constant by

$$L_U^{\mathrm{CO}} := 2L_\phi B_T^{\mathrm{CO}}\lambda^{-1/2}. \tag{C.22}$$

Define

$$\begin{aligned}\mathfrak{D}_{T,\mathrm{CO}} := {} & 1 + \lambda R_U^2 + G_K(T,\lambda) + \log(HT/\zeta) \\ & + \max_{j\in[H+1]} \log N_{\infty,\mathrm{CO}}\left(\varepsilon_{\mathrm{cov},T}; j, R_{T,\mathrm{CO}}^{\mathrm{base}}(B_T^{\mathrm{CO}}), B_T^{\mathrm{CO}}, 0\right).\end{aligned} \tag{C.23}$$

When the corresponding calibration is finite, Corollary C.5 and proposition C.4 give

$$B_T^{\mathrm{CO}} = \widetilde{O}\left(H\sqrt{\mathfrak{D}_{T,\mathrm{CO}}}\right), \qquad B_T^{\mathrm{base}} = \widetilde{O}\left(H\sqrt{\mathfrak{D}_T}\right).$$

The algorithms use the exact deterministic radii specified by the calibration rules.

C.4.1 Deterministic confidence-radius calibration

The FSO uncertainty radius is calibrated against the complete deterministic cover. We begin with the terminal FSO value-class convention

$$\log N_{\infty,\mathrm{FSO}}\left(\varepsilon_{\mathrm{cov},T}; H+1, R, B, C_{\mathrm{off}}\right) := 0$$

for every admissible (R, B, C_{off}).

For $\mathsf{r} > 0$, define

$$\begin{aligned}\mathcal{Q}_T(\mathsf{r}) := {} & 2\lambda R_U^2 + 8G_K(T,\lambda) + 8\log\left(HT/\zeta\right) + 20 \\ & + 8\max_{h\in[H]} \log N_{\infty,\mathrm{FSO}}\left(\varepsilon_{\mathrm{cov},T}; h+1, R_{T,\mathrm{FSO}}^{\mathrm{cap}}(\mathsf{r}), 4\mathsf{r}, 16L_\phi\mathsf{r}\lambda^{-1/2}\right)\end{aligned} \tag{C.24}$$

The deterministic action-value and continuation-value classes are nested in the coefficient and offset caps. Therefore $\mathsf{r} \mapsto \mathcal{Q}_T(\mathsf{r})$ is nondecreasing. Let $\mathsf{r}_j := 2^j H$ for $j \in \mathbb{N}_0$, and define

$$j_T^* := \min\left\{ j \in \mathbb{N}_0 : \mathcal{Q}_T(\mathsf{r}_j) \le \left(\frac{\mathsf{r}_j}{H}\right)^2 \right\}, \qquad B_T^{\mathrm{base}} := \mathsf{r}_{j_T^*}. \tag{C.25}$$

Consequently,

$$\mathcal{Q}_T\left(B_T^{\mathrm{base}}\right) \le \left(\frac{B_T^{\mathrm{base}}}{H}\right)^2.$$

Proposition C.4 (Deterministic dyadic radius calibration)**.** *If $\mathcal{Q}_T(\mathsf{r}) = o(\mathsf{r}^2/H^2)$ as $\mathsf{r} \to \infty$, then j_T^* in (C.25) is finite. This condition holds in the finite-rank and exponential-spectrum regimes considered below. Moreover,*

$$B_T^{\mathrm{base}} \le 2H\sqrt{1 + \mathcal{Q}_T\left(B_T^{\mathrm{base}}\right)} = \widetilde{O}\left(H\sqrt{\mathfrak{D}_T}\right). \tag{C.26}$$

Proof of Proposition C.4. The subquadratic condition implies that the defining inequality holds for every sufficiently large dyadic point, so $j_T^* < \infty$. If $j_T^* = 0$, then $B_T^{\mathrm{base}} = H$, and (C.26) follows from $\mathfrak{D}_T \ge 1$. If $j_T^* \ge 1$, minimality implies

$$\mathcal{Q}_T\left(B_T^{\mathrm{base}}/2\right) > \left(\frac{B_T^{\mathrm{base}}}{2H}\right)^2.$$

Monotonicity of the covering load gives $\mathcal{Q}_T(B_T^{\mathrm{base}}/2) \le \mathcal{Q}_T(B_T^{\mathrm{base}})$, which proves the first inequality in (C.26). The common step-index convention in (C.24) and (C.21) implies $\mathcal{Q}_T(B_T^{\mathrm{base}}) \lesssim \mathfrak{D}_T$, which proves the final order. □

The CO branch is calibrated separately, so its radius does not inherit the FSO batch-size dependence. We use the analogous terminal CO value-class convention

$$\log N_{\infty,\mathrm{CO}}\left(\varepsilon_{\mathrm{cov},T}; H+1, R, B, 0\right) := 0$$

for every admissible (R, B).

For $\mathsf{r} > 0$, define

$$\begin{aligned}\mathcal{Q}_{T,\mathrm{CO}}(\mathsf{r}) := \; & 2\lambda R_U^2 + 8G_K(T,\lambda) + 8\log\left(HT/\zeta\right) + 20 \\ & + 8\max_{h\in[H]} \log N_{\infty,\mathrm{CO}}\left(\varepsilon_{\mathrm{cov},T}; h+1, R_{T,\mathrm{CO}}^{\mathrm{base}}(\mathsf{r}), \mathsf{r}, 0\right).\end{aligned} \tag{C.27}$$

Using the same dyadic grid $\mathsf{r}_j = 2^j H$, define

$$j_{T,\mathrm{CO}}^* := \min\left\{ j \in \mathbb{N}_0 : \mathcal{Q}_{T,\mathrm{CO}}(\mathsf{r}_j) \le \left(\frac{\mathsf{r}_j}{H}\right)^2 \right\}, \qquad B_T^{\mathrm{CO}} := \mathsf{r}_{j_{T,\mathrm{CO}}^*}. \tag{C.28}$$

Corollary C.5 (Deterministic CO radius calibration)**.** *If $\mathcal{Q}_{T,\mathrm{CO}}(\mathsf{r}) = o(\mathsf{r}^2/H^2)$ as $\mathsf{r} \to \infty$, then $j_{T,\mathrm{CO}}^*$ in (C.28) is finite. Moreover,*

$$B_T^{\mathrm{CO}} \le 2H\sqrt{1 + \mathcal{Q}_{T,\mathrm{CO}}\left(B_T^{\mathrm{CO}}\right)} = \widetilde{O}\left(H\sqrt{\mathfrak{D}_{T,\mathrm{CO}}}\right). \tag{C.29}$$

Proof of Corollary C.5. The subquadratic condition implies that the defining inequality in (C.28) holds at every sufficiently large dyadic point. If $j_{T,\mathrm{CO}}^* = 0$, then $B_T^{\mathrm{CO}} = H$, and (C.29) follows from $\mathfrak{D}_{T,\mathrm{CO}} \ge 1$. If $j_{T,\mathrm{CO}}^* \ge 1$, minimality gives

$$\mathcal{Q}_{T,\mathrm{CO}}\left(B_T^{\mathrm{CO}}/2\right) > \left(\frac{B_T^{\mathrm{CO}}}{2H}\right)^2.$$

Monotonicity gives $\mathcal{Q}_{T,\mathrm{CO}}(B_T^{\mathrm{CO}}/2) \le \mathcal{Q}_{T,\mathrm{CO}}(B_T^{\mathrm{CO}})$, which proves the first inequality in (C.29). The common step-index convention in (C.27) and (C.23) then proves the final order. □

Specializations and eventwise containment. The deterministic CO class used for uniform uncertainty is $\mathcal{U}_{\mathrm{UCB}}(h, R^{\mathrm{base}}_{T,\mathrm{CO}}(B^{\mathrm{CO}}_T), B^{\mathrm{CO}}_T, 0)$. The deterministic FSO class used for uniform uncertainty is the capped class $\mathcal{U}_{\mathrm{UCB}}(h, R^{\mathrm{cap}}_{T,\mathrm{FSO}}(B^{\mathrm{base}}_T), B^{\mathrm{cap}}_T, C^{\mathrm{cap}}_{\mathrm{off}})$. The eventwise containment of the realized iterates in these classes is proved by backward induction in Lemmas G.11 and H.1 and corollary G.15; it is not asserted as a pathwise deterministic consequence of information gain.

Remark C.6 (Convention for L_ϕ, R_U, and the setting-specific base radii). Throughout the paper, L_ϕ denotes the feature Lipschitz constant of Assumption 2.4, so that $\|\phi(s,a,\mathcal{P}_1) - \phi(s,a,\mathcal{P}_2)\|_{\mathbb{H}} \le L_\phi \|\mathcal{P}_1 - \mathcal{P}_2\|_1$. The closure constant R_U in Assumption 2.5 satisfies $\|f\|_{\mathbb{H}} \le R_U H$ for every $f \in \mathcal{U}^*$ and is independent of T and λ. In the linear MFG of Definition 2.6 satisfying (H.9), Proposition H.2 gives $R_U = 1 + \sqrt{d_L}$. The oracle self-normalized event in Lemma H.1 gives the eventwise base-function radius $R^{\mathrm{base}}_{T,\iota}(\mathsf{r}) = \mathsf{r}\lambda^{-1/2}$ at oracle coefficient r. Consequently, the CO and realized FSO base radii are $R^{\mathrm{base}}_{T,\mathrm{CO}}(B^{\mathrm{CO}}_T)$ and $R^{\mathrm{base}}_{T,\mathrm{FSO}}(B^{\mathrm{FSO}}_T)$, respectively. These setting-specific radii determine the estimator-norm coordinates of the corresponding complete-UCB classes. By Corollary F.1 (ii), a CO iterate has joint-distribution Lipschitz constant L^{CO}_U, whereas an FSO iterate has joint-distribution Lipschitz constant L^{FSO}_U. The Bellman-backup closure used in the uncertainty proof remains a separate statement about $\mathbb{B}_h \widehat{W}^t_{h+1}$ and does not assert that the truncated optimistic iterate belongs to the RKHS.

Remark C.7 (Complete setting-specific value-envelope complexity). The CO and FSO uncertainty radii use the separate deterministic complexities $\mathfrak{D}_{T,\mathrm{CO}}$ and $\mathfrak{D}_T$, respectively. The corresponding covers contain the RKHS base function, scalar bonus coefficient, realizable posterior-standard-deviation shape, scalar offset when present, and the setting-specific follower feasible-set family. Under a finite-rank leader kernel, recall D_L from Section 4.1. Then

$$G^+_K(T,\lambda) = \widetilde{O}(d_L), \qquad \mathfrak{D}_{T,\mathrm{CO}} = \widetilde{O}(D_L), \qquad \mathfrak{D}_T = \widetilde{O}(D_L).$$

The shorthand $D_L \asymp d_L^2 + d_F^2$ may be used in rate displays. The safe leader-side scale is

$$\widetilde{O}\left(H^2\sqrt{T d_L D_L}\right).$$

When $d_F^2 = O(d_L^2)$, this scale is $\widetilde{O}(d_L^{3/2} H^2 \sqrt{T})$. The posterior-shape entropy is retained in the finite-rank leader scale stated above.

Remark C.8 (Spectral conditions and complete-cover sublinearity). For the polynomial-decay covering bound, assume

$$\sigma_i \le C_\sigma i^{-\gamma}, \qquad \|\psi_i\|_\infty \le C_e \sigma_i^{-\tau}, \qquad \tau \in [0, 1/2), \qquad \gamma(1-2\tau) > 1.$$

The first condition ensures that the stated RKHS-ball and posterior-shape entropy bounds are finite. The current dyadic calibration is automatically finite in the polynomial regime under the stronger sufficient condition $\gamma(1-2\tau) > 3$. Suppose

$$G^+_K(T,\lambda) = \widetilde{O}(T^{g_K}), \qquad \mathfrak{D}_T = \widetilde{O}(T^{g_D}).$$

Then the leader-side cumulative contribution is

$$\widetilde{O}\left(H^2 T^{(1+g_K+g_D)/2}\right).$$

Its average vanishes whenever $g_K + g_D < 1$. For a polynomial spectrum, both exponents must be computed from the same information-gain and complete value-envelope bounds. A condition on the information-gain exponent alone is insufficient.

C.5 A CONCRETE LIPSCHITZ-FEATURE EXAMPLE

The following kernel mean-embedding construction verifies Assumption 2.4.

Example C.9 (Kernel Mean Embedding as a Lipschitz feature)**.** *A concrete family satisfying Assumption 2.4 is the kernel mean embedding (KME): let $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a bounded, positive-definite kernel (e.g., Laplacian $K(z,z') = \exp(-\|z - z'\|_1)$ or Gaussian). Let ϕ_0 be a uniformly norm-bounded feature map and define $\phi(s,a,\mathcal{P}) := \phi_0(s,a) \otimes \int K(\cdot, z)\, \mathcal{P}(\mathrm{d}z)$. Write $\|\phi_0\|_\infty := \sup_{s,a} \|\phi_0(s,a)\|_{\mathbb{H}} < \infty$ and set $C := \sqrt{2\|K\|_\infty}$. Then for any $\mathcal{P}_1, \mathcal{P}_2$:*

$$\|\phi(s,a,\mathcal{P}_1) - \phi(s,a,\mathcal{P}_2)\|_{\mathbb{H}} = \|\phi_0(s,a)\|_{\mathbb{H}} \cdot \|K\mathcal{P}_1 - K\mathcal{P}_2\|_{\mathbb{H}} \le C\|\phi_0\|_\infty \cdot \|\mathcal{P}_1 - \mathcal{P}_2\|_1,$$

where the last step follows from the bounded kernel inequality $\|K\mathcal{P}_1 - K\mathcal{P}_2\|_{\mathbb{H}} \le \sqrt{2\|K\|_\infty} \cdot \|\mathcal{P}_1 - \mathcal{P}_2\|_1$. This verifies (2.9) with $L_\phi = C\|\phi_0\|_\infty$.

C.6 Supplementary material for the complete-observation setting

C.6.1 Response-consistent planning, full OVICO algorithm, and guarantee

The CO analysis provides the baseline regret decomposition and separates leader-side Bellman error from follower-response error before finite-sample plug-in effects are introduced.

Under CO, the leader observes the exact joint $\mathcal{P}_h^t = \mu_h^t \otimes \nu^{p_h^t, \theta_h^*}$ at each step h, and the step-h follower parameter θ_h^* is constrained by the response-consistent version space (3.1), defined separately for each step. As the true data-generating parameter θ_h^* satisfies every response-consistency constraint in (3.1), $\theta_h^* \in \Theta_h^{t-1}$ and hence each Θ_h^{t-1} is nonempty. For each fixed h, these sets are nested as t grows because additional observations only add constraints. The step-wise indexing is required because θ_h^* is step-dependent (Definition 2.6).

With $\{\Theta_h^{t-1}\}_{h\in[H]}$ and $\widehat{U}_h^t$, the optimistic local policy at step h of episode t is given by (3.2), which jointly optimizes over $p \in \Pi_{\mathrm{loc}}$ and the step-h parameter $\theta \in \Theta_h^{t-1}$. The leader's optimistic state-value estimator is then defined as the resulting optimum:

$$\widehat{W}_h^t(s,\mu) := \max_{p\in\Pi_{\mathrm{loc}}, \theta\in\Theta_h^{t-1}} \mathbb{E}_{a\sim p}[\widehat{U}_h^t(s,a,\mu\otimes\nu^{p,\theta}(\cdot|\cdot;s))]. \tag{C.30}$$

We formalize Optimistic Value Iteration for Complete Observation (OVICO) in Algorithm 2.

Algorithm 2: Optimistic Value Iteration for CO (OVICO).

Input: T; H; UCB radius β; KRR regularizer λ.

Initialize dataset $\mathcal{D}^0 \leftarrow \varnothing$;
for *episode* $t = 1, \ldots, T$ **do**
  Update version spaces $\{\Theta_h^{t-1}\}_{h\in[H]}$ via (3.1) using $\mathcal{D}^{t-1}$;
  Set $\widehat{W}_{H+1}^t \equiv 0$;
  for $h = H, H-1, \ldots, 1$ **do**
    Estimate optimistic action-value function $\widehat{U}_h^t$ via (C.10); Compute policy π_h^t via (3.2);
    Estimate optimistic value function $\widehat{W}_h^t$ via (C.30);
  Announce π^t; Collect data and update $\mathcal{D}^t \leftarrow \mathcal{D}^{t-1} \cup \{(s_h^t, a_h^t, \mathcal{P}_h^t, \pi_h^t, u_h^t)\}_{h\in[H]}$;

Output: Leader's commitment policy sequence $\{\pi^t\}_{t\in[T]}$.

We first invoke Lemma H.1, which characterizes the uncertainty on the leader's side, and then provide the theoretical guarantee (Theorem C.10) for OVICO (Algorithm 2) in the CO setting. Under the CO hypotheses of Lemma H.1 and the planning rule of Algorithm 2, with probability at least $1-\zeta$, the realized CO estimates satisfy

$$\left|\widehat{f}_h^t(z) - (\mathbb{B}_h \widehat{W}_{h+1}^t)(z)\right| \le \Gamma_h^t(z), \quad \text{for all } (h,t,z).$$

This uniform uncertainty bound supplies the confidence control used by the optimistic CO planner. The full oracle result, RKHS-norm control, and function-class containment are given in Section H.1.

Theorem C.10 (CO regret). *Let $\zeta \in (0, 1/5]$. Suppose the follower-side linear quantal-response model and the parameter domain $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$ are as specified in Definition 2.6. Suppose Assumptions 2.1, 2.2, 2.4, 2.5 and C.14 hold. Assume the conditions of Proposition C.3 and use its attained benchmark $\pi^* \in \Pi$. Assume $\lambda \ge (T+1)/T$ as required by Lemma H.1, and assume $J_{T,\mathrm{CO}}^* < \infty$ in (C.28). Run Algorithm 2 with radius B_T^{CO}. Then, with probability at least $1-3\zeta$, Algorithm 2 achieves*

$$\begin{aligned}\mathrm{Reg}_{\mathrm{CO}}(T) \lesssim\ & L_U^{\mathrm{CO}} \eta B_\Theta H \sqrt{d_F T \log(HT/\zeta)} + L_U^{\mathrm{CO}} C_{\mathrm{JDE}} B_\Theta^2 d_F H \log(HT/\zeta) \\ & + 2B_T^{\mathrm{CO}} B_\phi H \sqrt{T}\log(2eH/\zeta) + 4B_T^{\mathrm{CO}} H \sqrt{T G_K(T,\lambda)}.\end{aligned} \tag{C.31}$$

Remark C.11 (Constants and interpretation for Theorem C.10). The centered second-order softmax constant C_{JDE} is defined in (C.18). The CO action-value Lipschitz constant is $L_U^{\mathrm{CO}} = 2L_\phi B_T^{\mathrm{CO}} \lambda^{-1/2}$ by (C.22). The first two terms in (C.31) are follower-response terms, and the last two terms are leader Bellman-estimation terms. The finite-rank interpretation uses the complete CO covering complexity $\mathfrak{D}_{T,\mathrm{CO}}$.

Remark C.12 (Role of CO and comparison with the single-follower setting). For the executed local policy p at step h, CO recovers $\nu^{p,\theta_h^*}(b \mid x; s) = \mathcal{P}(x, b)/\mu(x)$ whenever $\mu(x) > 0$, so the cumulative parameter error vanishes in the population-aggregated Fisher metric. Nevertheless, the leader must learn the unknown leader reward and system transition model, and Lemma H.6 quantifies the resulting Bellman-estimation error. The follower-response and leader-estimation contributions are displayed exactly in (C.31), and their finite-rank specialization is stated in Section 4.1.

Single-follower versus mean-field followers. The linear Markov game of Chen et al. (2023a) has a single myopic follower and no endogenous population state. Both settings require learning the leader reward and transition model. In the present MFG, these primitives are indexed by the joint distribution $\mathcal{P} = \mu \otimes \nu$, the population distribution evolves endogenously, and the leader must extrapolate aggregate responses across commitment policies. These effects appear in (C.31) as separate follower-response and Bellman-estimation contributions, whose finite-rank scaling is given in (4.3). CO therefore isolates the difficulty of population-indexed planning, while FSO additionally introduces empirical-population error. It provides the natural benchmark for Section 3.

Parameter-estimation strategy (Definition 2.6). The unknown parameters are the follower-reward parameter θ_h^*, the leader-reward parameter ξ_h^*, and the transition parameter ω_h^*. Optimistic LSVI estimates the composite leader-side Bellman backup $u_h + P_h \widehat{W}_{h+1}^t$ directly through KRR, with its error controlled by the LBE term and the uncertainty quantifier in Lemma H.1 (Jin et al., 2020; Yang et al., 2020). The JDE term isolates the follower-response estimation error controlled through the response-consistent version space in (3.2).

Remark C.13 (Relation between confidence-set and point-estimator bonus planning). In Chen et al. (2023a, Algorithm 3), Scheme 4 jointly optimizes over the leader policy and a follower-parameter confidence set, whereas Scheme 5 fixes the MLE and adds a follower-side bonus. Our CO rule (3.2) follows Scheme 4 using the response-consistent version space Θ_h^{t-1}, and the FSO rule (3.7) uses the empirical Fisher confidence ellipsoid Θ_h^t. At each planner's own evaluation query, the selected optimistic follower parameter defines $\widehat{W}_h^t$, so the intermediate value-mismatch term in Lemma E.2 vanishes by construction. For FSO, converting the empirical planning query to the true mean field creates the separate batch-adaptive value mismatch controlled by (G.29).

C.7 Measurable planning, approximation conventions, and response regularity

This subsection records sufficient conditions for the CO and FSO planning oracles, approximation conventions, and the response regularity used in the analysis.

C.7.1 Oracle interpretation and approximation conventions

Assumption C.14 (Discrete optimistic planning oracle). *For $\iota \in \{\mathrm{CO}, \mathrm{FSO}\}$, let $\mathfrak{A}_{\mathrm{alg},\iota}$ be the corresponding setting-specific deterministic feasible-set family in (C.13). The admissible local-policy class $\Pi_{\mathrm{loc}} \subseteq \Delta(\mathcal{A})$ is chosen so that the following property holds for every $(h, t, s, \mu) \in [H] \times [T] \times \mathcal{S} \times \Delta(\mathcal{X})$ and every nonempty compact feasible set $\Theta' \in \mathfrak{A}_{\mathrm{alg},\iota}$ arising in setting ι. The pointwise objective $G_h^t(s, \mu; p, \theta) := \left\langle \widehat{U}_h^t\left(s, \cdot, \mu \otimes \nu^{p,\theta}(\cdot \mid \cdot; s)\right), p \right\rangle_{\mathcal{A}}$ attains its supremum over $\Pi_{\mathrm{loc}} \times \Theta'$ and admits a jointly Borel measurable maximizing selector $(s, \mu) \mapsto (p_h^t(s, \mu), \theta_h^t(s, \mu))$ realizing the supremum. For every episode t and step h, the planning correspondence admits a maximizing selector that is jointly measurable in the episode-start observable history and in $(s, \widetilde{\mu})$.*

Fix a measurable maximizing selector supplied by Assumption C.14 once and for all, and define the planning oracle to use this selector whenever the maximizer is nonunique. This fixed measurable selector is the deterministic tie-breaking convention. The full history-wise measurability requirement and sufficient topological conditions for pointwise measurable selection are given in Remark C.17.

Remark C.15 (Statistical-computational separation). The joint optimization over (p, θ) in the optimistic planner is generally non-convex: the logits $\eta\langle\varphi^p(s, x, \cdot), \theta\rangle$ are bilinear in (p, θ), and they are then composed with the nonlinear softmax map. Our results isolate the statistical sample complexity of learning under a planning oracle, following the common oracle-based convention in theoretical RL. Section C.13.5 gives separate MLE, Fisher, and KRR operation counts and identifies the planning and feature-evaluation costs excluded from those counts.

The main regret bounds use an exact benchmark in Π and an exact planning oracle over Π_{loc}. The contributions of benchmark approximation, planning approximation, and comparison with the unrestricted policy class are stated in Remark C.18.

Remark C.16 (Closed-ball parameter domain). Throughout the main analysis we take the follower-side parameter domain to be the closed Euclidean ball $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$. Bounded parameter domains and RKHS-norm constraints are common in linear and kernel RL and in Stackelberg learning (Jin et al., 2020; Yang et al., 2020; Chen et al., 2023a). Our closed-ball choice supplies a compact MLE domain and, together with the stated continuity conditions, ensures existence of MLE minimizers. The Borel measurable tie-breaking rule for the MLE in Section 3.2 exists and yields a $\mathcal{F}_{\mathrm{obs}}^{t-1}$-measurable minimizer because Θ is a compact metric space and the negative log-likelihood is Borel measurable in the leader's observed history and continuous in θ. Our closed-ball choice also supplies the compact follower-parameter component in the planning-attainment arguments.

Remark C.17 (Topological setup for the leader's action space, and existence of measurable optimistic selector). Throughout the main analysis, the leader's state space $\mathcal{S}$ and action space $\mathcal{A}$ are discrete measurable spaces, possibly countably infinite, while the follower spaces $\mathcal{X}$ and $\mathcal{B}$ are finite. The set of leader's randomized actions is

$$\Delta(\mathcal{A}) := \Big\{ p \in \ell_1(\mathcal{A}) : p(a) \geq 0,\ \sum_{a \in \mathcal{A}} p(a) = 1 \Big\}.$$

For every uniformly bounded function $f : \mathcal{A} \to \mathbb{R}$ (which holds throughout because $\widehat{U}_h^t$ is truncated to $[0, H - h + 1]$, etc.), the sum $\sum_{a \in \mathcal{A}} p(a) f(a)$ is absolutely convergent.

For a fixed episode-start history and a nonempty closed feasible set $\Theta' \subseteq \Theta$, the following conditions ensure attainment and a Borel measurable maximizing selector in (s, μ).

- *Finite $\mathcal{A}$.* $\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$ is compact (a finite-dimensional simplex), G_h^t is continuous in (p, θ) over $\Pi_{\mathrm{loc}} \times \Theta'$, and Θ' is closed; hence the supremum is attained by the Weierstrass extreme-value theorem. Moreover, $\Delta(\mathcal{A}) \times \Theta'$ is a compact metric space (hence Polish), and the map $(s, \mu, p, \theta) \mapsto G_h^t(s, \mu; p, \theta)$ is Carathéodory: Borel-measurable in (s, μ) for fixed (p, θ) and continuous in (p, θ) for fixed (s, μ). By the measurable maximum theorem for Carathéodory objectives (Aliprantis & Border, 2006, Theorem 18.19), the value function is measurable and the argmax correspondence is nonempty, compact-valued, and weakly measurable. Since the decision space is Polish, this argmax correspondence admits a Borel-measurable selector by the Kuratowski–Ryll-Nardzewski measurable selection theorem (Aliprantis & Border, 2006, Theorem 18.13).

- *Countably infinite $\mathcal{A}$ with exact compact planning.* We assume directly that $\Pi_{\mathrm{loc}} \subseteq \Delta(\mathcal{A})$ is compact in the ℓ_1 topology. Examples of local policy domains that are compact in the ℓ_1 topology include (a) any finite support class $\Pi_{\mathrm{loc}}^{(K)} = \{p \in \Delta(\mathcal{A}) : \mathrm{supp}(p) \subseteq \mathcal{A}_K\}$ for a fixed finite set $\mathcal{A}_K \subseteq \mathcal{A}$, and (b) any closed set $\Pi_{\mathrm{loc}} \subseteq \Delta(\mathcal{A})$ satisfying the uniform tail condition that for every $\epsilon > 0$ there exists a finite $\mathcal{A}_\epsilon \subseteq \mathcal{A}$ such that $\sup_{p \in \Pi_{\mathrm{loc}}} \sum_{a \notin \mathcal{A}_\epsilon} p(a) \leq \epsilon$. For any such compact policy domain, if $(s, \mu, p, \theta) \mapsto G_h^t(s, \mu; p, \theta)$ is Borel measurable in (s, μ) for every fixed (p, θ) and continuous in (p, θ) for every fixed (s, μ), then the Weierstrass theorem gives pointwise attainment, while the measurable maximum theorem and the Kuratowski–Ryll-Nardzewski measurable selection theorem yield a Borel measurable maximizing selector on the compact metric decision domain $\Pi_{\mathrm{loc}} \times \Theta'$.

Countably infinite leader actions. When $\mathcal{A}$ is countably infinite, the following mechanisms provide compact local-policy domains for the exact-attainment route in Assumption C.14:

(O1) Π_{loc} is a tight, sequentially compact subset of $\ell_1(\mathcal{A})$, for example $\Pi_{\mathrm{loc}} = \{p \in \Delta(\mathcal{A}) : \sum_a w(a) p(a) \leq M\}$ for a weight $w : \mathcal{A} \to [1, \infty)$ that diverges outside every finite set;

(O2) $\Pi_{\mathrm{loc}} = \{p \in \Delta(\mathcal{A}) : \mathrm{supp}(p) \subseteq \mathcal{A}_M\}$ for a fixed finite set $\mathcal{A}_M \subseteq \mathcal{A}$;

Discrete planning oracle for Step 1.4 of Algorithm 3. Under the continuity conditions above, when $\mathcal{A}$ is finite, $\Delta(\mathcal{A})$ is the standard finite-dimensional simplex and is compact, so the per-step planning argmax over $(p, \theta) \in \Delta(\mathcal{A}) \times \Theta_h^t$ is attained by Berge's maximum theorem. When $\mathcal{A}$ is countably infinite, $\Delta(\mathcal{A})$ is generally *not* compact in the ℓ_1 topology (mass can escape to infinity); the algorithm

therefore requires an explicit attainment assumption, formalized in Assumption C.14: there exists a class $\Pi_{\mathrm{loc}} \subseteq \Delta(\mathcal{A})$ over which the supremum

$$\sup_{p \in \Pi_{\mathrm{loc}},\, \theta \in \Theta_h^t} G_h^t(s, m; p, \theta)$$

is attained for every (s, m), with a Borel-measurable selector. For countably infinite $\mathcal{A}$, the sufficient conditions above for Assumption C.14 combine a compact local-policy domain with the stated Carathéodory property. The tight sequentially compact and fixed finite-support constructions above provide the two concrete compact-domain examples.

Under these exact-attainment conditions, the feasible correspondence

$$(s, m) \;\longmapsto\; \Pi_{\mathrm{loc}} \times \Theta_h^t$$

is weakly measurable with nonempty compact values, and the objective $G_h^t(s, m; p, \theta)$ is Carathéodory in $((s, m), (p, \theta))$. By the measurable maximum theorem for Carathéodory objectives (Aliprantis & Border, 2006, Theorem 18.19), the value function is measurable and the argmax correspondence is weakly measurable with nonempty compact values. Since the decision space is Polish, this argmax correspondence admits a Borel-measurable selector by the Kuratowski–Ryll-Nardzewski measurable selection theorem (Aliprantis & Border, 2006, Theorem 18.13); write the selector as $\big(g_h^t, \bar{\theta}_h^t\big)$, with $g_h^t : \mathcal{S} \times \Delta(\mathcal{X}) \to \Pi_{\mathrm{loc}}$ and $\bar{\theta}_h^t : \mathcal{S} \times \Delta(\mathcal{X}) \to \Theta_h^t$, realizing the pointwise maximum.

At the history-wise level, Assumption C.14 requires a maximizing selector

$$(\omega, s, \widetilde{\mu}) \longmapsto \big(p_h^t(\omega; s, \widetilde{\mu}), \vartheta_h^t(\omega; s, \widetilde{\mu})\big)$$

that is measurable with respect to $\mathcal{F}_{\mathrm{obs}}^{t-1} \otimes \mathcal{B}(\mathcal{S}) \otimes \mathcal{B}(\Delta(\mathcal{X}))$. We fix one such measurable selector once and for all and use it whenever the maximizer is nonunique; this is the deterministic tie-breaking convention. We adopt this measurable selector as the algorithm's output of Step 1.4, so the trajectory law of Algorithm 3 is a well-defined probability measure on the canonical product space.

Critical observation for the proof of Theorem 4.1. Although Π_{loc} may be infinite-dimensional when $\mathcal{A}$ is countably infinite, the proof of Theorem 4.1 does *not* require any uniform-in-p concentration over Π_{loc}: the empirical-to-population transition is handled by the *deterministic TV perturbation bound* of Lemma G.17 Stage 1 (eqs. (G.58)–(G.59)), which combines the pathwise bound $\sup_x e_h^t(x) \le 2$ with the conditional first-moment bound $\mathbb{E}^t[\|\mu_h^t - \widehat{\mu}_h^t\|_1] \le \bar{\epsilon}_N$ from Lemma G.2. The subsequent matrix Cauchy–Schwarz step uses the empirical Fisher $\widehat{\Gamma}_h^t(\theta^*)$ at the data-realized p_h^t, with the matrix elliptical-potential lemma applied directly to the additive recurrence of V_h^t (after a self-concordance comparison). Hence the analysis is fully compatible with countably infinite discrete $\mathcal{A}$.

Approximate planning and approximation of the unrestricted action simplex are accounted for outside the exact-attainment assumption in Assumption C.14. If the planner is only $\varepsilon_{\mathrm{opt}}$-accurate, or if Π_{loc} is used as a finite-support approximation to $\Delta(\mathcal{A})$ with truncation gap $\Delta_{\Pi_{\mathrm{loc}}}$, the regret displays acquire the additive slacks stated in Remark C.18.

Restricted-class approximation gap. The QSE and regret definitions use the benchmark over Π. Define

$$\Delta_{\Pi_{\mathrm{loc}}} := \sup_{\pi \in \Pi(\Delta(\mathcal{A}))} J(\pi) - \sup_{\pi \in \Pi} J(\pi). \tag{C.32}$$

The quantity $\Delta_{\Pi_{\mathrm{loc}}}$ measures the gap between the unrestricted benchmark over $\Pi(\Delta(\mathcal{A}))$ and this restricted benchmark. Regret relative to Π has no approximation-gap term, while comparison with the unrestricted local simplex adds $T\Delta_{\Pi_{\mathrm{loc}}}$. When $\mathcal{A}$ is finite and $\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$, one has $\Delta_{\Pi_{\mathrm{loc}}} = 0$.

Remark C.18 (Benchmark and planning slacks). A measurable $\varepsilon_{\mathrm{bench}}$-optimal benchmark means a policy $\pi^{\mathrm{bench}} \in \Pi$ satisfying $J(\pi^{\mathrm{bench}}) \ge \sup_{\pi \in \Pi} J(\pi) - \varepsilon_{\mathrm{bench}}$. Under this measurable $\varepsilon_{\mathrm{bench}}$-optimal benchmark convention, add $T\varepsilon_{\mathrm{bench}}$ to the regret upper bound. If each planning step uses a feasible $\varepsilon_{\mathrm{opt}}$-maximizing planning selector satisfying the same joint measurability requirements as in Assumption C.14 and the continuation backups retain the exact value envelopes, add $HT\varepsilon_{\mathrm{opt}}$. These contributions are additive when both approximations are used. Here $\varepsilon_{\mathrm{opt}}$ bounds the objective suboptimality in each joint maximization over (p, θ). When comparison is made with the unrestricted policy class $\Pi(\Delta(\mathcal{A}))$, additionally add $T\Delta_{\Pi_{\mathrm{loc}}}$, as defined in (C.32). The regret upper bounds in this paper are stated for an exact benchmark in Π and an exact planning oracle over the chosen local policy class Π_{loc}, with $\varepsilon_{\mathrm{bench}} = \varepsilon_{\mathrm{opt}} = 0$.

Planning accuracy. The planning-approximation conditions of Remark C.18 apply to the following accuracy choices.

If only an ϵ_{opt}-accurate planner is available and the continuation backups retain the exact value envelopes, the same telescoping argument adds at most $HT\epsilon_{\text{opt}}$ to regret; hence taking $\epsilon_{\text{opt}} = O(1/(HT))$ makes this computational error $O(1)$, and any ϵ_{opt} satisfying $HT\epsilon_{\text{opt}} = o(\sqrt{T})$ is lower order relative to the $\sqrt{T}$-type learning terms.

Remark C.19 (Why no policy-selector Lipschitz condition is needed). A naive function-class plug-in analysis (in the spirit of Π_{Lip}-restricted policy classes) would impose a uniform Lipschitz condition on the leader's policy selector $\mu \mapsto \pi(\cdot|s,\mu)$, with constant L_π independent of the horizon. Our pointwise planner avoids this entirely: Step 1.4 of Algorithm 3 optimizes pointwise over $p \in \Pi_{\text{loc}}$ at each (s,μ), and the regret analysis (Theorem 4.1) uses only the *value-level* Lipschitzness of $\widehat{W}_h^t(s,\cdot)$ established in Lemma C.23. The value-envelope bound uses the leader kernel-feature Lipschitz condition and the verified RKHS base-function norm cap. In particular, the proof never compares $g_h^t(s,\mu)$ with $g_h^t(s,\mu')$, only $G_h^t(s,\mu;p,\theta)$ with $G_h^t(s,\mu';p,\theta)$ for fixed (p,θ), so no continuity of the argmax selector g_h^t in μ is invoked. Thus, tie-breaking discontinuities or jumps of the measurable selector affect the realized policy but not the value-envelope stability used in the finite-sample plug-in analysis.

C.8 Supplementary discussion for the FSO algorithm

C.8.1 Paired observation protocol and batch-adaptive commitment values

The FSO analysis reuses the CO decomposition and adds empirical population concentration, follower-side MLE concentration, target-radius inflation, and the TV perturbation bound needed to control plug-in errors.

In contrast to the CO setting studied in Section 3.1, in the FSO setting, the leader observes only N follower samples per step, not the true population distribution. The FSO commitment-rule class used by OVI-FSO and by the main FSO value and regret analysis is defined in (C.33). At the beginning of episode t, the leader uses only $\mathcal{F}_{\text{obs}}^{t-1}$ to select and publicly commit to the full tuple $g^t = \{g_h^t\}_{h=1}^H$. This tuple remains fixed throughout the episode. At step h, the current follower-state batch supplies a realized argument of the already fixed map g_h^t and produces the realized local policy $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$. The leader discloses p_h^t before the followers act, but neither recomputes g^t nor revises the episode-start commitment. To formalize the observation setting, we adopt the *one-batch paired adaptive protocol* under FSO, as formulated in Assumption 2.3.

Equation (C.33) specifies the commitment-rule class analyzed under the FSO observation protocol. Pairing means retaining the correspondence between a sampled follower state and its subsequent action at the same step; it does not require tracking the follower's identity across steps or episodes. The same-cohort design is not essential: a two-cohort variant may use one cohort to form $\widehat{\mu}_h^t$ and a second cohort to provide paired state-action observations under the resulting fixed p_h^t; see Section C.13.3. Theorem 4.1 is stated for the one-batch paired protocol, and the state-conditioned follower-response likelihood still requires state-action pairing within the response cohort; observing only separate state and action marginals changes the observation model and can preclude a uniform $\widetilde{O}(\sqrt{T})$ guarantee, as shown in Proposition F.3.

FSO commitment rules and their value. The commitment rule class used by OVI-FSO and by the main FSO value and regret formulation is

$$\mathfrak{G}_{\text{FSO}} := \prod_{h=1}^{H} \{g_h : \mathcal{S} \times \Delta(\mathcal{X}) \to \Pi_{\text{loc}} : g_h \text{ is Borel measurable}\}. \tag{C.33}$$

Each step-h component is evaluated at the current leader state s_h and empirical mean field $\widehat{\mu}_h$. An online algorithm selects $g^t \in \mathfrak{G}_{\text{FSO}}$ using only $\mathcal{F}_{\text{obs}}^{t-1}$ and freezes it before episode t begins. The batch-adaptive value recursion evaluates the rules in this class.

The batch-adaptive value and initial return are defined in (2.6) and (2.7).

FSO regret against the CO benchmark. For an episode-start sequence $g^t \in \mathfrak{G}_{\mathrm{FSO}}$, the FSO regret against the exact-population CO benchmark $J(\pi^*)$ is

$$\mathrm{Reg}_{\mathrm{FSO}}(T) := \sum_{t=1}^{T} \big[J(\pi^*) - J_{\mathrm{FSO}}(g^t)\big], \qquad J_{\mathrm{FSO}}(g) = \mathbb{E}_{\mathrm{FSO}}\big[\textstyle\sum_h u_h(s_h, a_h, \mathcal{P}_h)\big]. \tag{C.34}$$

The second equality is by (2.7). $\pi^* \in \Pi$ is evaluated on the true population state in the CO benchmark, whereas each executable FSO rule g^t can use only its observable arguments $(s_h^t, \widehat{\mu}_h^t)$. The two values are therefore evaluated under their respective commitment-rule classes. Remark C.41 gives the exact decomposition into learning regret within the FSO commitment-rule class and the gap between its best value and the CO benchmark. That benchmark gap is conceptually distinct from the upper bound's plug-in residual.

Under Assumption 2.3, Lemma G.3 gives, with probability at least $1 - \zeta$,

$$\|\widehat{\mu}_h^t - \mu_h^t\|_1 \le \bar{\epsilon}_N, \qquad \|\widehat{\mathcal{P}}_h^t - \mathcal{P}_h^t\|_1 \le \bar{\epsilon}_N, \qquad (h,t) \in [H] \times [T].$$

Here $\bar{\epsilon}_N = \min\{\epsilon_N, 2\}$, where the exact untruncated scale ϵ_N is defined in (G.11) and satisfies $\epsilon_N = \widetilde{O}(\sqrt{|\mathcal{Z}|/N})$. The explicit bounds and numerical constants are recorded with Lemma G.3.

C.8.2 Full OVI-FSO estimation, planning, and execution

The likelihood, empirical Fisher matrix, and confidence ellipsoid are defined in (3.3), (3.4), and (3.5). The FSO algorithm uses the actual uncertainty radius B_T^{FSO} defined in (G.34) and the action-value Lipschitz constant $L_U^{\mathrm{FSO}} = 2L_\phi B_T^{\mathrm{FSO}} \lambda^{-1/2}$ from (G.35). Under (G.36), this radius absorbs the historical Bellman-target perturbation and satisfies $B_T^{\mathrm{FSO}} \le 4B_T^{\mathrm{base}}$ by Lemma G.12. The current-query correction uses the separate bonus $L_U^{\mathrm{FSO}} \bar{\epsilon}_N$ in Algorithm 3. The pointwise optimistic planning rule and its attained value are given by (3.7) and (3.8).

On the observed batch, write $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$ and $\bar{\theta}_h^t = \bar{\theta}_h^t(s_h^t, \widehat{\mu}_h^t)$ for the realized maximizing pair in (3.7). The measurable construction for arbitrary hypothetical batches is given in (G.4) and (G.5).

We formalize Optimistic Value Iteration for Finite-Sample Observation (OVI-FSO) in Algorithm 3.

Remark C.20 (Two modifications relative to OVICO). OVI-FSO differs from Algorithm 2 in two statistical components. *(M1) MLE confidence ellipsoid.* The exact version-space constraint is replaced by the empirical-Fisher ellipsoid $\Theta_h^t = \{\theta \in \Theta : \|\theta - \widehat{\theta}_h^t\|_{\widehat{F}_h^t}^2 \le \beta_{\mathrm{F}}^t\}$. Its simultaneous validity is established by Lemma G.5. The corresponding auxiliary V_h^t-norm bound follows from Lemma C.31. *(M2) Finite-sample uncertainty correction.* Historical target perturbations are absorbed into the actual radius B_T^{FSO} through (G.34). The current-query perturbation is controlled by the step-independent bonus $\Gamma_{\mathrm{FS}} = L_U^{\mathrm{FSO}} \bar{\epsilon}_N$. The baseline radius, deterministic cap, and actual radius are defined separately in (C.25), (C.20), and (G.34).

Policy-level and local response notation. The policy-level notation $\nu^{\pi,\theta}(\cdot|x; s, \mu)$ used in the Complete Observation analysis denotes the softmax response with logits $\eta\langle \sum_a \pi(a|s,\mu)\varphi(s,a,x,b),\, \theta\rangle$, where π is a Markov commitment policy. In the pointwise framework adopted in this section, the leader's relevant local policy at (s,μ) is the realized distribution $p \in \Delta(\mathcal{A})$ (a single point of the simplex, not a function), and the corresponding softmax response is $\nu^{p,\theta}(\cdot|x; s)$ from (C.2), with logits $\eta\langle \varphi^p(s,x,b), \theta\rangle$. The two notations are linked by the identity

$$\nu^{\pi,\theta}(\cdot|x; s, \mu) \;=\; \nu^{\pi(\cdot|s,\mu),\,\theta}(\cdot|x; s),$$

i.e., evaluating $\nu^{\pi,\theta}$ at a Markov commitment policy π amounts to applying the pointwise notation at $p = \pi(\cdot|s,\mu)$. In particular, when $\pi = g_h^\tau$ is the rule from Algorithm 3 and we evaluate at $(s_h^\tau, \widehat{\mu}_h^\tau)$, we have $\nu^{g_h^\tau,\theta}(\cdot|x; s_h^\tau, \widehat{\mu}_h^\tau) = \nu^{p_h^\tau,\theta}(\cdot|x; s_h^\tau)$ with $p_h^\tau = g_h^\tau(s_h^\tau, \widehat{\mu}_h^\tau)$. We use the pointwise $\nu^{p,\theta}$ form in all FSO analysis below.

Pointwise optimistic follower parameters. The maximizing parameter $\bar{\theta}_h^t(s,m) \in \Theta_h^t$ in (3.7) varies pointwise with (s,m) and is an *optimistic follower parameter* for that input: it certifies the optimistic value $\widehat{W}_h^t(s,m)$ via the constructive max over the MLE confidence ellipsoid Θ_h^t, but it is

Algorithm 3: Optimistic Value Iteration for Finite-Sample Observation (OVI-FSO).

Input: T; H; N; UCB radius $\beta = B_T^{\mathrm{FSO}}$; uniform MLE radius β_{F}^T; regularizers λ and λ_θ.

Initialize $\mathcal{D}^0 = \varnothing$; $\widehat{\theta}_h^1 = 0 \in \Theta$, $\widehat{F}_h^1 = \lambda_\theta I_{d_F}$, $\Theta_h^1 = \Theta$ for all $h \in [H]$;

for *episode* $t = 1, \dots, T$ **do**

  Set $\widehat{W}_{H+1}^t \equiv 0$;

  // **Phase 1: Backward sweep (offline planning)**

  for *step* $h = H, H-1, \dots, 1$ **do**

    (Step 1.1) Compute $\widehat{f}_h^t$ via (3.6) using the empirical KRR inputs $\widehat{z}_h^\tau = (s_h^\tau, a_h^\tau, \widehat{\mathcal{P}}_h^\tau)$ and targets $u_h^\tau + \widehat{W}_{h+1}^t(s_{h+1}^\tau, \widehat{\mu}_{h+1}^\tau)$ for $\tau < t$;

    (Step 1.2) Form Γ_h^t via (3.6) with radius B_T^{FSO} and set $\Gamma_{\mathrm{FS}} := L_U^{\mathrm{FSO}} \bar{\epsilon}_N$;

    (Step 1.3) Set $\widehat{U}_h^t = \mathrm{Trun}_{[0,H-h+1]}(\widehat{f}_h^t + \Gamma_h^t + \Gamma_{\mathrm{FS}})$;

    (Step 1.4) Obtain g_h^t and $\bar{\theta}_h^t$ via (3.7), and obtain $\widehat{W}_h^t$ via (3.8);

  Commit to and publicly announce the full rule tuple $g^t = \{g_h^t\}_{h=1}^H$;

  // **Phase 2: Forward roll-out (online execution)**

  for *step* $h = 1, 2, \dots, H$ **do**

    (Step 2.1) Observe $\{x_h^{t,n}\}_{n=1}^N$ from N followers, and form $\widehat{\mu}_h^t$ via (2.4);

    (Step 2.2) Evaluate the committed rule via $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$ and disclose p_h^t;

    (Step 2.3) Sample $a_h^t \sim p_h^t$ and observe u_h^t;

    (Step 2.4) Observe $\{b_h^{t,n}\}_{n=1}^N$ from the same N followers and form $\widehat{\mathcal{P}}_h^t$ via (2.5);

    (Step 2.5) The system evolves to (s_{h+1}^t, μ_{h+1}^t);

  Update dataset: $\mathcal{D}^t \leftarrow \mathcal{D}^{t-1} \cup \{(s_h^t, a_h^t, p_h^t, u_h^t, \widehat{\mu}_h^t, \widehat{\mathcal{P}}_h^t, \{(x_h^{t,n}, b_h^{t,n})\}_{n=1}^N)\}_{h=1}^H$;

  For each $h \in [H]$, compute $\widehat{\theta}_h^{t+1} \in \mathrm{argmin}_{\theta \in \Theta} \mathcal{L}_h^{t+1}(\theta)$ where $\mathcal{L}_h^{t+1}$ uses data $\mathcal{D}^t$ in (3.3); then update $\widehat{F}_h^{t+1}$ via (3.4) and Θ_h^{t+1} via (3.5);

Output: Commitment rules $\{g_h^t\}_{h\in[H]}^{t\in[T]}$ and their realized local policies $\{p_h^t\}_{h\in[H]}^{t\in[T]}$.

not announced to the followers and is not executed by the leader (the leader samples $a_h^t \sim p_h^t$, where $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$). Since followers are myopic, the step-h likelihood depends only on θ_h and not on cross-step parameters; hence different (s, m) may use different optimistic follower parameters without violating any cross-step coupling. This rectangularization is standard in optimistic RL with confidence-set planning (cf. Azar et al., 2017; Jin et al., 2020).

Execution protocol: one-batch paired adaptivity. Before episode t begins, the leader selects from $\mathcal{F}_{\mathrm{obs}}^{t-1}$ and announces the full tuple $g^t = \{g_h^t\}_{h=1}^H$, which remains fixed throughout the rollout. At step h, the leader first observes the N states $\{x_h^{t,n}\}_{n=1}^N$ and forms $\widehat{\mu}_h^t$; she then evaluates the committed rule to obtain and disclose the realized local policy $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$; finally the same followers generate actions $\{b_h^{t,n}\}_{n=1}^N$ according to the quantal response under this local policy. Three properties are used throughout the analysis:

(a) *No second batch is required*: the empirical mean field and the empirical joint distribution use the same N followers, so the per-step follower-observation cost is N rather than $2N$.

(b) *Adaptive dependence is conditionally harmless*: p_h^t may depend on the sampled states through $\widehat{\mu}_h^t$, but after conditioning on the state batch, p_h^t is fixed and the subsequent action draws are conditionally independent. The paired adaptive empirical concentration lemma (Lemma G.3) exploits this conditional structure.

(c) *Rule-level commitment and step-level disclosure are distinct*: the entire map tuple g^t is selected and committed before the episode. After observing the current sampled state distribution at step h, the leader evaluates that fixed tuple at the current observable input, realizes and discloses $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$, and the follower population at the true mean field μ_h^t responds to this realized local policy. This stepwise realization and disclosure of p_h^t is not a revision of g^t.

Thus all sample-size statements in the sequel are stated directly in terms of the one-batch size N.

Remark C.21 (Executed local policy and batch-adaptive analysis). At step h, the leader evaluates the jointly measurable rule from (G.4) at the realized state batch and executes $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$. The same realized local policy is evaluated under the empirical planning input and the true population input. The resulting planning mismatch is controlled by the two distinct shifts in (G.29). The bound uses the value-envelope and fixed-pair-objective Lipschitz constants separately and does not require a Lipschitz policy selector. The current-batch expectation is controlled by Lemma G.2, rather than by conditioning on a nonpredictable global empirical event.

Remark C.22 (Role of the uniform finite-sample bonus). The uniform bonus $\Gamma_{\mathrm{FS}} = L_U^{\mathrm{FSO}} \bar{\epsilon}_N$ controls the current-query conversion from the empirical joint distribution to the true joint distribution. Historical regression-target perturbations are handled separately by the radius inflation in (G.34). Because the bonus is nonnegative, it preserves the true-query optimism supplied by Corollary G.15. Its cumulative contribution to the leader Bellman error is $L_U^{\mathrm{FSO}} HT\bar{\epsilon}_N$ before the finite-sample terms are collected in (C.40). The bonus is constant in the query before truncation, although it can change the set of maximizers when the truncation ceiling becomes active. The measurable tie-breaking convention resolves such ties without affecting the optimistic value.

C.9 Lipschitz, stability, and optimism lemmas for the FSO regret analysis

For the FSO analysis below, write $\mathcal{E}_{\mathrm{or}}^t := \mathcal{E}_{\mathrm{or,FSO}}^t$, $\mathcal{E}_{\mathrm{or}} := \mathcal{E}_{\mathrm{or,FSO}}$, $\widehat{f}_h^{t,\mathrm{or}} := \widehat{f}_{h,\mathrm{FSO}}^{t,\mathrm{or}}$, and $\widehat{\Lambda}_{L,h,\mathrm{FSO}}^t = \widehat{\Lambda}_{L,h}^t$.

Lemma C.23 (Value-level Lipschitz continuity of the pointwise optimistic planner). *Consider the response model in (C.2), and assume Assumption 2.4. Fix $h \in [H]$, $t \in [T]$, and a nonempty feasible set $\Theta_h^t \in \mathfrak{A}_{\mathrm{alg,FSO}}$ that remains fixed when the hypothetical population argument is varied. Let $\widehat{U}_h^t \in \mathcal{U}_{\mathrm{UCB}}(h, R_{T,\mathrm{FSO}}^{\mathrm{base}}(B_T^{\mathrm{FSO}}), B_T^{\mathrm{FSO}}, L_U^{\mathrm{FSO}}\bar{\epsilon}_N)$, and let G_h^t and $\widehat{W}_h^t$ be its fixed-pair objective and value envelope over $\Pi_{\mathrm{loc}} \times \Theta_h^t$. Then, for every $s \in \mathcal{S}$ and $\mu, \mu' \in \Delta(\mathcal{X})$,*

$$\left|\widehat{W}_h^t(s,\mu) - \widehat{W}_h^t(s,\mu')\right| \le L_{\widehat{W}} \|\mu - \mu'\|_1, \tag{C.35}$$

where the step-independent choice is

$$L_{\widehat{W}} := C_W^{\mathrm{act}}\left(B_T^{\mathrm{FSO}}\right) \le C_{\mathrm{pair}}^{\mathrm{env}}\left(B_T^{\mathrm{FSO}}\right) = L_U\left(B_T^{\mathrm{FSO}}\right) = L_U^{\mathrm{FSO}}. \tag{C.36}$$

No factor $H - h + 1$ is introduced. The comparison holds with the same feasible set Θ_h^t at μ and μ', and it does not require any regularity of the maximizing selector.

Remark C.24 (Radius-indexed Lipschitz envelope of the optimistic estimator). For any member of the radius-indexed UCB class with RKHS base-function cap $\mathsf{r}\lambda^{-1/2}$ and bonus coefficient at most r, the KRR component and the RKHS bonus contribute at most $\mathsf{r}L_\phi\lambda^{-1/2}$ each. The scalar correction contributes zero and truncation is nonexpansive, so $L_U(\mathsf{r}) = 2L_\phi\mathsf{r}\lambda^{-1/2}$. For the actual FSO iterate, the required class membership follows on $\mathcal{E}_{\mathrm{or}}^t \cap \mathcal{E}_{\mathrm{emp}}^{<t}$ from Corollary G.15; at $\mathsf{r} = B_T^{\mathrm{FSO}}$, write $L_U^{\mathrm{FSO}} := L_U(B_T^{\mathrm{FSO}})$.

Lemma C.25 (Plug-in stability of the pointwise argmax value). *Suppose the conditions of Lemma C.23 hold. Fix h, t and condition on the prefix determining Θ_h^t. Let*

$$(p_{\widehat{\mu}}, \bar{\theta}_{\widehat{\mu}}) \in \operatorname*{argmax}_{(p,\theta)\in\Pi_{\mathrm{loc}}\times\Theta_h^t} G_h^t(s, \widehat{\mu}; p, \theta)$$

denote any pointwise optimizer of G_h^t at $(s, \widehat{\mu})$ over the admissible class $\Pi_{\mathrm{loc}} \times \Theta_h^t$. Then for every $\mu, \widehat{\mu} \in \Delta(\mathcal{X})$,

$$0 \le \widehat{W}_h^t(s,\mu) - G_h^t(s,\mu; p_{\widehat{\mu}}, \bar{\theta}_{\widehat{\mu}}) \le 2C_{\mathrm{pair}}^{\mathrm{env}}(B_T^{\mathrm{FSO}})\|\mu - \widehat{\mu}\|_1. \tag{C.37}$$

The conclusion uses only value-level Lipschitz constants and does not require a Lipschitz maximizing selector.

Corollary C.26 (Current-batch state-bias bound). *Let $(p_h^t, \bar{\theta}_h^t)$ be the realized optimizer obtained from the episode-t hypothetical-batch selector. On the predictable historical event used in Lemma G.9,*

$$\sum_{t,h} \mathbb{E}^t\left[\widehat{W}_h^t(s_h^t, \mu_h^t) - G_h^t(s_h^t, \mu_h^t; p_h^t, \bar{\theta}_h^t)\right]$$
$$\le HT\bar{\epsilon}_N\left[C_W^{\mathrm{act}}(B_T^{\mathrm{FSO}}) + C_G^{\mathrm{act}}(B_T^{\mathrm{FSO}})\right] \le 2HT\bar{\epsilon}_N C_{\mathrm{pair}}^{\mathrm{env}}(B_T^{\mathrm{FSO}}).$$

Lemma C.27 (Per-episode pointwise optimism under FSO). *Fix $t \in [T]$. On $\mathcal{E}^t_{\mathrm{or}} \cap \mathcal{E}^{<t}_{\mathrm{emp}} \cap \mathcal{E}_{\mathrm{MLE,main}}$, for every h and (s, μ),*

$$\widehat{W}^t_h(s,\mu) \geq W^{\pi^*}_h(s,\mu). \tag{C.38}$$

Consequently, $\widehat{W}^t_1(s_0,\mu_0) \geq J(\pi^*)$ and $\sum_{t=1}^T [J(\pi^*) - \widehat{W}^t_1(s_0,\mu_0)] \leq 0$ on the intersection of the predictable historical events. The benchmark local policy belongs to the same class Π_{loc} used by the planner.

Remark C.28 (Comparator optimism and executed-path uncertainty). Lemma C.27 uses a confidence inequality uniform over the true evaluation domain, so it evaluates the comparator query directly and never applies an elliptical-potential argument to a comparator trajectory. After pointwise optimism removes the comparator term, the LBE analysis uses elliptical potential only for empirical queries generated by the executed path and stored in the design operator.

Remark C.29 (Where the current-batch mean-field bias enters). The episode-start rule is fixed before the current batch and evaluated at $\widehat{\mu}^t_h$. The mismatch at μ^t_h is the sum of the value-envelope shift and the fixed empirical-optimizer objective shift; the empirical optimization gap is zero. Lemma C.25 and corollary C.26 and (G.29) control these shifts inside the current-batch conditional expectation. No current-batch pathwise event or recursive population-dynamics propagation is used.

C.10 Supplementary material for the FSO LBE and JDE

Remark C.30 (Event composition for Lemma G.16). Episode t uses the predictable historical event $\mathcal{E}^t_{\mathrm{or}} \cap \mathcal{E}^{<t}_{\mathrm{emp}}$. The current true-to-empirical query error remains inside $\mathbb{E}^t[\cdot]$ and is controlled by Lemma G.2. Converting expected bonuses to realized empirical-design bonuses additionally uses $\mathcal{E}_{\text{LBE-mart}}$.

Lemma C.31 (Formal PSD-equivalence of algorithmic $\widehat{F}^t_h$ and auxiliary additive Fisher matrix V^t_h). *The algorithmic empirical Fisher matrix $\widehat{F}^t_h$ from (3.4) (with all historical contributions evaluated at the current MLE $\widehat{\theta}^t_h$) and the auxiliary additive Fisher matrix V^t_h from (G.15) (with each historical contribution evaluated at the contemporaneous MLE $\widehat{\theta}^\tau_h$, admitting strict additive recurrence) satisfy the PSD sandwich*

$$C_{\mathrm{sc}}^{-1} V^t_h \preceq \widehat{F}^t_h \preceq C_{\mathrm{sc}} V^t_h \quad \text{for all } (h,t) \in [H] \times [T], \tag{C.39}$$

where $C_{\mathrm{sc}} = e^{4\eta B_\Theta}$ is the deterministic pointwise Fisher comparison factor from Lemma G.1.

The two factors C_{sc} (linear) and C^2_{sc} (quadratic) come from *one* self-concordance comparison $\widehat{\Gamma}(\theta^*) \to \widehat{\Gamma}(\widehat{\theta})$ inside the elliptical-potential trace bound, and *one* self-concordance comparison from the algorithmic $\widehat{F}$-norm radius β^t_{F} to the analytic V-norm radius $C_{\mathrm{sc}}\beta^t_{\mathrm{F}}$ used in matrix Cauchy–Schwarz (cf. (G.16)).

Remark C.32 (Why the TV perturbation bound resolves the $T^{3/4}$ obstruction). A naive approach to Stage 3 would be to apply Jensen on the population weight μ_h rather than $\widehat{\mu}^t_h$, giving the population Fisher $\bar{\Sigma}^{p^t_h,\theta^*_h}_{s^t_h,\mu^t_h}$ in place of $\widehat{\Gamma}^t_h(\theta^*_h)$. Then a deterministic operator-norm bound $\left\|\bar{\Sigma}_{\mu^t_h} - \widehat{\Gamma}^t_h\right\|_{\mathrm{op}} \leq \bar{\epsilon}_N R_\Sigma$ would translate to a trace residual $\sum_t \mathrm{Tr}[(V^t_h)^{-1}]\bar{\epsilon}_N R_\Sigma \leq T d_F R_\Sigma \bar{\epsilon}_N/\lambda_\theta$, which after Cauchy–Schwarz over t produces a term of order $T\sqrt{\beta^T_{\mathrm{F}} d_F \bar{\epsilon}_N/\lambda_\theta}$. At the threshold scale $\bar{\epsilon}_N = O(T^{-1/2})$ implied by (G.37), up to fixed problem constants and logarithms, this route contains a $T^{3/4}$ temporal factor and is not dominated by $\sqrt{T}$. Our TV perturbation bound (G.58)–(G.59) avoids this by performing the empirical-to-population transition *before* Jensen and Cauchy–Schwarz: the correction proportional to $\bar{\epsilon}_N$ is additive outside every square root and is $O(H\sqrt{T})$ under the same threshold scaling, which preserves the target $\sqrt{T}$ temporal order. This is the structural reason for the cleanness of (G.57): additive residuals from the TV perturbation bound tolerate $\bar{\epsilon}_N = O(T^{-1/2})$, whereas the multiplicative route would require $\bar{\epsilon}_N = O(T^{-1})$, corresponding to N of order T^2 up to problem scales and logarithms.

C.11 Supplementary discussion for the FSO main theorem

The full RKHS guarantee is stated in Theorem 4.1, and its four-term rate interpretation is (4.2) under the additional conditions specified there. Corollary 4.3 evaluates this complexity for finite-rank

leader kernels. We record the resulting data-abundant regime, expectation conversion, and complete finite-time comparison below.

C.11.1 Data-abundant interpretation and expectation conversion

Remark C.33 (Data-abundant finite-rank regime). Suppose $d_F \lesssim \sqrt{d_L}$, $T \gtrsim d_F^2$, and $N = \widetilde{\Omega}(T|\mathcal{Z}|)$. Suppose also that either $R_\Sigma = 0$ or $N\eta^2 R_\Sigma \geq \lambda_\theta$, and retain the regularizer and radius-absorption conditions of Corollary 4.3. Then $D_L = \Theta(d_L^2)$ and $\mathfrak{D}_T = \widetilde{O}(d_L^2)$, so

$$\mathrm{Reg}_{\mathrm{FSO}}(T) = \widetilde{O}\Big(d_L^{3/2} H^2 \sqrt{T}\Big).$$

The d_L^2 posterior-shape entropy yields an additional factor $\sqrt{d_L}$ relative to $\widetilde{O}(d_L H^2 \sqrt{T})$.

Remark C.34 (High-probability versus in-expectation comparison). Since $\mathrm{Reg}_{\mathrm{FSO}}(T) \leq HT$ deterministically and Theorem 4.1 holds with probability at least $1 - 5\zeta$, we have $\mathbb{E}[\mathrm{Reg}_{\mathrm{FSO}}(T)] \leq$ [right-hand side of (4.1)] $+ 5\zeta HT$. Choosing $\zeta = (5HT)^{-1}$ changes only logarithmic factors.

C.11.2 Overview of the complete finite-time results

Algorithm and rate overview. We propose OVI-FSO, an optimistic value iteration algorithm for FSO, combining *(i)* RKHS-based LSVI for the leader, *(ii)* an MLE-based confidence set for the followers' reward parameter, and *(iii)* a plug-in estimator for the population distribution from empirical samples.

Let d_L be the rank of the leader kernel, let d_F be the follower feature dimension, and set $D_L = 1 + d_L + d_L^2 + d_F^2$. Under the conditions of Corollary 4.3, OVI-FSO satisfies the finite-rank regret bound summarized in Table 7. If $d_F \lesssim \sqrt{d_L}$, $T \gtrsim d_F^2$, and $N = \widetilde{\Omega}(T|\mathcal{Z}|)$, this bound reduces to $\widetilde{O}(d_L^{3/2} H^2 \sqrt{T})$.

The CO lower bound $\Omega(d_L H \sqrt{T})$ shows that the square-root dependence on T is unavoidable. Where both CO bounds apply in this dimension regime, the current complete-cover CO upper bound leaves factors $\sqrt{d_L}$ and H in the leader-side dimension and horizon, respectively, up to logarithmic factors. For the FSO commitment-rule class in (C.33), Theorem C.46 gives $\Omega(T/\sqrt{N})$ expected regret on a fixed-dimensional family, so the linear T exponent in the sample requirement is unavoidable within this class.

Table 7: Summary of the main finite-time guarantees. Here $D_L = 1 + d_L + d_L^2 + d_F^2$ is recalled from Section 4.1. Logarithmic and fixed problem-dependent constants are suppressed.

Setting	Main guarantee	Interpretation
Finite-rank CO	$\widetilde{O}\big(H^2\sqrt{T d_L D_L} + \sqrt{d_F}\, H^2 \sqrt{T D_L} + d_F H^2 \sqrt{D_L}\big)$	Complete-cover CO upper bound
Finite-rank FSO	$\widetilde{O}\big(H^2\sqrt{T d_L D_L} + d_F H^2 \sqrt{T D_L} + d_F^2 H^2 \sqrt{D_L} + H^2 T \sqrt{D_L}\sqrt{\lvert\mathcal{Z}\rvert/N}\big)$	Complete-cover FSO upper bound
Data-abundant finite-rank FSO	$\widetilde{O}(d_L^{3/2} H^2 \sqrt{T})$ under $d_F \lesssim \sqrt{d_L}$, $T \gtrsim d_F^2$, and $N = \widetilde{\Omega}(T\lvert\mathcal{Z}\rvert)$	Simplified rate under the corollary conditions
FSO lower bound	$\Omega(T/\sqrt{N})$	$N = \widetilde{\Omega}(T)$ is necessary for expected $\widetilde{O}(\sqrt{T})$ within $\mathfrak{G}_{\mathrm{FSO}}$
Finite-rank CO lower bound	$\Omega(d_L H \sqrt{T})$	$\sqrt{T}$ is unavoidable; gaps remain

Remark C.35 (Explicit constants in (4.1)). The leader radius is B_T^{FSO} from (G.34), and the setting-specific action-value Lipschitz constant is $L_U^{\mathrm{FSO}} = L_U(B_T^{\mathrm{FSO}})$. The explicit sample-size condition in (G.37) is sufficient for the radius-absorption conditions in (G.36). The constant C_{sc} is the common self-concordance envelope from (C.18), and β_{F}^T is the final confidence radius from Lemma G.5. Define

$$C_{\mathrm{FS}}(\mathsf{r}) := 2\mathsf{r} L_\phi \lambda^{-1/2} + L_U(\mathsf{r}) + 2C_{\mathrm{pair}}^{\mathrm{env}}(\mathsf{r}) + 2L_U(\mathsf{r}). \tag{C.40}$$

At the actual FSO radius, the finite-sample coefficient satisfies $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}}) = 6L_U^{\mathrm{FSO}}$.

With the continuous convention $L_{T,N} = 0$ when $R_\Sigma = 0$, define the exact trace factor by

$$L_{T,N} := \frac{R_\Sigma/\lambda_\theta}{\log(1 + N\eta^2 R_\Sigma/\lambda_\theta)} \log\left(1 + \frac{TN\eta^2 R_\Sigma}{\lambda_\theta d_F}\right). \tag{C.41}$$

In the degenerate case $R_\Sigma = 0$, the follower response has no parameter variation and the corresponding JDE contribution is zero. Define the per-step JDE increment envelope by

$$R_Y := 4\eta B_\Theta \sqrt{R_\Sigma} + 4C_{\mathrm{JDE}} B_\Theta^2 R_\Sigma. \tag{C.42}$$

The exact terms linear in $HT\bar{\epsilon}_N$ are collected by $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})$ from (C.40). Historical target perturbations are already contained in B_T^{FSO} and are not added again. The simplified rate is stated separately in (4.2) under its explicit additional conditions.

Remark C.36 (Explicit d_F-dependence inside β_{F}^t). Since $\Theta = \overline{\mathbb{B}(0, B_\Theta)} \subset \mathbb{R}^{d_F}$, the exact scale in (G.19) is

$$\varepsilon_{\mathrm{net}} = \frac{1}{1 + TN[\eta(1 + B_A^\Theta)B_\varphi + \bar{L}_H]}, \qquad \bar{L}_H := 2\eta B_\varphi + 8\eta^2(1 + \eta B_A^\Theta)^2 B_\varphi^2.$$

The standard Euclidean covering bound (see, e.g., Vershynin, 2018, Corollary 4.2.13) gives

$$\log \mathcal{N}_\Theta(\varepsilon_{\mathrm{net}}) \le d_F \log\big(1 + 2B_\Theta\big(1 + TN[\eta(1 + B_A^\Theta)B_\varphi + \bar{L}_H]\big)\big).$$

Substituting this cover bound into (G.13) gives

$$\beta_{\mathrm{F}}^t \le 8C_{\mathrm{sc}}(1 + \eta B_A^\Theta)^2 \Big[1 + \log(H/\zeta) + d_F \log\big(1 + 2B_\Theta\big(1 + TN[\eta(1 + B_A^\Theta)B_\varphi + \bar{L}_H]\big)\big)\Big] + 4C_{\mathrm{sc}}\lambda_\theta B_\Theta^2.$$

Thus $\beta_{\mathrm{F}}^T = \widetilde{O}(d_F)$ under fixed problem constants. The T dependence enters through the covering scale and not through a separate union bound over t. This is the scaling used in Theorem 4.1 and corollary 4.3.

Remark C.37 (Dependence on the inverse-temperature parameter η). The main FSO theorem uses the well-specified follower MLE confidence radius and the common self-concordance envelope in (C.18). Under $\eta = O(\log|\mathcal{B}|)$ and fixed B_Θ, factors of the form $\exp(O(\eta B_\Theta))$ are polynomial in $|\mathcal{B}|$. If B_Θ grows with dimension, this remains an explicit problem-dependent factor and is not hidden as a logarithm. The fully rational limit $\eta \to \infty$ is not covered by this simplification.

Remark C.38 (Sample-size conditions and rate simplification). The explicit condition (G.37) is the deterministic sufficient condition used by the OVI-FSO radius-absorption proof. Independently, Theorem C.46 proves that its linear dependence on T is unavoidable, up to logarithmic factors, for uniform expected $\widetilde{O}(\sqrt{T})$ regret over $\mathfrak{G}_{\mathrm{FSO}}$. The theorem does not establish necessity of the displayed H, $|\mathcal{Z}|$, dimension, or regularization factors, and it does not apply to perfect-recall controllers. The alternative $R_\Sigma = 0$ or $N\eta^2 R_\Sigma \ge \lambda_\theta$ is used only to simplify the exact trace factor $L_{T,N}$.

Remark C.39 (Finite-N interpretation of the FSO decomposition). For every pair (T, N) satisfying the hypotheses of Theorem 4.1, dividing (4.1) by T gives a finite-horizon average-regret bound with current-query and batch-mismatch contribution $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})H\bar{\epsilon}_N$. The dependence of B_T^{FSO} and C_{FS} must be retained when interpreting this contribution. For fixed N and $L_\phi > 0$, if the regularization parameter is fixed or uniformly bounded above, $q_T = \sqrt{T}\,\bar{\epsilon}_N c_q$ eventually violates the radius-absorption condition. Thus the displayed upper bound does not establish an asymptotic fixed-N approximation floor or a necessary sample-size growth rate for each individual model. Separately, Theorem C.46 rules out uniform expected no-regret over the stated FSO commitment-rule class at fixed N.

Remark C.40 (Finite-rank dimension regimes). Under the hypotheses of Corollary 4.3, the complete complexity is $D_L = 1 + d_L + d_L^2 + d_F^2$. If $\sqrt{d_L} \ll d_F \lesssim d_L$, the follower first-order term can dominate the leader term; if $d_F \gg d_L$, the d_F^2 contribution also enlarges D_L. These regimes are described by the exact bound (4.4).

Table 8: Finite-rank regret comparison under the complete deterministic-cover analysis. Here $D_L = 1 + d_L + d_L^2 + d_F^2$ is recalled from Section 4.1.

Setting (with QR)	Reference	Regret	Conditions
1 leader + 1 follower (linear Markov game)	Chen et al. (2023a)	$\widetilde{O}(d^{3/2}H^2\sqrt{T} + d^2H^2)$	Theorem 6.1; fixed response constants
Stackelberg MFG (CO, finite rank)	Theorem C.10	$\widetilde{O}\big(H^2\sqrt{Td_LD_L} + \sqrt{d_F}\,H^2\sqrt{TD_L} + d_FH^2\sqrt{D_L}\big)$	exact population observation
Stackelberg MFG (FSO, finite rank)	Corollary 4.3	$\widetilde{O}\big(H^2\sqrt{Td_LD_L} + d_FH^2\sqrt{TD_L} + d_F^2H^2\sqrt{D_L} + H^2T\sqrt{D_L}\sqrt{\lvert\mathcal{Z}\rvert/N}\big)$	corollary regularizer and radius conditions
Stackelberg MFG (FSO, data abundant)	Remark C.33	$\widetilde{O}(d_L^{3/2}H^2\sqrt{T})$	$d_F \lesssim \sqrt{d_L}$, $T \gtrsim d_F^2$, $N = \widetilde{\Omega}(T\lvert\mathcal{Z}\rvert)$, plus corollary conditions
Stackelberg MFG (LB, FSO)	Theorem C.46	$\Omega(T/\sqrt{N})$	fixed-dimensional family; rules in $\mathfrak{G}_{\mathrm{FSO}}$
Stackelberg MFG (LB, CO)	Theorem C.44	$\Omega(d_L H\sqrt{T})$	$\sqrt{T}$ necessity only

Under the respective hypotheses of the finite-rank CO and FSO bounds, setting $d_L = d_F$, $T \gtrsim d_F^2$, and $N = \widetilde{\Omega}(T|\mathcal{Z}|)$ gives $\widetilde{O}(d_F^{3/2}H^2\sqrt{T})$ for CO and $\widetilde{O}(d_F^2H^2\sqrt{T})$ for FSO, with the other model parameters fixed. Identifying d_F with the common feature dimension in Chen et al. (2023a), the CO bound matches their leading dimension dependence, while the current FSO bound has an additional factor $\sqrt{d_F}$.

Remark C.41 (Identity-form benchmark-gap decomposition). The FSO regret (C.34) compares OVI-FSO against the Stackelberg benchmark policy π^* at the true mean field. Here π^* is the benchmark defined in Definition C.1: when $\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$ it is an unrestricted benchmark, and otherwise it is a benchmark over Π. The quantity $\Delta_{\Pi_{\mathrm{loc}}}$ directly measures the gap to the unrestricted local-simplex benchmark, so an unrestricted comparison adds $T\Delta_{\Pi_{\mathrm{loc}}}$ without attributing this gap to the planning assumption. With

$$J^\star_{\mathrm{FSO}}(N) := \sup_{g \in \mathfrak{G}_{\mathrm{FSO}}} J_{\mathrm{FSO}}(g),$$

one has the algebraic identity

$$\mathrm{Reg}_{\mathrm{FSO}}(T) = \underbrace{\sum_{t=1}^{T}\big[J^\star_{\mathrm{FSO}}(N) - J_{\mathrm{FSO}}(g^t)\big]}_{\text{learning regret within the FSO commitment-rule class}} + \underbrace{T\big[J(\pi^*) - J^\star_{\mathrm{FSO}}(N)\big]}_{\text{FSO benchmark gap}}. \tag{C.43}$$

This identity holds by definition and is purely interpretive: the proof of Theorem 4.1 bounds $\mathrm{Reg}_{\mathrm{FSO}}(T)$ directly against $J(\pi^*)$ and does not pass through $J^\star_{\mathrm{FSO}}(N)$.

(i) For an optimal exact-population benchmark over the same local-policy domain, the benchmark gap is nonnegative. Conditional on the exact system state (s, μ), an FSO commitment rule induces a distribution over admissible local policies, and the expectation of their Bellman objectives is at most the pointwise supremum over those policies. Backward induction therefore gives $J_{\mathrm{FSO}}(g) \le J(\pi^*)$ for every FSO commitment rule g, without requiring linearity in the local policy or an enlarged CO policy class. For an $\varepsilon_{\mathrm{bench}}$-optimal benchmark, the same argument gives $J(\pi^*) - J^\star_{\mathrm{FSO}}(N) \ge -\varepsilon_{\mathrm{bench}}$.

(ii) The finite-sample residual $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})HT\bar{\epsilon}_N$ in (4.1) may be heuristically associated with finite-sample observation effects in the data-poor regime, but it is *not* formally identified with the benchmark gap; a formal quantitative bound on the benchmark gap would additionally require either a policy-Lipschitz assumption on π^* (explicitly avoided in this paper) or a separate value-level comparison argument that we defer to future work. The decomposition (C.43) serves only as an interpretive lens.

C.12 Lower-bound proof details and open questions

The following model-class identifiability condition ensures that the map $\theta \mapsto \big(\nu^{\delta_a,\theta}(b \mid x; s)\big)_{x\in\mathcal{X},\, b\in\mathcal{B}}$ is injective for each fixed (s, a).

Assumption C.42 (Identifiability). *The follower-feature dimension satisfies* $d_F \le |\mathcal{X}|(|\mathcal{B}|-1)$*, and the static feature-difference covariance matrix* $\Lambda_{\text{static}} := \sum_{x,b\neq b'}(\varphi(s,a,x,b)-\varphi(s,a,x,b'))(\varphi(s,a,x,b)-\varphi(s,a,x,b'))^\top$ *is strictly positive definite uniformly over* (s,a)*.* [1]

The lower-bound constructions also satisfy the following Lipschitz regularity condition of Φ_h and $\mathcal{T}_{L,h}$.

Assumption C.43 (Lipschitz continuity of mean-field dynamics). *There exist finite constants* $L_\Phi, L_\mathcal{T} \ge 0$ *such that, uniformly over* $h \in [H]$*, for every fixed* (s,a) *and all* $(\mu,\nu),(\widehat{\mu},\widehat{\nu})$*,*

$$\big\|\Phi_h(s,a,\mu,\nu)-\Phi_h(s,a,\widehat{\mu},\widehat{\nu})\big\|_1 \le L_\Phi\big(\|\mu-\widehat{\mu}\|_1+\|\nu-\widehat{\nu}\|_1\big),$$
$$\big\|\mathcal{T}_{L,h}(\cdot\mid s,a,\mu,\nu)-\mathcal{T}_{L,h}(\cdot\mid s,a,\widehat{\mu},\widehat{\nu})\big\|_1 \le L_\mathcal{T}\big(\|\mu-\widehat{\mu}\|_1+\|\nu-\widehat{\nu}\|_1\big).$$

C.12.1 Complete lower-bound statements and the common minimax class

This section complements the upper bounds with two lower bounds that isolate distinct sources of difficulty. Both lower bounds concern expected regret measured against the optimal CO value.

First, we construct a finite-rank linear-MFG family that reduces leader-side learning under CO to a hypercube linear bandit and yields an $\Omega(d_L H\sqrt{T})$ minimax lower bound. Because CO supplies at least as much population information as FSO, this leader-side learning hardness also applies to the less informative FSO setting. The construction makes the follower side trivial, so its regret is caused by leader exploration rather than follower-parameter estimation or finite-sample plug-in error.

Second, under the FSO setting, a fixed-dimensional construction gives an $\Omega(T/\sqrt{N})$ lower bound for the FSO commitment-rule class $\mathfrak{G}_{\text{FSO}}$ in (C.33), even when the reward, transition, and response primitives are revealed. It establishes the necessity of the linear dependence $N = \widetilde{\Omega}(T)$ for a uniform expected $\widetilde{O}(\sqrt{T})$ guarantee within this FSO commitment-rule class.

Finally, we retain the individual parameter ranges of the two results and then combine them into a joint minimax lower bound on a common model class. Together, these results establish the worst-case T exponent of the FSO sample requirement and the unavoidable $\sqrt{T}$ leader-learning dependence. The lower-bound constructions also satisfy the identifiability and mean-field-dynamics regularity conditions in Assumptions C.42 and C.43, whose full statements are collected in Section C.12.

Theorem C.44 (Leader-side minimax lower bound). *For any* $d_L \ge 5$*,* $H \ge 2$*, and* $T \ge (d_L-3)^2$*, there exists a family* $\mathfrak{M}_{\text{LB}}$ *of linear MFGs in the sense of Definition 2.6 satisfying Assumptions 2.1, 2.2, 2.4, 2.5, C.14, C.42 and C.43. Every model in this family has leader-side feature dimension* d_L*, leader state-space size* $|\mathcal{S}|=4$*, leader action-space size* $|\mathcal{A}|=2^{d_L-3}$*, deterministic initial system state* $\rho_0=\delta_{(s_1,\mu_0)}$*, and a decoupled trivial follower side (*$|\mathcal{X}|=|\mathcal{B}|=2$*, follower-feature dimension* $d_F=1$*, and* $\theta_h^*=0$ *for every* $h\in[H]$*). For every learning algorithm* $\mathfrak{A}$ *under CO,*

$$\sup_{\mathcal{M}\in\mathfrak{M}_{\text{LB}}} \mathbb{E}_\mathcal{M}[\text{Reg}_\mathfrak{A}(T)] \ge c_{\text{LB}} d_L H\sqrt{T}, \tag{C.44}$$

where $c_{\text{LB}} > 0$ *is an absolute constant.*

The same bound holds under FSO for every per-step sample size $N \ge 1$ *under the state-sampling and conditional follower-action sampling laws specified in Assumption 2.3. For this lower bound, the learner may use any possibly randomized, model-uniform causal commitment rule based on the full observable history.*

Remark C.45 (Role of the hypercube action set). The choice $|\mathcal{A}|=2^{d_L-3}$ corresponds to the vertex set of the hypercube used in the linear-bandit lower-bound construction of Lattimore & Szepesvári (2020, Chapter 24) and exposes the d_L-3 hard leader directions simultaneously. The action set is finite, so the planning oracle in Assumption C.14 is satisfied. The follower-side joint space has $|\mathcal{Z}|=|\mathcal{X}||\mathcal{B}|=4$, independently of d_L.

The construction trivializes the follower side and isolates the difficulty of learning the leader transition model. It proves the stated $\Omega(d_L H\sqrt{T})$ lower bound for the CO problem and matches the upper

[1] This condition restricts the lower-bound model class and is verified by the constructions for Theorems C.44 and C.46; it is not required by the upper-bound proofs.

bound in its $\sqrt{T}$ dependence up to logarithmic factors. For this $d_F = 1$ family, comparison with the current complete-cover finite-rank CO upper bound in (4.3) leaves a factor $\sqrt{d_L}$ in dimension and a factor H in the horizon, up to logarithmic factors. The leader-side lower bound is independent of N and does not certify optimality of the finite-sample plug-in residual.

Theorem C.46 (FSO sample-size lower bound)**.** *For every integer $N \geq 1$, there exists a $(19, 1)$-linear leader–follower MFG with $H = 3$, $|\mathcal{S}| = 4$, and $|\mathcal{A}| = |\mathcal{X}| = |\mathcal{B}| = 2$ satisfying Assumptions 2.1 to 2.5, C.14, C.42 and C.43. The CO benchmark is attained by a policy in Π.*

The conclusion continues to hold if the learner is given the reward, transition, and response primitives, including θ_h^*, in advance. The realized population distributions are not observed; the learner receives only the FSO batches prescribed by Assumption 2.3. Revealing the fixed common initial population μ_0 would not change the conclusion.

For every possibly randomized learning algorithm whose episode-t commitment rule is selected before episode t from completed observable episodes and satisfies $g^t \in \mathfrak{G}_{\mathrm{FSO}}$, one has, for every $T \geq 1$,

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{FSO}}(T)] \geq c_{\mathrm{samp}} \frac{T}{\sqrt{N}}, \qquad c_{\mathrm{samp}} := \frac{1}{16}\left(1 - \frac{1}{\sqrt{8}}\right). \tag{C.45}$$

Consequently, a uniform expected $\widetilde{O}(\sqrt{T})$ regret guarantee over this FSO commitment-rule class requires $N = \widetilde{\Omega}(T)$.

Remark C.47 (Scope of the FSO lower bound). Theorem C.46 applies to every statistical procedure whose executed controller remains in $\mathfrak{G}_{\mathrm{FSO}}$, including plug-in, debiased, orthogonalized, cross-fitted, and randomized procedures. The theorem identifies the sharp T exponent of the sample requirement for the FSO commitment-rule class; the sharp dependence on H, $|\mathcal{Z}|$, d_F, and d_L remains open.

A common model class. For integers $d_L \geq 19$ and $H \geq 3$, let $\mathfrak{M}_{\mathrm{j}}(d_L, H)$ consist of the $(d_L, 1)$-linear MFGs satisfying the structural, sampling, and planning assumptions listed in Theorem C.46, with $\mathcal{S} = \{s_1, s_2^+, s_2^-, s_3\}$, $\mathcal{A} = \{-1, +1\}^{d_L - 3}$, $\mathcal{X} = \{0, 1\}$, $\mathcal{B} = \{-1, +1\}$, and $\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$. Each model has a deterministic initial system state with leader component s_1, an attained CO benchmark, and the finite-rank norm bounds (H.9). Here d_L is the ambient representation dimension, so zero-padded feature representations are allowed. The feature maps are supplied to the learner as known inputs and may vary between members of the class. Let $\mathfrak{A}_{\mathrm{FSO}}$ comprise all possibly randomized, model-uniform learning algorithms under the original FSO observation protocol that use these known inputs and completed observable episodes to select $g^t \in \mathfrak{G}_{\mathrm{FSO}}$ before episode t, with g^t fixed throughout that episode.

The resulting joint minimax guarantee is stated in Corollary 5.1. Its absolute constant can be taken as $c_{\mathrm{j}} := \frac{1}{2}\min\{c_{\mathrm{LB}}, c_{\mathrm{samp}}\}$, using the constants in Theorems C.44 and C.46.

Interpretation. The joint conclusion follows by taking the worst case over a class containing both hard constructions; the two component bounds may be witnessed by different models. The leader-side bound already holds under CO, whereas the finite-sample bound persists even with known model primitives for the FSO commitment-rule class $\mathfrak{G}_{\mathrm{FSO}}$. Comparison with the complete-cover finite-rank CO upper bound establishes the optimal $\sqrt{T}$ temporal dependence for this family of CO instances, up to logarithmic factors. Under the conditions of Corollary 4.3 and with the remaining model parameters fixed, the necessary scaling $N = \widetilde{\Omega}(T)$ for a uniform expected $\widetilde{O}(\sqrt{T})$ regret guarantee in $\mathfrak{G}_{\mathrm{FSO}}$ is sharp in its T exponent. These results determine the T dependence for the stated classes; matching dependence on all dimension, horizon, or local-space factors remains open. Closing the remaining upper–lower gaps calls for sharper uncertainty analysis or stronger lower-bound constructions.

Proof roadmap for Theorem C.44. We reduce to the hypercube-packing linear bandit of Lattimore & Szepesvári (2020), embedded as a stochastic leader-side transition with deterministic rewards respecting Definition 2.6 strictly. The reduction proceeds via Le Cam's two-point method (Tsybakov, 2009) combined with the divergence decomposition technique of Lattimore & Szepesvári (2020): a hypercube of $2^{d_L - 3}$ leader-side transition-parameter vectors is packed into the unit ball, and any algorithm's expected regret is lower-bounded by transferring the bandit lower bound across the packing via a KL-divergence chain rule. The detailed proof of Theorem C.44 is given in Section F.3.

Matching lower bounds for follower-parameter learning, for the complete dimension/horizon/local-space dependence, and for unrestricted perfect-recall FSO controllers remain open. The finite-sample lower bound in Theorem C.46 resolves only the T exponent for the FSO commitment-rule class.

Remark C.48 (Leader-side dimension and horizon gaps). The lower bound $\Omega(d_L H\sqrt{T})$ matches the $\sqrt{T}$ dependence of the complete-cover upper bound. Where both CO bounds apply with $d_F \lesssim \sqrt{d_L}$ and $T \gtrsim d_F^2$, the upper bound (4.3) retains factors $\sqrt{d_L}$ and H in dimension and horizon, respectively, up to logarithmic factors. Variance-aware refinements may improve the horizon dependence, but such a refinement for the present plug-in analysis remains open.

C.13 Modeling rationale and auxiliary discussions

This subsection collects auxiliary discussions on the modeling choices used in the main analysis. The topics are: (a) finite local spaces and linear representation, and (b) the random-sampling paradigm versus a fixed-cohort alternative. It also includes a discussion of Assumption 2.1. The final two subsubsections discuss FSO sample complexity, recovery of the CO rate under the stated conditions, comparison with Chen et al. (2023a), and the computational and statistical limitations of the main theorem. For the related ecological-inference question of whether separate state and action marginals suffice, see Section F.3.1.

The following discussion distinguishes the bounded-feature regularity used throughout the analysis from the auxiliary geometric identifiability conditions imposed only on the lower-bound constructions.

Roles of Assumptions 2.1 and C.42. The bounded-feature condition in Assumption 2.1 normalizes the scale in both the upper and lower bound analyses, whereas Assumption C.42 supplies auxiliary geometric conditions verified by the lower-bound constructions. Specifically:

- **Boundedness and Scale Normalization:** The uniform bound on the feature vectors is a standard regularity condition for finite-sample statistical analysis. This bound provides the norm control used in the finite-sample concentration arguments, including vector martingale concentration inequalities. Furthermore, it inherently normalizes the scale of the learning problem, ensuring that the estimation error and the resulting regret bounds are dimensionally consistent and not artificially inflated by arbitrary feature scaling. Such boundedness constraints are commonly used in the linear bandit and linear RL literature (Abbasi-Yadkori et al., 2011; Jin et al., 2020).
- **Geometric Identifiability in the Lower-Bound Constructions:** For a fixed (s, a) and the pure commitment $p = \delta_a$, the log-odds satisfy

$$\log \frac{\nu^{\delta_a,\theta}(b \mid x; s)}{\nu^{\delta_a,\theta}(b' \mid x; s)} = \eta\langle \varphi(s, a, x, b) - \varphi(s, a, x, b'), \theta\rangle.$$

  Positive definiteness of Λ_{static} therefore makes θ identifiable from the full collection of these response probabilities. Since each local state contributes at most $|\mathcal{B}| - 1$ independent log-odds, this fixed-action rank condition requires $d_F \le |\mathcal{X}|(|\mathcal{B}| - 1)$. For a mixed commitment, the relevant differences are those of the averaged features φ^p, whose rank must be assessed separately. Identifiability from sampled behavior also depends on which local states are observed. These are distinctions between structural identifiability and statistical information, as in discrete-choice modeling (McFadden, 1973). Bounded parameter and feature domains and likelihood-curvature control also appear in logistic-bandit analysis (Faury et al., 2020), but the auxiliary rank condition here is verified by our lower-bound constructions and is not required by our upper-bound confidence or regret arguments.

C.13.1 Motivation and the two information regimes

Stackelberg games between a leader and a large population of strategically responsive agents arise pervasively across operations, economics, and public policy: dynamic pricing, demand-response in smart grids, congestion pricing, mechanism design, and ride-share incentives (Acemoglu & Ozdaglar, 2007; Elie et al., 2019; Maharjan et al., 2013; Bimpikis et al., 2019; Chen, 2016). When the population is large and the agents are homogeneous, the interaction is naturally modeled as a *mean-field game* (MFG) (Caines et al., 2017; Lasry & Lions, 2007). Combining the MFG paradigm

with the Stackelberg leader-follower structure yields a class of games that is both practically motivated and analytically structured (Aurell et al., 2022; Bensoussan et al., 2017). As a stylized example motivated by ride-sharing pricing models (Bimpikis et al., 2019; Chen, 2016), consider a platform that sets regional prices for a large population of drivers. In our model, each driver responds locally according to a *quantal-response* behavioral model, while the platform observes a sampled batch of drivers per pricing round and seeks to maximize fare revenue while learning their price sensitivity. This paper studies the fundamental question of how a leader can efficiently *learn* to play optimally in such a game when the parameters are unknown and she observes the population through finite samples.

Problem setup and information structure. We consider a class of finite-horizon, nonstationary episodic Stackelberg MFGs in which a leader interacts with an infinite population of homogeneous, boundedly rational followers. We model the followers as myopic and use a quantal-response model to characterize bounded rationality (Simon, 1955; McFadden, 1973; McKelvey & Palfrey, 1995; 1998). The leader is modeled as a value-maximizing planner and the solution concept is a *quantal Stackelberg equilibrium* (QSE) (Chen et al., 2023a), combining Stackelberg commitment (Başar & Olsder, 1998) with quantal-response behavior characterizing followers' bounded rationality (McKelvey & Palfrey, 1995; 1998). The leader has *incomplete information*: she knows neither the agents' reward functions nor the system's transition dynamics, and most crucially, cannot observe the followers' realized rewards. We study two observational regimes: *Complete Observation* (CO), where the leader directly observes the exact joint state-action distribution of the followers' population, and *Finite-Sample Observation* (FSO), where the leader observes state-action pairs from a finite sample of N followers per step and hence observes empirical distributions rather than any true population distribution. In particular, although the environment starts from a fixed deterministic population distribution μ_0, its numerical value is not revealed to the FSO leader.

Before episode t, the FSO leader uses only observations from episodes $1, \dots, t-1$ to select and publicly commit to the full commitment-rule tuple $g^t = \{g_h^t\}_{h=1}^H$, which is then kept fixed throughout the episode. At step h, after observing the current follower-state batch, the leader evaluates the committed rule and discloses the realized local policy $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$; this stepwise disclosure is not a new policy optimization or a revision of the episode-start commitment. CO serves as a theoretical benchmark, whereas FSO is our main focus and is both more practically realistic and technically challenging. Notation tables are provided in Section A for reference.

C.13.2 Rationale for finite local spaces and linear representation

Finite local spaces $\mathcal{X}$ and $\mathcal{B}$ place the population distributions in finite-dimensional simplices and make the softmax normalization a finite sum. For a fixed leader state-action pair, a tabular follower reward has $|\mathcal{X}||\mathcal{B}|$ entries; the linear representation uses shared features of dimension d_F to control follower-parameter estimation complexity. These features support prediction at rarely observed inputs when sufficient information is available in the relevant feature-difference directions, while the empirical population-estimation error still depends on the local space sizes.

Observed types and heterogeneous rewards. A finite observed type can be included in the local state by writing $x = (q, \tau)$, where $\tau \in \{1, \dots, K\}$. Let $\psi_\tau(s, a, q, b) \in \mathbb{R}^{d_\tau}$ be the type-specific feature, and let ι_τ inject its coordinates into $\bigoplus_{j=1}^K \mathbb{R}^{d_j}$. The representation

$$\varphi(s, a, (q, \tau), b) = \iota_\tau \psi_\tau(s, a, q, b), \qquad \theta_h^* = (\theta_{h,1}^*, \dots, \theta_{h,K}^*),$$
$$d_F = \sum_{\tau=1}^K d_\tau, \qquad \|\theta_h^*\|_2^2 = \sum_{\tau=1}^K \|\theta_{h,\tau}^*\|_2^2$$

encodes type-specific reward parameters with $\|\varphi\|_2 \le 1$ whenever $\|\psi_\tau\|_2 \le 1$. The type must be observed as part of each sampled state, and the combined parameter and true rewards must satisfy the original parameter-ball and reward bounds. The likelihood and confidence-set constructions retain their form, and the theorems apply when the enlarged model satisfies their remaining assumptions. Type complexity enters through d_F, the local-space cardinalities, and the affected model constants.

Unknown inverse temperature. The response probabilities depend on $\eta\theta_h^*$, so behavioral observations generally do not identify η and θ_h^* separately. If a known finite upper bound $\bar{\eta} \ge \eta > 0$

is available, define $\bar{\theta}_h^* = \frac{\eta}{\bar{\eta}}\theta_h^*$ and $\bar{r}_h = \frac{\eta}{\bar{\eta}} r_h$. Then $\bar{\eta}\langle \varphi^p, \bar{\theta}_h^* \rangle = \eta \langle \varphi^p, \theta_h^* \rangle$, so the response law is unchanged, while $0 \le \bar{r}_h \le 1$ and $\|\bar{\theta}_h^*\|_2 \le B_\Theta$. The learner can estimate $\bar{\theta}_h^*$ using the known inverse temperature $\bar{\eta}$ without separately recovering the original reward scale. Applying the regret theorem still requires its closure and planning conditions for the candidate response family generated using $\bar{\eta}$, with constants calibrated at that value.

C.13.3 Finite-sample observation paradigms: random sampling vs. fixed cohort

Under the one-batch paired adaptive protocol of Assumption 2.3, the follower states are i.i.d. from μ_h^t conditional on the past and the current system state (s_h^t, μ_h^t). The already committed rule is evaluated on the state batch to obtain $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$; conditional on the state batch and this realized policy, the follower actions are mutually independent with laws $\nu^{p_h^t, \theta_h^*}(\cdot \mid x_h^{t,n}; s_h^t)$. Each state is paired with its subsequent action within the step, and persistent identifiers across steps are unnecessary. This conditional structure underlies Lemma G.3 and the chronological martingale analysis.

A two-cohort alternative uses the first cohort to form $\widehat{\mu}_h^t$ and evaluate the committed rule, then draws an independent second cohort from the same current population to obtain paired state-action observations under the fixed p_h^t. Conditional on the past, the current system state, and the first cohort, the second-cohort pairs are i.i.d. Separate concentration bounds for the two empirical measures preserve the $N^{-1/2}$ sample-size dependence; the stated regret theorem applies to the one-batch protocol. Both designs retain state-action pairing, whose role is distinguished from marginal observation in Proposition F.3.

A fixed cohort need not provide fresh i.i.d. state draws from the current μ_h^t at each step. Temporal dependence alone does not invalidate the conditional action-likelihood martingale argument, but controlling the cohort's empirical state-distribution error can require additional finite-particle, mixing-time, or propagation-of-chaos arguments. Under the adopted random-sampling protocol, the empirical snapshot error scales as $O(\sqrt{|\mathcal{Z}|/N})$ up to logarithmic factors, while the leader's learning remains sequential.

C.13.4 Discussion of FSO sample complexity, recovery of CO, and comparison with Chen et al. (2023a)

Sample complexity and the price of finite observation. The exact finite-sample contribution is the fourth term in (4.4), and (G.37) is a sufficient condition for the OVI-FSO radius-absorption argument. Separately, Theorem C.46 proves an $\Omega(T/\sqrt{N})$ lower bound for every rule in $\mathfrak{G}_{\mathrm{FSO}}$, matching the temporal exponent of this condition. The lower bound does not match every structural coefficient and does not extend to perfect-recall controllers. Debiased or orthogonalized estimators therefore cannot uniformly weaken the T exponent while their executed rules remain in $\mathfrak{G}_{\mathrm{FSO}}$, although they may improve other factors or restricted subclasses.

Vanishing plug-in term and fixed-N interpretation. As $N \to \infty$, $\bar{\epsilon}_N \to 0$ and the explicit finite-sample plug-in term in (4.1) vanishes. The remaining follower-estimation terms are compared with the CO bound under the dimension and regularity conditions of Corollary 4.3; they do not disappear merely by deleting the plug-in term. For hardware-bounded constant $N \equiv N_0$, the finite-horizon average-performance interpretation in Remark C.39 applies only while the theorem's hypotheses, including radius absorption, hold. This interpretation concerns cumulative average performance rather than last-iterate convergence.

Comparison with Chen et al. (2023a). They analyze a single-follower linear Markov game setting where the leader observes b_h^t exactly; structurally, this is not the $N = 1$ instance of FSO, which still has $\widehat{\mu}$ plug-in error. Both works use Hoeffding-style sub-Gaussian self-normalized concentration on the leader side, through their Lemma C.1 invoked in Appendix D.2 and through our Lemma E.4 with G_K from Yang et al., 2020. Both routes yield an $H^2\sqrt{T}$ horizon-time scale. A Bernstein-type refinement from $H^2\sqrt{T}$ to $H^{3/2}\sqrt{T}$ would improve the horizon dependence by a factor $\sqrt{H}$; establishing such a refinement here remains open.

C.13.5 Discussion and limitations of the main theorem

Statistical price of finite-sample population observation. The OVI-FSO upper bound contains a first-order $T\bar{\epsilon}_N$ accumulation. The FSO lower bound identifies a worst-case reason that the $T/\sqrt{N}$ order cannot be improved uniformly within $\mathfrak{G}_{\mathrm{FSO}}$: in the lower-bound construction, the two population distributions induce statistically close distributions of the current batch but correspond to different optimal local policies. Debiased or cross-fitted estimators may improve the coefficients in the plug-in residual or yield sharper guarantees on restricted subclasses, but they cannot yield a uniform $N = O(T^{1-\varepsilon})$, expected-$\widetilde{O}(\sqrt{T})$ guarantee for any fixed $\varepsilon > 0$ without enlarging the controller class or strengthening the model assumptions.

Separated complexity profile and RKHS treatment. Our framework treats the potentially large or countably infinite leader state space $\mathcal{S}$ through $G_K(T, \lambda)$ and the complete value-class covering complexity, while the follower side is finite dimensional with parameter dimension d_F and finite local spaces $(\mathcal{X}, \mathcal{B})$. This separated profile avoids explicit dependence on $|\mathcal{S}|$ in the statistical complexity of leader value estimation. The auxiliary inequality $d_F \leq |\mathcal{X}|(|\mathcal{B}| - 1)$ is imposed only on the lower-bound constructions through Assumption C.42.

Computational considerations. The joint $\arg\max$ in Step 1.4 of Algorithm 3 is generally non-convex in p and is treated as a planning oracle, as in Chen et al. (2023a); Jin et al. (2020); Yang et al. (2020). The experiments use the restricted pure-action point-estimator implementation described in Section D. The formal regret guarantee in Theorem 4.1 concerns the joint scheme (3.7).

Arithmetic and storage costs. For episode t, let $n = t - 1$ and $m = Nn$ denote the numbers of stored episodes and follower pairs per step. Assume cached policy-averaged follower features φ^p, fixed historical empirical kernel inputs and leader regularizer, and unit-cost kernel evaluations. With at most I_t projected full-gradient MLE iterations, one empirical-Fisher construction and factorization, and an incremental KRR Cholesky update and target solve per step, the costs are

$$\begin{aligned} \text{Arithmetic:} \quad & O\big(H\left[n^2 + I_t m|\mathcal{B}|d_F + m|\mathcal{B}|d_F^2 + d_F^3\right]\big), \\ \text{Storage:} \quad & O\big(H\left[n^2 + m|\mathcal{B}|d_F + d_F^2\right]\big). \end{aligned}$$

Refactorizing the dense KRR Gram matrices from scratch replaces the Hn^2 arithmetic term by Hn^3. The storage bound covers feature caches and dense matrices; total runtime additionally includes feature and kernel-input construction, Bellman-target and query evaluation, and planning.

Limitations. (a) *High-rationality regime*: the polynomial-in-$|\mathcal{B}|$ simplification uses $\eta = O(\log|\mathcal{B}|)$ with fixed B_Θ (Remark C.37); as $\eta \to \infty$, the response covariance $\bar{\Sigma}^{p,\theta^*}$ can become degenerate and the softmax comparison constants can become ill-conditioned, so this limit falls outside our scope.

(b) *Sample-size N condition*: the detailed condition (G.37) is sufficient and algorithm specific, while Theorem C.46 establishes necessity of its linear T exponent only for controllers in the FSO commitment-rule class. The minimax requirement for unrestricted perfect-recall FSO controllers remains open.

(c) *H-factor gap*: the current Hoeffding-style leader-side LBE analysis leaves the horizon gap described in Remark C.48.

D Experiment

This appendix provides a controlled synthetic experiment illustrating the scaling behavior suggested by Theorem 4.1. The finite-feature experiment examines learning over episodes, training-batch effects, and estimator diagnostics. The setting is deliberately small enough to admit a dynamic-programming benchmark, so that the plots in Figures 1 and 2 serve as controlled numerical diagnostics.

D.1 Experimental setup

We construct a synthetic Stackelberg MFG and evaluate a response-aware finite-feature implementation with pure-action, point-estimator planning, as specified below.

The construction makes both the leader reward $u_h(s, a, \mathcal{P})$ and the leader-state marginal transition $\mathbb{P}_h^L(\cdot \mid s, a, \mathcal{P})$ linear in a common feature map $\phi_h^L(s, a, \mathcal{P}) \in \mathbb{R}^{d_L}$ of $\mathcal{P} = \mu \otimes \nu^{p,\theta^*} \in \Delta(\mathcal{X} \times \mathcal{B})$. Here $\mathbb{P}_h^L(s' \mid s, a, \mathcal{P}) := \mathbb{P}_h\big(\{s'\} \times \Delta(\mathcal{X}) \mid s, a, \mathcal{P}\big)$ is the leader-state marginal of the system transition kernel. Concretely, with $\phi_h^L(s, a, \mathcal{P})[1] \equiv 1$ serving as a constant feature, the construction takes the affine form $\phi_h^L(s, a, \mathcal{P}) = \mathbf{c}_\phi(s, a) + \mathbf{M}_\phi(s, a)\,\mathrm{vec}(\mathcal{P})$, $u_h(s, a, \mathcal{P}) = \langle \phi_h^L(s, a, \mathcal{P}),\, \xi^* \rangle$ and $\mathbb{P}_h^L(s' \mid s, a, \mathcal{P}) = \langle \phi_h^L(s, a, \mathcal{P}),\, \omega^*(s') \rangle$. Here $\mathbf{c}_\phi(s, a) \in \mathbb{R}^{d_L}$, $\mathbf{M}_\phi(s, a) \in \mathbb{R}^{d_L \times |\mathcal{Z}|}$, and $\mathrm{vec}(\mathcal{P}) \in \mathbb{R}^{|\mathcal{Z}|}$ is the flattened joint distribution. Because ϕ_h^L is affine in $\mathcal{P}$, transition non-negativity over the simplex $\Delta(\mathcal{X} \times \mathcal{B})$ is equivalent to non-negativity at every vertex (a one-hot in (x, b)); this is enforced deterministically at instance generation by enumerating the $|\mathcal{X}|\,|\mathcal{B}|$ vertices and rescaling the perturbation in ω^* until all vertex probabilities are non-negative with margin $\geq 10^{-4}$. The agent's plug-in feature is $\phi_h^L(s, a, \widetilde{\mathcal{P}}_h^t)$ with $\widetilde{\mathcal{P}}_h^t = \widehat{\mu}_h^t \otimes \nu^{p,\widehat{\theta}_h^t}$, where the predicted follower response uses the agent's current MLE estimate $\widehat{\theta}_h^t$; thus the follower MLE genuinely enters the leader's planning through the joint-distribution feature, in line with the response-aware optimism of Algorithm 3.

The training batch sizes are $N \in \{20, 60, 200\}$, with eight random seeds per batch size. This response-aware finite-feature experiment illustrates temporal and finite-sample trends motivated by Theorem 4.1 and corollary 4.3.

Dimensions and hyperparameters. We set $|\mathcal{S}| = 4$, $|\mathcal{A}| = 3$, $|\mathcal{X}| = 3$, $|\mathcal{B}| = 2$, $H = 3$, $d_L = 12$, $d_F = 4$, and the inverse-temperature parameter $\eta = 0.4$. The algorithm uses ridge regularization $\lambda = 1.0$ for the leader-side KRR and $\lambda_\theta = 1.0$ for the follower-side MLE, projection radius $B_\Theta = 4$, and UCB scale $\beta = 2.0$; these constants reside in a single `Config` class for reproducibility. Both the follower features $\varphi(s, a, x, b) \in \mathbb{R}^{d_F}$ and the leader joint feature $\phi_h^L(s, a, \mathcal{P}) \in \mathbb{R}^{d_L}$ are normalized at instance generation so that $\|\varphi\|_2 \leq 1$ and the maximum of $\|\phi_h^L(s, a, \mathcal{P})\|_2$ over $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\mathcal{P} \in \Delta(\mathcal{X} \times \mathcal{B})$ is bounded by a small constant; for the joint feature the maximum is attained at a vertex of $\Delta(\mathcal{X} \times \mathcal{B})$ (since ϕ_h^L is affine in $\mathcal{P}$), so the bound is verified deterministically by enumeration of the $|\mathcal{X}|\,|\mathcal{B}|$ vertices.

Subclass simplifications relative to the general algorithm. For computational tractability and code transparency, the implementation uses a restricted finite-rank variant of OVI-FSO. The theorem is stated relative to the admissible commitment class Π_{loc} in Assumption C.14 and theorem 4.1; in the experiment, that class is restricted further to pure actions and the dynamic-programming benchmark uses the same restriction.

Specifically: (i) the local policy is restricted to $p = e_a$ for $a \in \mathcal{A}$, so planning reduces to a discrete maximum over $|\mathcal{A}|$ candidates;

(ii) the data-generating instance uses a shared parameter $\theta_h^* \equiv \theta^*$ and the implementation pools observations across steps into one MLE. The pooled estimator is an experimental time-homogeneous simplification; the step-specific MLE theorem Lemma G.5 is not invoked to justify this pooling operation.

(iii) the leader reward and leader-state marginal transition are linear in the joint feature by construction, and transition nonnegativity is enforced at every vertex of $\Delta(\mathcal{X} \times \mathcal{B})$; these properties establish the marginal model but do not verify the full system-state Bellman-closure condition;

(iv) feature norms are normalized at instance generation as described above so that the bounded-feature assumption holds with explicit constants;

(v) the follower MLE is updated every episode with no sparse schedule, and the leader-side KRR is rebuilt every episode using the agent's current $\widehat{\theta}_h^t$ to recompute the joint-feature plug-in $\phi_h^L(s, a, \widehat{\mu}_h^t \otimes \nu^{e_a, \widehat{\theta}_h^t})$ on every stored trajectory. This response-aware point-estimator construction uses $\widehat{\theta}_h^t$ in leader planning through the joint-distribution feature, which distinguishes the implementation from a leader-only LSVI-UCB baseline.

These simplifications keep the experiment within a controlled finite-feature setting but define a restricted numerical diagnostic rather than a full implementation of the general planner.

Benchmark. The benchmark J_{DP} is computed by backward dynamic programming on a Cartesian grid discretization of $\Delta(\mathcal{X})$, using 91 grid points from a 13-point per-axis lattice on the $|\mathcal{X}| = 3$

simplex. It is the optimal value of the discretized pure-action benchmark problem; the approximation error relative to the continuous-population benchmark is not quantified here. All reported performance differences use J_{DP} as the benchmark. Write $g^{t,N}$ for the rule learned at episode t using training batch size N. For realized training returns R_t, the discrepancy between $\sum_{t=1}^{T}(J_{\mathrm{DP}} - R_t)$ and $\sum_{t=1}^{T}[J_{\mathrm{DP}} - J_{\mathrm{FSO},N}(g^{t,N})]$ is a martingale sum with zero expectation. The evaluation curve integrates periodic Monte-Carlo evaluations of the frozen learned rule, using the same dynamics ordering as in training, $N_{\mathrm{eval}} = 200$ followers per step, and 20 rollouts per evaluation point. Its target at episode t is $J_{\mathrm{FSO},200}(g^{t,N})$, which can differ from $J_{\mathrm{FSO},N}(g^{t,N})$ when $N \neq 200$. Monte-Carlo estimation and interpolation between evaluation points introduce further numerical approximation. We use this evaluation curve for slope estimation, so the fitted slopes describe the stated evaluation protocol.

Reproducibility. A single self-contained Python script `ovi_fso_experiment.py` (no GPU required, runtime on the order of a few minutes on a standard laptop) reproduces every figure in this section. We use $n_{\mathrm{seeds}} = 8$ independent random instantiations of the environment, where each seed determines both the sampled synthetic instance and the algorithm's exploration randomness; shaded bands in all plots denote $\pm$ one standard error of the mean (SEM) across these eight runs. The shaded bands should therefore be interpreted as across-instance-and-randomness variability, rather than as repeated runs on a single fixed instance: this protocol assesses whether the scaling pattern is typical across randomly generated instances, not the algorithmic uncertainty for any one fixed problem.

D.2 Results

We focus on three theory-linked quantities that are observable in this synthetic instance.

(R1) Observed finite-window temporal scaling. Figure 1 plots cumulative DP-benchmark performance differences from the evaluation protocol above, with $N_{\mathrm{eval}} = 200$ for every training batch size N. At the largest tested training batch size N, the empirical log–log slope over the latter half of the recorded run falls between 0.4 and 0.6, providing a finite-window comparison with the square-root reference. The dashed line in both panels is the single-parameter reference $c\sqrt{T}$ anchored at the largest-N curve at the final recorded episode.

(R2) Euclidean MLE error in the synthetic instance. The left panel of Figure 2 plots $\|\widehat{\theta} - \theta^*\|_2$ against T on logarithmic axes. The empirical curve decreases approximately as $c/\sqrt{T}$ on this instance. The formal result Lemma G.5 controls the estimator in empirical Fisher geometry and yields a Euclidean rate only under an additional excitation lower bound. Newton's method on the empirical Fisher matrix stabilizes after the initial episodes in the reported runs.

(R3) Effect of the training batch size. The finite-sample contribution in (4.1) is $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})HT\bar{\epsilon}_N$, and its finite-rank specialization is displayed in (4.4). Because $\bar{\epsilon}_N$ decreases with N, the theorem permits a larger residual at smaller follower sample sizes. The small-N curves in Figure 1 typically sit above the larger-N curves. The contribution proportional to $T\bar{\epsilon}_N$, with its displayed radius-dependent coefficient, is part of the upper bound for admissible pairs (T, N). Seed-to-seed variability remains visible because each seed samples a fresh synthetic instance and a fresh exploration trajectory.

Diagnostics. The right panel of Figure 2 reports the average leader-side parameter error $\frac{1}{H}\sum_h \|\widehat{\xi}_h - \xi^*\|_2$. This diagnostic tracks a bootstrapped action-value fit rather than the immediate leader reward parameter alone: the LSVI-UCB target $\widehat{\xi}_h$ encodes future state-value contributions through the Bellman recursion. The generally decreasing and stabilizing trend indicates stable bootstrapped fitting in this synthetic instance.

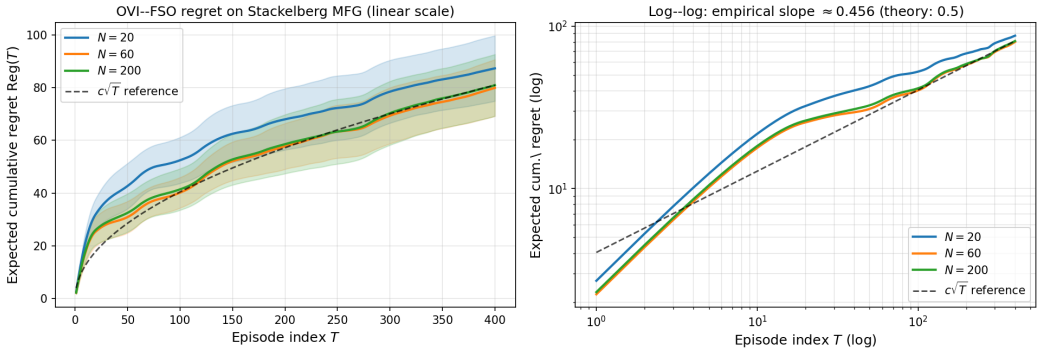

Figure 1: **Cumulative DP-benchmark performance differences in the restricted numerical subclass.** The curves use periodic evaluation with $N_{\text{eval}} = 200$ for each training batch size N. Mean (solid) $\pm$ one standard error of the mean (shaded) over $n_{\text{seeds}} = 8$ independent runs. Linear scale (*left*); log-log scale (*right*). The dashed curve is the reference $c\sqrt{T}$. At the largest tested N, the fitted slope is close to 0.5 over the reported finite window. The small-N curves typically lie above the larger-N curves.

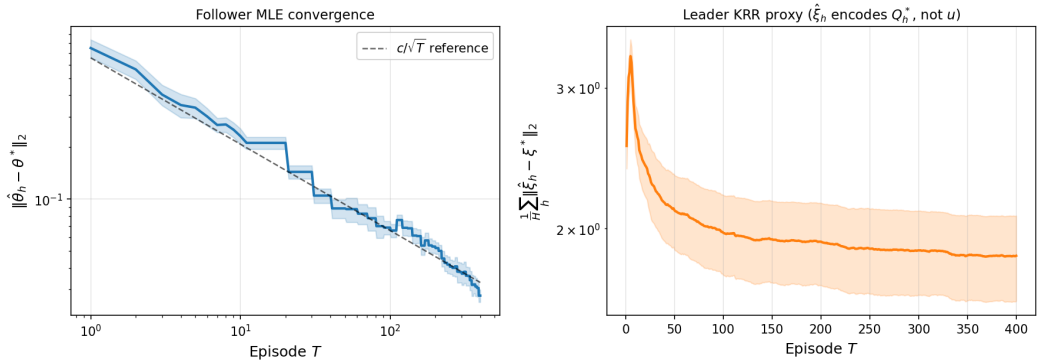

Figure 2: **Estimator diagnostics for the largest-N runs.** *Left*: Euclidean follower-MLE error $\|\widehat{\theta} - \theta^*\|_2$ on logarithmic axes, with the empirical reference $c/\sqrt{T}$. *Right*: leader KRR proxy $H^{-1}\sum_h \|\widehat{\xi}_h - \xi^*\|_2$ on a logarithmic y axis. The latter need not converge to zero because $\widehat{\xi}_h$ encodes a bootstrapped action-value target rather than the immediate reward parameter alone.

D.3 Discussion

The fitted slopes describe the recorded finite windows and can be affected by warmup, instance geometry, stochastic variation, and finite-sample shifts; they do not establish an asymptotic exponent.

Software availability. A single Python file (`ovi_fso_experiment.py`) reproducing both figures in this section, together with all hyperparameters, the DP benchmark, the algorithm, the environment, and the evaluation machinery, is provided in the supplementary material. The script runs without modification on Python 3.9+ with `numpy`, `scipy`, and `matplotlib` as the only dependencies.

E AUXILIARY LEMMAS

Remark E.1 (Norm convention). Throughout the paper, $\|\cdot\|_1$ denotes the vector ℓ_1-norm: for any signed measure ρ on a finite set $\mathcal{Z}$, $\|\rho\|_1 = \sum_{z\in\mathcal{Z}} |\rho(z)|$. In particular, for a conditional probability kernel $\nu \in (\Delta(\mathcal{B}))^{\mathcal{X}}$, we use the μ-weighted form $\sum_x \mu(x)\,\|\nu(\cdot\,|\,x)\|_1$, which equals $\|\mu\otimes\nu\|_1 = 1$. The same weighting convention applies to a conditional-kernel difference $\Delta\nu = \nu_1 - \nu_2$, for which $\|\mu\otimes\Delta\nu\|_1 = \sum_x \mu(x)\,\|\Delta\nu(\cdot\,|\,x)\|_1 \le \max_x \|\Delta\nu(\cdot\,|\,x)\|_1$. Equality in this inequality holds if and only if $\|\Delta\nu(\cdot\,|\,x)\|_1$ attains its maximum at every x with $\mu(x) > 0$. The weighted form arises naturally from the product structure of $\mu\otimes\Delta\nu$.

Lemma E.2 (Performance difference lemma for MFG, adapted from Lemma B.1 of Chen et al. (2023a)). *For any fixed policy π, let $\widetilde{\nu}$ be an estimate of the quantal response ν^π and let $\widetilde{U}$ and $\widetilde{W}$ be estimates of U^π and W^π, respectively. Use the terminal conventions $\widetilde{W}_{H+1} \equiv 0$ and $(T^{\pi,\widetilde{\nu}}_{H+1}\widetilde{U}_{H+1}) \equiv 0$. Assume that there exists a constant $L_U > 0$ such that, for every $h \in [H]$, the action-value estimate $\widetilde{U}_h(s,a,\mathcal{P})$ is L_U-Lipschitz in $\mathcal{P}$ with respect to the vector ℓ_1-norm:*

$$|\widetilde{U}_h(s,a,\mathcal{P}) - \widetilde{U}_h(s,a,\mathcal{P}')| \le L_U\,\|\mathcal{P}-\mathcal{P}'\|_1\,, \qquad \forall (s,a,\mathcal{P},\mathcal{P}'). \tag{E.1}$$

In the CO application, set $\widetilde{U}_h = \widehat{U}^t_h$ and $L_U = L_U^{\mathrm{CO}}$; in the FSO application, set $\widetilde{U}_h = \widehat{U}^t_h$ and $L_U = L_U^{\mathrm{FSO}}$. Both choices follow from Corollary F.1 and the setting-specific iterate-membership statements. Although the optimistic iterate $\widehat{U}^t_h$ need not belong to $\mathcal{U}^*$ because of the UCB bonus and truncation, it still satisfies the Lipschitz property (E.1). Then, we have the following results:

$$\begin{aligned}
&\mathbb{E}_{(s_1,\mu_1)\sim\rho_0}\big[(T^{\pi,\widetilde{\nu}}_1\widetilde{U}_1)(s_1,\mu_1)\big] - J(\pi)\\
&\quad\le \underbrace{\sum_{h=1}^H \mathbb{E}\Big[(\widetilde{U}_h - u_h)(s_h,a_h,\mathcal{P}_h) - (T^{\pi,\widetilde{\nu}}_{h+1}\widetilde{U}_{h+1})(s_{h+1},\mu_{h+1})\Big]}_{\textit{Leader's Bellman error}}\\
&\qquad + \underbrace{L_U\sum_{h=1}^H \mathbb{E}\big[\sum_{x\in\mathcal{X}} \mu_h(x)\,\|(\widetilde{\nu}_h - \nu^\pi_h)(\cdot\,|\,x;s_h,\mu_h)\|_1\big]}_{\textit{Quantal response error}},
\end{aligned}\tag{E.2}$$

$$\begin{aligned}
&\mathbb{E}_{(s_1,\mu_1)\sim\rho_0}\big[\widetilde{W}_1(s_1,\mu_1)\big] - J(\pi)\\
&\quad\le \underbrace{\sum_{h=1}^H \mathbb{E}\big[(\widetilde{U}_h - u_h)(s_h,a_h,\mathcal{P}_h) - \widetilde{W}_{h+1}(s_{h+1},\mu_{h+1})\big]}_{\textit{Leader's Bellman error}} + \underbrace{\sum_{h=1}^H \mathbb{E}\big[\widetilde{W}_h(s_h,\mu_h) - (T^{\pi,\widetilde{\nu}}_h\widetilde{U}_h)(s_h,\mu_h)\big]}_{\textit{Value mismatch error}}\\
&\qquad + \underbrace{L_U\sum_{h=1}^H \mathbb{E}\big[\sum_{x\in\mathcal{X}} \mu_h(x)\,\|(\widetilde{\nu}_h - \nu^\pi_h)(\cdot\,|\,x;s_h,\mu_h)\|_1\big]}_{\textit{Quantal response error}},
\end{aligned}\tag{E.3}$$

where the expectation is taken with respect to the randomness of the trajectory generated by (π,ν^π) under the true model M^*.

Proof of Lemma E.2. Recall that the operator $T^{\pi,\nu}_h$ is given by

$$(T^{\pi,\nu}_h U)(s,\mu) = \langle\, U(s,\cdot,\mu\otimes\nu_h(s,\mu))\,,\,\pi_h(\cdot\,|\,s,\mu)\,\rangle_{\mathcal{A}}.$$

By the definitions of U^π and W^π in (2.2), $J(\pi)$ can be written as

$$J(\pi) = \mathbb{E}_{(s_1,\mu_1)\sim\rho_0}[W^\pi_1(s_1,\mu_1)] = \mathbb{E}_{(s_1,\mu_1)\sim\rho_0}[(T^{\pi,\nu}_1 U^\pi_1)(s_1,\mu_1)],$$

where we write $\nu = \nu^\pi$ to simplify the notation.

For any $h\in[H]$, we have

$$\begin{aligned}
&\mathbb{E}\big[(T^{\pi,\widetilde{\nu}}_h\widetilde{U}_h)(s_h,\mu_h)\big] - \mathbb{E}[(T^{\pi,\nu}_h U^\pi_h)(s_h,\mu_h)]\\
&\quad= \mathbb{E}\big[(T^{\pi,\widetilde{\nu}}_h\widetilde{U}_h)(s_h,\mu_h)\big] - \mathbb{E}\big[T^{\pi,\nu}_h\widetilde{U}_h(s_h,\mu_h)\big] + \mathbb{E}\big[T^{\pi,\nu}_h\widetilde{U}_h(s_h,\mu_h)\big] - \mathbb{E}[(T^{\pi,\nu}_h U^\pi_h)(s_h,\mu_h)]\\
&\quad= \mathbb{E}\big[(T^{\pi,\widetilde{\nu}}_h - T^{\pi,\nu}_h)\widetilde{U}_h(s_h,\mu_h)\big] + \mathbb{E}\big[(T^{\pi,\nu}_h\widetilde{U}_h - T^{\pi,\nu}_h U^\pi_h)(s_h,\mu_h)\big].
\end{aligned}\tag{E.4}$$

For the second term of the RHS of (E.4), we have

$$\begin{aligned}\mathbb{E}\big[(T_h^{\pi,\nu}\widetilde{U}_h - T_h^{\pi,\nu}U_h^\pi)(s_h,\mu_h)\big] &= \mathbb{E}\big[\widetilde{U}_h(s_h,a_h,\mathcal{P}_h) - U_h^\pi(s_h,a_h,\mathcal{P}_h)\big]\\ &= \mathbb{E}\big[\widetilde{U}_h(s_h,a_h,\mathcal{P}_h) - u_h(s_h,a_h,\mathcal{P}_h) - T_{h+1}^{\pi,\nu}U_{h+1}^\pi(s_{h+1},\mu_{h+1})\big]\\ &= \mathbb{E}\big[(\widetilde{U}_h - u_h)(s_h,a_h,\mathcal{P}_h) - T_{h+1}^{\pi,\widetilde{\nu}}\widetilde{U}_{h+1}(s_{h+1},\mu_{h+1})\big]\\ &\quad + \mathbb{E}\big[T_{h+1}^{\pi,\widetilde{\nu}}\widetilde{U}_{h+1}(s_{h+1},\mu_{h+1}) - T_{h+1}^{\pi,\nu}U_{h+1}^\pi(s_{h+1},\mu_{h+1})\big],\end{aligned} \tag{E.5}$$

where the second equality follows from the Bellman equation of U^π, and the expectation $\mathbb{E}$ is taken with respect to the randomness of the trajectory generated by (π,ν^π) under the true model M^*. Combining (E.4) and (E.5), we have

$$\begin{aligned}&\mathbb{E}\big[(T_h^{\pi,\widetilde{\nu}}\widetilde{U}_h)(s_h,\mu_h)\big] - \mathbb{E}[(T_h^{\pi,\nu}U_h^\pi)(s_h,\mu_h)]\\ &\quad= \mathbb{E}\big[(T_h^{\pi,\widetilde{\nu}} - T_h^{\pi,\nu})\widetilde{U}_h(s_h,\mu_h)\big] + \mathbb{E}\big[(\widetilde{U}_h - u_h)(s_h,a_h,\mathcal{P}_h) - T_{h+1}^{\pi,\widetilde{\nu}}\widetilde{U}_{h+1}(s_{h+1},\mu_{h+1})\big]\\ &\qquad+ \mathbb{E}\big[T_{h+1}^{\pi,\widetilde{\nu}}\widetilde{U}_{h+1}(s_{h+1},\mu_{h+1})\big] - \mathbb{E}\big[T_{h+1}^{\pi,\nu}U_{h+1}^\pi(s_{h+1},\mu_{h+1})\big].\end{aligned} \tag{E.6}$$

By applying recursion to (E.6) over $h\in[H]$, we obtain that

$$\begin{aligned}&\mathbb{E}_{(s_1,\mu_1)\sim\rho_0}\big[(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big] - J(\pi) = \mathbb{E}\big[(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big] - \mathbb{E}[(T_1^{\pi,\nu}U_1^\pi)(s_1,\mu_1)]\\ &\quad= \sum_{h=1}^H \mathbb{E}\big[(T_h^{\pi,\widetilde{\nu}} - T_h^{\pi,\nu})\widetilde{U}_h(s_h,\mu_h)\big] + \sum_{h=1}^H \mathbb{E}\big[\big(\widetilde{U}_h - u_h\big)(s_h,a_h,\mathcal{P}_h) - (T_{h+1}^{\pi,\widetilde{\nu}}\widetilde{U}_{h+1})(s_{h+1},\mu_{h+1})\big].\end{aligned} \tag{E.7}$$

Finally, by the Lipschitz hypothesis (E.1) on $\widetilde{U}_h$ in $\mathcal{P}$ with envelope L_U, we have

$$\begin{aligned}&\mathbb{E}\big[(T_h^{\pi,\widetilde{\nu}} - T_h^{\pi,\nu})\widetilde{U}_h(s_h,\mu_h)\big]\\ &\quad= \mathbb{E}\left[\sum_{a_h\in\mathcal{A}}\pi_h(a_h\,|\,s_h,\mu_h)\left[\widetilde{U}_h(s_h,a_h,\mu_h\otimes\widetilde{\nu}_h(s_h,\mu_h)) - \widetilde{U}_h(s_h,a_h,\mu_h\otimes\nu_h(s_h,\mu_h))\right]\right]\\ &\quad\le \mathbb{E}\left[\sum_{a_h\in\mathcal{A}}\pi_h(a_h\,|\,s_h,\mu_h)\,L_U\,\|\mu_h\otimes\widetilde{\nu}_h(s_h,\mu_h) - \mu_h\otimes\nu_h(s_h,\mu_h)\|_1\right]\\ &\quad= L_U\,\mathbb{E}\left[\|\mu_h\otimes\widetilde{\nu}_h(s_h,\mu_h) - \mu_h\otimes\nu_h(s_h,\mu_h)\|_1\right].\end{aligned}$$

We now expand the vector ℓ_1-norm of the joint-distribution difference using the product structure $\mu_h\otimes\nu = \mu_h(x)\nu(b\,|\,x)$:

$$\begin{aligned}\|\mu_h\otimes\widetilde{\nu}_h(s_h,\mu_h) - \mu_h\otimes\nu_h(s_h,\mu_h)\|_1 &= \sum_{x\in\mathcal{X}}\sum_{b\in\mathcal{B}}|\mu_h(x)\widetilde{\nu}_h(b\,|\,x;s_h,\mu_h) - \mu_h(x)\nu_h(b\,|\,x;s_h,\mu_h)|\\ &= \sum_{x\in\mathcal{X}}\mu_h(x)\sum_{b\in\mathcal{B}}|(\widetilde{\nu}_h-\nu_h)(b\,|\,x;s_h,\mu_h)| = \sum_{x\in\mathcal{X}}\mu_h(x)\,\|(\widetilde{\nu}_h-\nu_h)(\cdot\,|\,x;s_h,\mu_h)\|_1.\end{aligned}$$

This is the natural μ_h-weighted form of the joint-distribution error. Therefore,

$$\mathbb{E}\big[(T_h^{\pi,\widetilde{\nu}} - T_h^{\pi,\nu})\widetilde{U}_h(s_h,\mu_h)\big] \le L_U\mathbb{E}\left[\sum_{x\in\mathcal{X}}\mu_h(x)\,\|(\widetilde{\nu}_h-\nu_h)(\cdot\,|\,x;s_h,\mu_h)\|_1\right]. \tag{E.8}$$

Combining (E.7) and (E.8), we have established (E.2).

The decomposition uses a joint-distribution Lipschitz constant for the action-value estimate rather than only its supremum-norm bound. For the algorithmic applications, the relevant Lipschitz constant is L_U^{CO} or L_U^{FSO} as specified by the observation model. The complete deterministic-cover analysis retains the kernel information gain and the posterior-shape entropy explicitly. It therefore does not justify replacing the setting-specific Lipschitz constant by a plain function of $G_K(T,\lambda)$.

Then, we turn to the proof of (E.3). To this end, it suffices to incorporate the value mismatch error in (E.3) into (E.2). Specifically, we have

$$\begin{aligned}
&\sum_{h=1}^{H}\mathbb{E}\big[(\widetilde{U}_h-u_h)(s_h,a_h,\mathcal{P}_h)-T_{h+1}^{\pi,\widetilde{\nu}}\widetilde{U}_{h+1}(s_{h+1},\mu_{h+1})\big]+\mathbb{E}\big[\widetilde{W}_1(s_1,\mu_1)-(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big]\\
&\quad=\sum_{h=1}^{H}\mathbb{E}\big[(\widetilde{U}_h-u_h)(s_h,a_h,\mathcal{P}_h)\big]-\sum_{h=2}^{H}\mathbb{E}\big[T_h^{\pi,\widetilde{\nu}}\widetilde{U}_h(s_h,\mu_h)\big]+\mathbb{E}\big[\widetilde{W}_1(s_1,\mu_1)-(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big]\\
&\quad=\sum_{h=1}^{H}\mathbb{E}\big[(\widetilde{U}_h-u_h)(s_h,a_h,\mathcal{P}_h)-\widetilde{W}_{h+1}(s_{h+1},\mu_{h+1})\big]+\sum_{h=2}^{H}\mathbb{E}\big[\widetilde{W}_h(s_h,\mu_h)-T_h^{\pi,\widetilde{\nu}}\widetilde{U}_h(s_h,\mu_h)\big]\\
&\qquad+\mathbb{E}\big[\widetilde{W}_1(s_1,\mu_1)-(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big]\\
&\quad=\sum_{h=1}^{H}\mathbb{E}\big[(\widetilde{U}_h-u_h)(s_h,a_h,\mathcal{P}_h)-\widetilde{W}_{h+1}(s_{h+1},\mu_{h+1})\big]+\sum_{h=1}^{H}\mathbb{E}\big[\widetilde{W}_h(s_h,\mu_h)-T_h^{\pi,\widetilde{\nu}}\widetilde{U}_h(s_h,\mu_h)\big].
\end{aligned}\tag{E.9}$$

By combining (E.2) and (E.9), we have that

$$\begin{aligned}
&\mathbb{E}\big[\widetilde{W}_1(s_1,\mu_1)\big]-J(\pi)=\mathbb{E}\big[\widetilde{W}_1(s_1,\mu_1)-(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big]+\mathbb{E}\big[(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big]-J(\pi)\\
&\quad\le\mathbb{E}\big[\widetilde{W}_1(s_1,\mu_1)-(T_1^{\pi,\widetilde{\nu}}\widetilde{U}_1)(s_1,\mu_1)\big]+\sum_{h=1}^{H}\mathbb{E}\Big[(\widetilde{U}_h-u_h)(s_h,a_h,\mathcal{P}_h)-(T_{h+1}^{\pi,\widetilde{\nu}}\widetilde{U}_{h+1})(s_{h+1},\mu_{h+1})\Big]\\
&\qquad+L_U\sum_{h=1}^{H}\mathbb{E}\Big[\sum_{x\in\mathcal{X}}\mu_h(x)\,\big\|(\widetilde{\nu}_h-\nu_h^{\pi})(\cdot\,|\,x;s_h,\mu_h)\big\|_1\Big]\\
&\quad=\sum_{h=1}^{H}\mathbb{E}\big[(\widetilde{U}_h-u_h)(s_h,a_h,\mathcal{P}_h)-\widetilde{W}_{h+1}(s_{h+1},\mu_{h+1})\big]+\sum_{h=1}^{H}\mathbb{E}\big[\widetilde{W}_h(s_h,\mu_h)-(T_h^{\pi,\widetilde{\nu}}\widetilde{U}_h)(s_h,\mu_h)\big]\\
&\qquad+\sum_{h=1}^{H}L_U\mathbb{E}\Big[\sum_{x\in\mathcal{X}}\mu_h(x)\,\big\|(\widetilde{\nu}_h-\nu_h)(\cdot\,|\,x;s_h,\mu_h)\big\|_1\Big],
\end{aligned}$$

where the first inequality is by (E.2) and the second equality is by (E.9). Hence, we have established (E.3) and complete the proof of Lemma E.2. □

Lemma E.3 (Martingale concentration inequality derived from Freedman's inequality, Corollary G.8 in Chen et al. (2023a))**.** *Let $X_1, X_2, \ldots, X_T$ be a sequence of integrable real-valued random variables adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^{T}$. Suppose that $|X_t - \mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}]| \le R$ almost surely for every $t \in [T]$. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\sum_{t=1}^{T}X_t-\sum_{t=1}^{T}\mathbb{E}\left[X_t\,|\,\mathcal{F}_{t-1}\right]\right|\le R\sqrt{T}\log(2e\delta^{-1}).\tag{E.10}$$

Moreover, suppose that $X_t \in [0,R]$ almost surely for every $t \in [T]$. Then, with probability at least $1-\delta$,

$$\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[X_t\,|\,\mathcal{F}_{t-1}\right]-2R\log(2\delta^{-1})\le\sum_{t=1}^{T}X_t\le\frac{3}{2}\sum_{t=1}^{T}\mathbb{E}\left[X_t\,|\,\mathcal{F}_{t-1}\right]+2R\log(2\delta^{-1}).\tag{E.11}$$

Lemma E.4 (Concentration of Self-Normalized Processes in RKHS, Chowdhury & Gopalan (2017))**.** *Let $\mathbb{H}$ be an RKHS defined over an auxiliary domain $\mathcal{D} \subseteq \mathbb{R}^m$ with kernel function $K(\cdot,\cdot): \mathcal{D}\times\mathcal{D} \to \mathbb{R}$. Let $\{x_\tau\}_{\tau=1}^{\infty} \subset \mathcal{D}$ be a discrete time stochastic process that is predictable with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$, meaning that x_τ is $\mathcal{F}_{\tau-1}$ measurable for every $\tau \ge 1$. Let $\{\epsilon_\tau\}_{\tau=1}^{\infty}$ be a real-valued stochastic process such that (i) ϵ_τ is $\mathcal{F}_\tau$ measurable and (ii) ϵ_τ is zero-mean and σ-sub-Gaussian conditional on $\mathcal{F}_{\tau-1}$, i.e.,*

$$\mathbb{E}[\epsilon_\tau\,|\,\mathcal{F}_{\tau-1}]=0,\qquad\mathbb{E}\big[e^{\lambda\epsilon_\tau}\,\big|\,\mathcal{F}_{\tau-1}\big]\le e^{\lambda^2\sigma^2/2},\quad\forall\lambda\in\mathbb{R}.$$

Moreover, for any $t \geq 2$, let $E_t = (\epsilon_1, \ldots, \epsilon_{t-1})^\top \in \mathbb{R}^{t-1}$ and let $K_t \in \mathbb{R}^{(t-1)\times(t-1)}$ be the Gram matrix of $\{x_\tau\}_{\tau\in[t-1]}$. For $t = 1$, let E_1 be the empty vector and let K_1 be the 0×0 matrix. Then for any $\eta > 0$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds simultaneously for all $t \geq 1$ that

$$E_t^\top \left[(K_t + \eta \cdot I)^{-1} + I\right]^{-1} E_t \leq \sigma^2 \cdot \log\det\left[(1+\eta)\cdot I + K_t\right] + 2\sigma^2 \cdot \log(1/\delta).$$

Proof of Lemma E.4. For finite or countable $\mathcal{S}$ and $\mathcal{A}$, identify their elements with integers and embed $\Delta(\mathcal{Z})$ in $\mathbb{R}^{|\mathcal{Z}|}$. This realizes the query domain as a Borel subset of $\mathbb{R}^{|\mathcal{Z}|+2}$; transporting the kernel to this domain preserves every Gram matrix in the stated inequality. If the concentration result is stated for a kernel on all of Euclidean space, extend the transported kernel by zero whenever either argument lies outside this Borel subset. The stated inequality follows from Chowdhury & Gopalan (2017, Theorem 1). □

Lemma E.5 (Kernel elliptical potential, Abbasi-Yadkori et al., 2011; Yang et al., 2020)**.** *Let $\Lambda_1 = \lambda I_{\mathbb{H}}$ with $\lambda \geq 1$, and define*

$$\Lambda_t := \lambda I_{\mathbb{H}} + \sum_{\tau<t} \phi_\tau \otimes \phi_\tau.$$

If $\|\phi_t\|_{\mathbb{H}} \leq 1$, then

$$\sum_{t=1}^{T} \|\phi_t\|^2_{\Lambda_t^{-1}} \leq 2\log\det\left(I + \frac{K_T}{\lambda}\right). \tag{E.12}$$

Proof of Lemma E.5. Let K_t be the $t \times t$ Gram matrix of $\phi_1, \ldots, \phi_t$, and let K_0 be the empty Gram matrix with determinant one. The Schur-complement identity for finite Gram matrices gives

$$\frac{\det(I_t + K_t/\lambda)}{\det(I_{t-1} + K_{t-1}/\lambda)} = \left(1 + \|\phi_t\|^2_{\Lambda_t^{-1}}\right).$$

The assumptions imply $\|\phi_t\|^2_{\Lambda_t^{-1}} \leq 1$. Use $x \leq 2\log(1+x)$ for $x \in [0, 1]$ and sum over t. The logarithms of the finite Gram determinant ratios telescope and give the right-hand side of (E.12). □

Lemma E.6 (Elliptical Potential Lemma for matrices, Lemma G.10 in Chen et al. (2023a); cf. Proposition 1 in Carpentier et al. (2020))**.** *Suppose $U_0 = \lambda I_d$ with $\lambda > 0$, $U_t = U_{t-1} + X_t$ for $X_t \in \mathbb{S}^d_+$, and $\mathrm{Tr}\,(X_t) \leq L$ for some $L > 0$. Then*

$$\sum_{t=1}^{T} \sqrt{\mathrm{Tr}\left(U_{t-1}^{-1} X_t\right)} \leq \sqrt{\frac{LT/\lambda}{\log(1 + L/\lambda)} \cdot d\log\left(1 + \frac{LT}{\lambda d}\right)},$$

and also

$$\sum_{t=1}^{T} \mathrm{Tr}\left(U_{t-1}^{-1} X_t\right) \leq \frac{L/\lambda}{\log(1 + L/\lambda)} \cdot d\log\left(1 + \frac{LT}{\lambda d}\right).$$

Proposition E.7 (Entropy of the realizable posterior-standard-deviation family)**.** *Suppose the kernel satisfies the corresponding spectral assumptions of Yang et al. (2020, Assumption 4.3), including the eigenfunction sup-norm condition, and suppose $\lambda \in [c_1, c_2]$ for fixed $0 < c_1 \leq c_2 < \infty$. Let $\mathfrak{F}^{\mathrm{post}}_{\lambda,T}$ be the realizable posterior family in (C.11). There are constants $C_5, C_6 > 0$, depending only on the fixed spectral constants and on c_1, c_2, such that the following bounds hold for every $\delta \in (0, 1)$. In the finite-spectrum case of rank d_L,*

$$\log N_\infty\left(\delta, \mathfrak{F}^{\mathrm{post}}_{\lambda,T}\right) \leq C_5 d_L^2 \left[\log(1/\delta) + C_6\right].$$

In the exponentially decaying case,

$$\log N_\infty\left(\delta, \mathfrak{F}^{\mathrm{post}}_{\lambda,T}\right) \leq C_5 \left[\log(1/\delta) + C_6\right]^{1+2/\gamma}.$$

In the polynomially decaying case, if $\tau \in [0, 1/2)$ and $\gamma(1 - 2\tau) > 1$, then

$$\log N_\infty\left(\delta, \mathfrak{F}^{\mathrm{post}}_{\lambda,T}\right) \leq C_5 \delta^{-4/[\gamma(1-2\tau)-1]} \left[\log(1/\delta) + C_6\right].$$

Proof of Proposition E.7. Let $\{(\sigma_j, \psi_j)\}_{j\geq 1}$ be the Mercer eigenpairs and let $\{e_j\}_{j\geq 1}$ be an orthonormal basis of ℓ_2. Use the canonical feature representation

$$\phi(z) = \sum_{j\geq 1} \sqrt{\sigma_j}\psi_j(z)e_j.$$

Every realizable posterior operator satisfies

$$0 \preceq \Upsilon \preceq \lambda^{-1} I.$$

It is therefore enough to cover the ambient positive-semidefinite shape class

$$\mathfrak{F}_\lambda^{\mathrm{PSD}} := \left\{ z \longmapsto \sqrt{\langle \phi(z), \Upsilon\phi(z)\rangle} : 0 \preceq \Upsilon \preceq \lambda^{-1} I \right\}.$$

Let P_m denote the projection onto $\mathrm{span}\{e_1, \ldots, e_m\}$ and define

$$r_m := \frac{C_e}{\sqrt{\lambda}} \left(\sum_{j>m} \sigma_j^{1-2\tau} \right)^{1/2}.$$

The eigenfunction condition $\|\psi_j\|_\infty \leq C_e \sigma_j^{-\tau}$ gives

$$\sup_z \|(I - P_m)\phi(z)\|_2 \leq C_e \left(\sum_{j>m} \sigma_j^{1-2\tau} \right)^{1/2}.$$

For every admissible Υ, the reverse triangle inequality for the seminorm $\|v\|_\Upsilon = \sqrt{\langle v, \Upsilon v\rangle}$ yields

$$\sup_z |\|\phi(z)\|_\Upsilon - \|P_m\phi(z)\|_\Upsilon| \leq r_m. \tag{E.13}$$

The compression $A_m = P_m \Upsilon P_m$ belongs to the finite-dimensional order interval

$$\mathcal{I}_m := \left\{ A \in \mathbb{S}^m : 0 \preceq A \preceq \lambda^{-1} I_m \right\}.$$

Let $q_m = m(m+1)/2$. Endow $\mathbb{S}^m$ with the operator norm and choose a maximal η-separated subset of $\mathcal{I}_m$. It is an η-net with proper centers. The operator-norm balls of radius $\eta/2$ around its centers are disjoint and are contained in the operator-norm ball of radius $\lambda^{-1} + \eta/2$. Lebesgue-volume comparison in the q_m-dimensional space $\mathbb{S}^m$ gives

$$\log N\left(\eta, \mathcal{I}_m, \|\cdot\|_{\mathrm{op}}\right) \leq q_m \log\left(1 + \frac{2}{\lambda\eta}\right). \tag{E.14}$$

For $A, \widetilde{A} \in \mathcal{I}_m$ and $w_z = P_m\phi(z)$, the normalization $\sup_z K(z,z) \leq 1$ gives

$$\left| \sqrt{w_z^\top A w_z} - \sqrt{w_z^\top \widetilde{A} w_z} \right| \leq \sqrt{\left| w_z^\top (A - \widetilde{A}) w_z \right|} \leq \|A - \widetilde{A}\|_{\mathrm{op}}^{1/2}. \tag{E.15}$$

Choose the smallest $m = m_\delta$ satisfying $r_m \leq \delta/4$, and use the matrix mesh $\eta = \delta^2/16$. Equations (E.13) and (E.15) give an external $\delta/2$-cover of $\mathfrak{F}_\lambda^{\mathrm{PSD}}$. Every matrix-net center is extended by zero on $\mathrm{range}(P_m)^\perp$, so it defines an ambient positive-semidefinite shape satisfying $0 \preceq \widetilde{\Upsilon} \preceq \lambda^{-1} I$. Retain only ambient $\delta/2$-balls that intersect $\mathfrak{F}_{\lambda,T}^{\mathrm{post}}$ and choose one realizable posterior shape from each occupied ball. Every realizable posterior shape is within distance δ of the selected realizable center in its occupied ball. This produces a proper δ-cover of $\mathfrak{F}_{\lambda,T}^{\mathrm{post}}$ without increasing the number of centers.

In the finite-rank case take $m_\delta = d_L$, and (E.14) gives $O(d_L^2 \log(C/\delta))$. Under exponential decay, take

$$m_\delta = O\left([\log(C/\delta)]^{1/\gamma} \right),$$

which gives $O([\log(C/\delta)]^{1+2/\gamma})$. Under polynomial decay, set $a := \gamma(1-2\tau) - 1 > 0$. The tail sum is $O(m^{-a})$, so take

$$m_\delta = O\left(\delta^{-2/a} \right).$$

Equation (E.14) then gives $O(\delta^{-4/a} \log(C/\delta))$. This proof is the positive-semidefinite specialization of the operator-net method in Yang et al. (2020, Lemma D.3 and Appendix E.3), with every truncation and mesh scale stated explicitly. □

Lemma E.8 (Perturbation of raw confidence ellipsoids). *For $i \in \{1,2\}$, let*

$$E_i := \left\{ c_i + \sqrt{\beta_i}\Lambda_i^{-1/2} u : \|u\|_2 \le 1 \right\},$$

where $\Lambda_i \succeq \lambda_\theta I$, $0 \le \beta_i \le \bar{\beta}$, and $\lambda_\theta > 0$. Then

$$d_H(E_1, E_2) \le \|c_1 - c_2\|_2 + \lambda_\theta^{-1/2} \left| \sqrt{\beta_1} - \sqrt{\beta_2} \right| + \frac{\sqrt{\bar{\beta}}}{2\lambda_\theta^{3/2}} \|\Lambda_1 - \Lambda_2\|_{\mathrm{op}}.$$

Proof of Lemma E.8. Couple the two ellipsoids by using the same u in their radial parameterizations. The resolvent representation of the inverse square root gives

$$\|\Lambda_1^{-1/2} - \Lambda_2^{-1/2}\|_{\mathrm{op}} \le \frac{1}{2\lambda_\theta^{3/2}} \|\Lambda_1 - \Lambda_2\|_{\mathrm{op}}.$$

The triangle inequality and $\|\Lambda_i^{-1/2}\|_{\mathrm{op}} \le \lambda_\theta^{-1/2}$ prove the bound in both Hausdorff directions. □

Lemma E.9 (Entry into a norm ball along a feasible chord). *Let $\mathcal{C}$ be a convex subset of a Hilbert space. Let $c \in \mathcal{C}$ satisfy $\|c\| \le B$, and let $y \in \mathcal{C}$ satisfy $\|y\| > B$. Let z be the first point on the segment from y to c that belongs to the closed ball of radius B. Then $z \in \mathcal{C}$, $\|z\| \le B$, and*

$$\|y - z\| \le \sqrt{\|y\|^2 - B^2}.$$

Proof of Lemma E.9. The segment from y to c intersects the closed ball because its endpoint c belongs to the ball. Continuity of the norm gives a first entry parameter in $[0,1]$, and convexity gives $z \in \mathcal{C}$. Let $L := \|c - y\|$, $v := (c - y)/L$, $q := \|y\|^2 - B^2$, and $a := -\langle y, v \rangle$. The case $L = 0$ is impossible. The inequality $\|c\|^2 = \|y + Lv\|^2 \le B^2$ gives $a \ge (L^2 + q)/(2L) \ge \sqrt{q}$. The first-entry distance $r := \|y - z\|$ is the smaller root of $r^2 - 2ar + q = 0$. Consequently,

$$r = \frac{q}{a + \sqrt{a^2 - q}} \le \sqrt{q}.$$

□

Lemma E.10 (Clipping nearby convex sets by a norm ball). *Let C_1 and C_2 be nonempty closed convex sets satisfying $C_i \cap \overline{\mathbb{B}(0, B_\Theta)} \neq \varnothing$. If $d_H(C_1, C_2) \le \Delta < \infty$, then*

$$d_H\left(C_1 \cap \overline{\mathbb{B}(0, B_\Theta)}, C_2 \cap \overline{\mathbb{B}(0, B_\Theta)} \right) \le \Delta + \sqrt{2B_\Theta \Delta + \Delta^2}.$$

Proof of Lemma E.10. Fix $x \in C_1 \cap \overline{\mathbb{B}(0, B_\Theta)}$ and choose $y \in C_2$ with $\|x - y\| \le \Delta$. If $\|y\| \le B_\Theta$, use y itself. Otherwise choose any $c \in C_2 \cap \overline{\mathbb{B}(0, B_\Theta)}$ and apply Lemma E.9 to the segment from y to c. Since $\|y\| \le B_\Theta + \Delta$, the entry point z satisfies

$$\|x - z\| \le \Delta + \sqrt{(B_\Theta + \Delta)^2 - B_\Theta^2}.$$

Exchange C_1 and C_2 to obtain the reverse Hausdorff direction. □

Lemma E.11 (Transport between clipped affine feasible sets). *For $i \in \{1,2\}$, let*

$$A_i = v_i + \mathrm{range}(P_i),$$

where P_i is an orthogonal projector, $P_i v_i = 0$, and $A_i \cap \overline{\mathbb{B}(0, B_\Theta)} \neq \varnothing$. Define

$$\Delta_{\mathrm{aff}} := \|v_1 - v_2\|_2 + 2B_\Theta \|P_1 - P_2\|_{\mathrm{op}}.$$

Then

$$d_H\left(A_1 \cap \overline{\mathbb{B}(0, B_\Theta)}, A_2 \cap \overline{\mathbb{B}(0, B_\Theta)} \right) \le \Delta_{\mathrm{aff}} + \sqrt{2B_\Theta \Delta_{\mathrm{aff}} + \Delta_{\mathrm{aff}}^2}.$$

Proof of Lemma E.11. For $x \in A_1 \cap \overline{\mathbb{B}(0, B_\Theta)}$, set

$$y := v_2 + P_2(x - v_1) \in A_2.$$

Because v_1 is the closest point in A_1 to the origin, $\|v_1\|_2 \le B_\Theta$. Hence

$$\|x - y\|_2 \le \|v_1 - v_2\|_2 + \|(P_1 - P_2)(x - v_1)\|_2 \le \Delta_{\mathrm{aff}}.$$

If y lies outside the ball, apply Lemma E.9 inside A_2. Exchange the two affine sets to prove the reverse direction. □

Lemma E.12 (Setting-specific Hausdorff covering and envelope Lipschitzness for $\mathfrak{A}_{\mathrm{alg},\iota}$)**.** *Equip $\mathfrak{A}_{\mathrm{alg,CO}}$ and $\mathfrak{A}_{\mathrm{alg,FSO}}$ with the Hausdorff metric induced by the Euclidean norm on Θ. If $B_\Theta = 0$, then*

$$\log N^{\mathfrak{A}}_{\mathrm{Haus},\iota}(\delta) = 0, \qquad \delta > 0, \quad \iota \in \{\mathrm{CO}, \mathrm{FSO}\},$$

and the feasible-set covering term is omitted. In this case, the remaining envelope-Lipschitz statement is immediate. Suppose $B_\Theta > 0$. There are numerical constants $c_{\mathfrak{A}}, c'_{\mathfrak{A},\mathrm{CO}}, c'_{\mathfrak{A},\mathrm{FSO}} > 0$ such that, for every $\delta \in (0, B_\Theta]$,

$$\log N^{\mathfrak{A}}_{\mathrm{Haus},\iota}(\delta) \le c_{\mathfrak{A}} d_F^2 \log\left(c'_{\mathfrak{A},\iota}/\delta\right), \qquad \iota \in \{\mathrm{CO}, \mathrm{FSO}\}. \tag{E.16}$$

The constant $c'_{\mathfrak{A},\mathrm{CO}}$ depends on B_Θ and the affine-subspace geometry but is independent of N, M_F, and β_{F}^T. The constant $c'_{\mathfrak{A},\mathrm{FSO}}$ may depend polynomially on B_Θ, M_F, β_{F}^T, and λ_θ^{-1}. For every $U \in \mathcal{U}_{\mathrm{UCB}}(h, R, B, C_{\mathrm{off}})$ and every fixed setting ι, the map $\Theta' \mapsto V^{\iota}_{\Theta',U}$ is $L_g(R,B)$-Lipschitz from $(\mathfrak{A}_{\mathrm{alg},\iota}, d_H)$ into the supremum norm, where $L_g(R,B)$ is defined once in (C.16). The scalar offset C_{off} does not enter this Lipschitz constant. At the operational value-cover scale, set

$$\delta_{\mathrm{op}}(R,B) := B_\Theta \wedge \frac{\varepsilon_{\mathrm{cov},T}}{2(1 \vee L_g(R,B))}.$$

Then $\delta_{\mathrm{op}}(R,B) \in (0, B_\Theta]$, and (E.16) applies at this scale for both settings. The CO constant is independent of N, whereas the FSO constant may contain M_F and β_{F}^T inside the logarithm.

Proof of Lemma E.12. If $B_\Theta = 0$, every nonempty clipped feasible set equals $\{0\}$, so the claim is immediate. Fix $B_\Theta > 0$ and a target Hausdorff radius $\delta \in (0, B_\Theta]$. Set

$$\Delta_\delta := \frac{\delta^2}{16(B_\Theta + \delta)}.$$

Then

$$\Delta_\delta + \sqrt{2B_\Theta \Delta_\delta + \Delta_\delta^2} \le \delta.$$

For CO, stratify the affine family by its dimension and parameterize each stratum by the closest-to-origin offset v and orthogonal projector P. Let $q := d_F(d_F+1)/2$. Use final pairwise tolerances

$$\|v - v'\|_2 \le \frac{\Delta_\delta}{2}, \qquad \|P - P'\|_{\mathrm{op}} \le \frac{\Delta_\delta}{4B_\Theta}.$$

Lemma E.11 then gives Hausdorff error at most δ. Set the ambient cell radii to

$$r_v := \frac{\Delta_\delta}{4}, \qquad r_P := \frac{\Delta_\delta}{8B_\Theta}.$$

The Euclidean ball containing v has radius B_Θ, and the orthogonal projectors lie in the operator-norm unit ball in the q-dimensional normed space $\mathbb{S}^{d_F}$. Summing over the $d_F + 1$ rank strata, the number of occupied CO cells is at most

$$(d_F + 1)\left(1 + \frac{2B_\Theta}{r_v}\right)^{d_F}\left(1 + \frac{2}{r_P}\right)^{q}.$$

For FSO, let $\bar\beta := \beta_{\mathrm{F}}^T$. Set $r_c := \Delta_\delta/6$. When $\bar\beta = 0$, every raw ellipsoid is the singleton $\{c\}$, the Λ and β meshes are omitted, and the number of occupied cells is at most $(1 + 2B_\Theta/r_c)^{d_F}$. When $\bar\beta > 0$, use final pairwise tolerances

$$\|c - c'\|_2 \le \frac{\Delta_\delta}{3}, \qquad \left|\sqrt{\beta} - \sqrt{\beta'}\right| \le \frac{\sqrt{\lambda_\theta}\Delta_\delta}{3}, \qquad \|\Lambda - \Lambda'\|_{\mathrm{op}} \le \frac{2\lambda_\theta^{3/2}\Delta_\delta}{3\sqrt{\bar\beta}}.$$

Lemmas E.8 and E.10 then give Hausdorff error at most δ. For $\bar{\beta} > 0$, set

$$r_{\sqrt{\beta}} := \frac{\sqrt{\lambda_\theta}\Delta_\delta}{6}, \qquad r_\Lambda := \frac{\lambda_\theta^{3/2}\Delta_\delta}{3\sqrt{\bar{\beta}}}.$$

The center, symmetric-matrix and square-root-radius grids have at most

$$\left(1 + \frac{2B_\Theta}{r_c}\right)^{d_F} \left(1 + \frac{2M_F}{r_\Lambda}\right)^{q} \left(1 + \frac{2\sqrt{\bar{\beta}}}{r_{\sqrt{\beta}}}\right)$$

cells. The ambient cell radii r_c, $r_{\sqrt{\beta}}$ and r_Λ are one half of the displayed final pairwise tolerances.

Retain only occupied cells and choose one realizable nonempty feasible set from each occupied cell as the proper covering center. This occupied-cell construction preserves the positive-definite, orthogonal-projector, nonempty-intersection and confidence-radius constraints. The factor two between ambient cell radius and final pairwise tolerance accounts for the distance between an arbitrary realizable parameter and the selected realizable representative. The displayed counts and $\Delta_\delta = \delta^2/[16(B_\Theta+\delta)]$ prove (E.16) after absorbing fixed polynomial scale factors into $c'_{\mathfrak{A},\iota}$.

Finally, the softmax sensitivity bound and the joint-distribution Lipschitz bound for U give

$$|h_U(s,\mu;p,\theta_1) - h_U(s,\mu;p,\theta_2)| \le L_g(R,B)\|\theta_1 - \theta_2\|_2.$$

The scalar offset is common to the two objectives and does not change this Lipschitz constant. Taking a maximizer in one compact feasible set, transporting it through the Hausdorff correspondence, and exchanging the sets proves the envelope Lipschitz claim. □

Role of the parameter domain in the covering argument. The Euclidean-ball geometry is used in the clipping estimates underlying Lemma E.12. More general compact parameter sets $\Theta' \subseteq \overline{\mathbb{B}(0, B_\Theta)}$ can be handled if their corresponding algorithmic feasible-set family $\{\Theta' \cap A : A \text{ affine}\} \cup \{\Theta' \cap \widetilde{E}(\bar{\theta}, \Lambda, \beta)\}$ admits the same $O(d_F^2 \log(1/\epsilon))$ Hausdorff entropy bound, where $\widetilde{E}(\bar{\theta}, \Lambda, \beta) := \{\theta \in \mathbb{R}^{d_F} : \|\theta - \bar{\theta}\|_\Lambda^2 \le \beta\}$ is the raw Mahalanobis ellipsoid.

Lemma E.13 (ℓ_∞ covering of the complete action-value and value-envelope classes)**.** *Fix $\iota \in \{\mathrm{CO}, \mathrm{FSO}\}$, $h \in [H]$, $R > 0$, $B \ge 0$, $C_{\mathrm{off}} \ge 0$, and $\epsilon \in (0,1)$. When $B > 0$, the complete deterministic action-value class in (C.12) satisfies*

$$\log N_\infty^{\mathcal{U}}(\epsilon; h, R, B, C_{\mathrm{off}}) \le \log N_\infty\left(\frac{\epsilon}{4}, \mathbb{B}_{\mathbb{H}}(R)\right) + \log N\left(\frac{\epsilon\sqrt{\lambda}}{4}, [0, B]\right) + \log N_\infty\left(\frac{\epsilon}{4B}, \mathfrak{F}_{\lambda,T}^{\mathrm{post}}\right) + \log N\left(\frac{\epsilon}{4}, [0, C_{\mathrm{off}}]\right). \tag{E.17}$$

When $B = 0$, the posterior bonus vanishes identically and the corresponding decomposition is

$$\log N_\infty^{\mathcal{U}}(\epsilon; h, R, 0, C_{\mathrm{off}}) \le \log N_\infty\left(\frac{\epsilon}{2}, \mathbb{B}_{\mathbb{H}}(R)\right) + \log N\left(\frac{\epsilon}{2}, [0, C_{\mathrm{off}}]\right). \tag{E.18}$$

When $B_\Theta > 0$, set

$$\delta_\epsilon(R,B) := B_\Theta \wedge \frac{\epsilon}{2(1 \vee L_g(R,B))}$$

and the corresponding setting-specific value-envelope class satisfies

$$\log N_{\infty,\iota}(\epsilon; h, R, B, C_{\mathrm{off}}) \le \log N_\infty^{\mathcal{U}}\left(\frac{\epsilon}{2}; h, R, B, C_{\mathrm{off}}\right) + \log N_{\mathrm{Haus},\iota}^{\mathfrak{A}}\left(\delta_\epsilon(R,B)\right). \tag{E.19}$$

When $B_\Theta = 0$, the same inequality holds with the Hausdorff term defined to be zero and omitted. When $B_\Theta > 0$ and $L_g(R,B) = 0$, the Hausdorff term may also be omitted. Under the finite-spectrum assumptions of rank d_L, there exist a universal numerical constant $C \ge 1$ and finite deterministic scales $C_{\mathrm{CO}}, C_{\mathrm{FSO}} \ge 1$, independent of ϵ, R, B, and C_{off}, such that the following bound holds. For $\iota \in \{\mathrm{CO}, \mathrm{FSO}\}$, set $C_\iota = C_{\mathrm{CO}}$ when $\iota = \mathrm{CO}$ and $C_\iota = C_{\mathrm{FSO}}$ when $\iota = \mathrm{FSO}$. Then

$$\log N_{\infty,\iota}(\epsilon; h, R, B, C_{\mathrm{off}}) \le C\left(d_L + d_L^2 + d_F^2 + 1\right) \log\left(C_\iota\left(1 + R + B + C_{\mathrm{off}}\right)/\epsilon\right).$$

The scale C_{CO} is independent of the FSO batch size N, whereas C_{FSO} may contain M_F and β_{F}^T inside the logarithm. Under exponential eigenvalue decay and $B > 0$, the RKHS-ball and posterior-shape contributions have orders $[\log(R/\epsilon) + C]^{1+1/\gamma}$ and $[\log(4B/\epsilon) + C]^{1+2/\gamma}$, respectively, in addition to the scalar and setting-specific d_F^2 terms.

Under polynomial eigenvalue decay with $\tau \in [0, 1/2)$ and $\gamma(1-2\tau) > 1$, the RKHS-ball contribution has order $(R/\epsilon)^{2/[\gamma(1-2\tau)-1]}[\log(R/\epsilon) + C]$. Under the same polynomial condition and $B > 0$, the posterior-shape contribution has order $(4B/\epsilon)^{4/[\gamma(1-2\tau)-1]}[\log(4B/\epsilon) + C]$, in addition to the scalar and setting-specific d_F^2 terms. When $B = 0$, the posterior-shape contribution is absent in both spectral cases.

Proof of Lemma E.13. Since $\sup_z K(z,z) \le 1$ and $\Lambda(z_{1:m}) \succeq \lambda I_{\mathbb{H}}$, every posterior-standard-deviation shape is uniformly bounded by $\lambda^{-1/2}$. Scalar truncation is 1-Lipschitz and is therefore nonexpansive in the supremum norm. When $B > 0$, cover the RKHS base function, scalar bonus coefficient, realizable posterior shape, and scalar offset at accuracies $\epsilon/4$, $\epsilon\sqrt{\lambda}/4$, $\epsilon/(4B)$, and $\epsilon/4$, respectively. The triangle inequality proves (E.17). When $B = 0$, the bonus coefficient is fixed at zero and no posterior shape affects the function. Covering only the RKHS base function and scalar offset at accuracies $\epsilon/2$ and $\epsilon/2$ proves (E.18). Replacing U by an $\epsilon/2$ supremum-norm approximation changes each objective and its supremum by at most $\epsilon/2$. For $B_\Theta > 0$, replacing Θ' by a Hausdorff approximation of radius $\delta_\epsilon(R, B)$ changes the supremum by at most $L_g(R, B)\delta_\epsilon(R, B) \le \epsilon/2$. For $B_\Theta = 0$, every nonempty clipped feasible set is $\{0\}$ and no feasible-set cover is needed. This proves (E.19).

For positive semidefinite matrices A and B,

$$\left|\sqrt{\phi^\top A\phi} - \sqrt{\phi^\top B\phi}\right| \le \sqrt{|\phi^\top(A-B)\phi|} \le \|\phi\|_2 \|A-B\|_{\mathrm{op}}^{1/2}.$$

Therefore, a function-class mesh of size δ is obtained from an operator mesh of size $O(\delta^2)$. The space of symmetric $d_L \times d_L$ matrices has dimension $d_L(d_L+1)/2$, so the resulting log covering number is $O(d_L^2 \log(1/\delta))$. The finite-rank conclusion follows from the d_L-dimensional RKHS-ball cover, the $d_L(d_L+1)/2$-dimensional positive-semidefinite posterior-operator cover, the two scalar covers, and (C.15).

For $B > 0$, the exponential and polynomial conclusions follow by adding the RKHS-ball entropy at resolution $\epsilon/4$ to the realizable posterior-shape entropy at resolution $\epsilon/(4B)$ in Proposition E.7. For $B = 0$, the posterior-shape term is absent and only the RKHS-ball, scalar-offset, and setting-specific feasible-set contributions remain. □

F VALUE-FUNCTION PROPERTIES AND PROOFS OF MAIN RESULTS

This appendix collects additional value-function properties and proofs of the CO regret and information-theoretic lower bounds.

F.1 VALUE-FUNCTION LIPSCHITZ COROLLARY

Corollary F.1 (RKHS and complete-UCB Lipschitz envelopes)**.** (i) *Any $f \in \mathbb{H}$ with $\|f\|_{\mathbb{H}} \le R$ is RL_ϕ-Lipschitz in the joint-distribution input under the ℓ_1 norm.* (ii) *For $R > 0$ and $B \ge 0$, define*

$$L_U(R, B) := L_\phi\left(R + B\lambda^{-1/2}\right).$$

Every $U \in \mathcal{U}_{\mathrm{UCB}}(h, R, B, C_{\mathrm{off}})$ is $L_U(R, B)$-Lipschitz in the joint-distribution input. The scalar offset C_{off} does not change this Lipschitz constant. For the CO and realized FSO iterates, use the setting-specific Lipschitz constants L_U^{CO} and L_U^{FSO} defined in (C.22) and (G.35), respectively.

Proof of Corollary F.1. (i) The reproducing property and the feature Lipschitz assumption give

$$|f(z) - f(z')| \le RL_\phi \|\mathcal{P} - \mathcal{P}'\|_1.$$

(ii) For the posterior bonus, the reverse triangle inequality and $\Lambda^{-1} \preceq \lambda^{-1} I_{\mathbb{H}}$ give

$$|\beta\|\phi(z)\|_{\Lambda^{-1}} - \beta\|\phi(z')\|_{\Lambda^{-1}}| \le B\lambda^{-1/2} L_\phi \|\mathcal{P} - \mathcal{P}'\|_1.$$

The scalar offset is constant in z, and scalar truncation is nonexpansive. Adding the two Lipschitz constants proves the claim. □

Remark F.2 (Uniform scope of the complete-UCB Lipschitz bound). The conclusion of Corollary F.1 applies to every element of the four-parameter class $\mathcal{U}_{\mathrm{UCB}}(h, R, B, C_{\mathrm{off}})$. The one-argument shorthand $L_U(\mathsf{r})$ is defined once in (G.33); the corresponding radius-indexed classes are introduced immediately before (G.38).

F.2 Complete-observation main regret theorem

Proof of Theorem C.10. Let $\mathcal{E}_{\mathrm{FRE}}$ denote the event on which (H.28) holds, and let $\mathcal{E}_{\text{LBE-mart}}$ denote the CO leader bonus-martingale event used in Lemma H.6. Work on $\mathcal{E}_{\mathrm{or,CO}} \cap \mathcal{E}_{\text{LBE-mart}} \cap \mathcal{E}_{\mathrm{FRE}}$.

Step 1: Regret decomposition. On $\mathcal{E}_{\mathrm{or,CO}}$, Lemma H.5, equation (H.16), gives

$$\mathrm{Reg}_{\mathrm{CO}}(T) \le \mathrm{LBE} + L_U^{\mathrm{CO}}\mathrm{JDE}. \tag{F.1}$$

Step 2: Leader Bellman-estimation term. On $\mathcal{E}_{\mathrm{or,CO}} \cap \mathcal{E}_{\text{LBE-mart}}$, Lemma H.6 (the leader LBE lemma), equation (H.22), gives

$$\mathrm{LBE} \le 4B_T^{\mathrm{CO}} H\sqrt{TG_K(T,\lambda)} + 2B_T^{\mathrm{CO}} B_\phi H\sqrt{T}\log(2eH/\zeta)\,. \tag{F.2}$$

Step 3: Follower-response term. On $\mathcal{E}_{\mathrm{FRE}}$, Lemma H.9 (the joint-distribution-error lemma), equation (H.28), gives

$$L_U^{\mathrm{CO}}\mathrm{JDE} \lesssim L_U^{\mathrm{CO}}\eta B_\Theta H\sqrt{d_F T\log(HT/\zeta)} + L_U^{\mathrm{CO}} C_{\mathrm{JDE}} B_\Theta^2 d_F H\log(HT/\zeta). \tag{F.3}$$

Step 4: Assembly and probability. Substituting (F.2) and (F.3) into (F.1) proves (C.31). By Lemmas H.1, H.6 and H.9, each of $\mathcal{E}_{\mathrm{or,CO}}$, $\mathcal{E}_{\text{LBE-mart}}$, and $\mathcal{E}_{\mathrm{FRE}}$ has failure probability at most ζ. The union bound gives probability at least $1-3\zeta$. □

F.3 Information-theoretic lower bounds

Proof of Theorem C.44. We construct a linear MFG whose leader-side dynamics implement a $(d_L - 3)$-dimensional stochastic-transition linear bandit, amplified through $H-1$ absorbing steps of deterministic reward. The follower side is non-degenerate (so that all follower-side assumptions are verified non-vacuously) but informationally decoupled from the leader.

Step 1: Leader-side construction. Set the leader state space $\mathcal{S} = \{s_1, s_+, s_-, s_{\mathrm{end}}\}$, with s_+, s_-, s_{end} absorbing for all $h \ge 2$. Write $k := d_L - 3$ for brevity; we have $k \ge 2$ under $d_L \ge 5$. The leader action space is the hypercube $\mathcal{A} = \{-1,+1\}^k$, so $|\mathcal{A}| = 2^k = 2^{d_L-3}$. The unknown parameter is

$$\xi^* \in \Xi^* := \left\{-\frac{1}{2\sqrt{T}}, +\frac{1}{2\sqrt{T}}\right\}^k.$$

Step-1 transition (stochastic). From s_1, under any action $a \in \mathcal{A}$ and any $\mathcal{P}$:

$$\mathbb{P}_1\big(s_+ \mid s_1, a, \mathcal{P}\big) = \frac{1+\langle a, \xi^*\rangle}{2}, \quad \mathbb{P}_1\big(s_- \mid s_1, a, \mathcal{P}\big) = \frac{1-\langle a, \xi^*\rangle}{2}, \tag{F.4}$$

and $\mathbb{P}_1(s_{\mathrm{end}}|s_1,a,\mathcal{P}) = \mathbb{P}_1(s_1|s_1,a,\mathcal{P}) = 0$. Under the assumption $T \ge k^2 = (d_L-3)^2$, we have $|\langle a,\xi^*\rangle| \le k\cdot\frac{1}{2\sqrt{T}} \le 1/2$ uniformly over $a \in \mathcal{A}$, $\xi^* \in \Xi^*$, so the probabilities in (F.4) are valid.

Steps $h \ge 2$ (absorbing). For $s \in \{s_+, s_-, s_{\mathrm{end}}\}$, $\mathbb{P}_h(s|s,a,\mathcal{P}) = 1$.

Leader rewards (deterministic, consistent with Definition 2.6). For all h: $u_h(s_+,a,\mathcal{P}) = 1$, $u_h(s_-,a,\mathcal{P}) = 0$, $u_h(s_1,a,\mathcal{P}) = \frac{1}{2}$, $u_h(s_{\mathrm{end}},a,\mathcal{P}) = 0$. All rewards lie in $\{0,\frac{1}{2},1\} \subseteq [0,1]$, in compliance with the leader-reward range of Definition 2.6. Note that rewards depend *only* on s (not on a or $\mathcal{P}$), so the randomness in the leader's return comes *entirely* from the stochastic transition in (F.4), not from any noisy reward channel; the action-independent value $u_h(s_1) = \frac{1}{2}$ does not affect the bandit reduction in Step 4 (regret is invariant to action-independent constants), but it is required by the dimension-tight linear representation in Step 2.

All transitions have population component δ_{μ_0}, consistent with the stationary follower construction in Step 3. Write $\mathbb{P}_h(s', d\mu' \mid s, a, \mathcal{P}) = p_h(s' \mid s, a, \mathcal{P})\delta_{\mu_0}(d\mu')$, where $p_h(s' \mid s, a, \mathcal{P}) = \langle \phi(s, a, \mathcal{P}), v_{h,s'} \rangle$. Step 2 specifies the coefficient vectors for every $(h, s, a, \mathcal{P})$, including the unreachable cases, and verifies nonnegativity and normalization on the full domain.

Step 2: Verification of Definition 2.6 linearity. With $k = d_L - 3$ fixed in Step 1, we partition the d_L feature coordinates into *two* disjoint groups: k coordinates for the bandit action and three coordinates for the absorbing-state indicators $\{s_+, s_-, s_{\text{end}}\}$. The allocation is tight: $k + 3 = (d_L - 3) + 3 = d_L$. The additive constant $\frac{1}{2}$ in the transition (F.4) is supplied by the $\{s_+, s_-\}$ coordinates of $\phi(s_1, \cdot)$ (see below).

Partition $[d_L] = [k] \cup \{+, -, \text{end}\}$, and define $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{X} \times \mathcal{B}) \to \mathbb{R}^{d_L}$ by

$$\phi(s_1, a, \mathcal{P}) = \Big(\frac{a_1}{2\sqrt{k}}, \ldots, \frac{a_k}{2\sqrt{k}}, \frac{1}{2}, \frac{1}{2}, 0\Big)^\top, \quad \phi(s_+, a, \mathcal{P}) = (0, \ldots, 0, 1, 0, 0)^\top,$$
$$\phi(s_-, a, \mathcal{P}) = (0, \ldots, 0, 0, 1, 0)^\top, \quad \phi(s_{\text{end}}, a, \mathcal{P}) = (0, \ldots, 0, 0, 0, 1)^\top.$$

Feature-norm bound. Direct computation:

$$\|\phi(s_1, a, \mathcal{P})\|_2^2 = \sum_{i=1}^{k} \Big(\frac{a_i}{2\sqrt{k}}\Big)^2 + \Big(\frac{1}{2}\Big)^2 + \Big(\frac{1}{2}\Big)^2 + 0 = k \cdot \frac{1}{4k} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4} \le 1,$$

and $\|\phi(s_\bullet, a, \mathcal{P})\|_2 = 1$ for $\bullet \in \{+, -, \text{end}\}$ (one-hot vectors). Hence $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to B_{d_L}$ in strict compliance with the feature-norm bound in Definition 2.6.

Transition linearity (with explicit signed-measure notation). The signed-measure factor $\omega_h^*(s', d\mu')$ in (2.11) takes the form

$$\omega_h^*(s', d\mu') = v_{h,s'} \cdot \delta_{\mu_0}(d\mu'), \qquad v_{h,s'} \in \mathbb{R}^{d_L}, \tag{F.5}$$

i.e., the signed measure is supported on the singleton $\{\mu_0\}$, with vector-valued "mass" $v_{h,s'}$. The kernel evaluates as $\mathbb{P}_h(s', d\mu' | s, a, \mathcal{P}) = \langle \phi(s, a, \mathcal{P}), v_{h,s'} \rangle \cdot \delta_{\mu_0}(d\mu')$.

Step-1 vectors (uniform across all starting states). Set

$$v_{1,s_+} = \big(\sqrt{k}\,\xi_1^*, \ldots, \sqrt{k}\,\xi_k^*, 1, 0, 0\big)^\top, \quad v_{1,s_-} = \big(-\sqrt{k}\,\xi_1^*, \ldots, -\sqrt{k}\,\xi_k^*, 0, 1, 0\big)^\top,$$
$$v_{1,s_{\text{end}}} = e_{\text{end}} = (0, \ldots, 0, 0, 0, 1)^\top, \quad v_{1,s_1} = \mathbf{0}.$$

By the linearity (F.5), these uniquely determine $\mathbb{P}_1(\cdot | s, a, \mathcal{P})$ for *every* starting state $s \in \mathcal{S}$ (recall ω_h^* in Definition 2.6 depends only on (h, s'), not on s). Direct verification:

- Reachable transitions from s_1: $\mathbb{P}_1(s_+, d\mu' | s_1, a, \mathcal{P}) = \langle \phi(s_1, a, \mathcal{P}), v_{1,s_+} \rangle \cdot \delta_{\mu_0}(d\mu') = \frac{1 + \langle a, \xi^* \rangle}{2} \cdot \delta_{\mu_0}(d\mu')$, recovering (F.4); analogously for s_-. Also $\mathbb{P}_1(s_{\text{end}} | s_1, a, \mathcal{P}) = \langle \phi(s_1, a, \mathcal{P}), e_{\text{end}} \rangle = 0$ and $\mathbb{P}_1(s_1 | s_1, a, \mathcal{P}) = \langle \phi(s_1, a, \mathcal{P}), \mathbf{0} \rangle = 0$, both consistent with Step 1.
- Self-loops at unreachable absorbing states $s \in \{s_+, s_-, s_{\text{end}}\}$ at $h = 1$: $\mathbb{P}_1(s | s, a, \mathcal{P}) = \langle \phi(s, a, \mathcal{P}), v_{1,s} \rangle = \langle e_s, v_{1,s} \rangle = (v_{1,s})_{[s]} = 1$ (here $(v_{1,s})_{[s]}$ denotes the coordinate of $v_{1,s}$ corresponding to the one-hot position of $\phi(s, \cdot)$); the other three transitions out of each such s are zero, as direct computation $\langle e_s, v_{1,s'} \rangle = (v_{1,s'})_{[s]}$ shows (e.g., $\mathbb{P}_1(s_- | s_+, a, \mathcal{P}) = (v_{1,s_-})_{k+1} = 0$).

Step-h vectors ($h \ge 2$). For $s' \in \{s_+, s_-, s_{\text{end}}\}$ set $v_{h,s'} = e_{s'} \in \mathbb{R}^{d_L}$ (one-hot indicator), and set $v_{h,s_1} = \mathbf{0}$. Then $\langle \phi(s, a, \mathcal{P}), v_{h,s'} \rangle = \mathbb{1}\{s' = s\}$ for the absorbing s' (giving the self-loops of *Steps* $h \ge 2$ above), and $\langle \phi(s_1, a, \mathcal{P}), \mathbf{0} \rangle = 0$ (so s_1 has no inflow at $h \ge 2$ from any starting state; cf. *Sum-to-one verification* below).

Sum-to-one verification for unreachable cases. Since the signed-measure vectors $v_{h,s'}$ specified in *Step-1 vectors* (above) and *Step-h vectors* are uniquely determined by (h, s') (independent of the starting state s, in strict compliance with Definition 2.6 that $\phi(s, a, \mathcal{P})$ does not depend on s'), the four transition probabilities $\mathbb{P}_h(s' | s, a, \mathcal{P})$ out of any (h, s) are completely fixed. We now verify sum-to-one for the unreachable cases (which is automatic for all (h, s) given the above v's, as we now demonstrate).

- *Unreachable absorbing states at $h = 1$ ($s \in \{s_+, s_-, s_{\text{end}}\}$).* With $\phi(s, a, \mathcal{P}) = e_s$:

$$\mathbb{P}_1(s'|s, a, \mathcal{P}) \;=\; \langle e_s, v_{1,s'} \rangle \;=\; (v_{1,s'})_{[s]},$$

  where $[s] \in \{k{+}1, k{+}2, k{+}3\}$ denotes the one-hot coordinate of $\phi(s, \cdot)$. By inspection of the $v_{1,s'}$ defined above, $(v_{1,s_+})_{[s_+]} = 1$, $(v_{1,s_-})_{[s_-]} = 1$, $(v_{1,s_{\text{end}}})_{[s_{\text{end}}]} = 1$ are the only nonzero values at one-hot positions, and $(v_{1,s_1})_{[s]} = 0$ since $v_{1,s_1} = \mathbf{0}$. Hence $\mathbb{P}_1(s|s, a, \mathcal{P}) = 1$ (self-loop) and $\mathbb{P}_1(s'|s, a, \mathcal{P}) = 0$ for $s' \neq s$; the four probabilities sum to 1.

- *Unreachable s_1 at $h \geq 2$.* With $\phi(s_1, a, \mathcal{P}) = (a_1/(2\sqrt{k}), \ldots, a_k/(2\sqrt{k}), \frac{1}{2}, \frac{1}{2}, 0)$ and $v_{h,s_\bullet} = e_{s_\bullet}$ for $s_\bullet \in \{s_+, s_-, s_{\text{end}}\}$, $v_{h,s_1} = \mathbf{0}$:

$$\mathbb{P}_h(s_+|s_1, a, \mathcal{P}) = \langle \phi(s_1, a, \mathcal{P}), e_+ \rangle = \frac{1}{2}, \quad \mathbb{P}_h(s_-|s_1, a, \mathcal{P}) = \langle \phi(s_1, a, \mathcal{P}), e_- \rangle = \frac{1}{2},$$
$$\mathbb{P}_h(s_{\text{end}}|s_1, a, \mathcal{P}) = \langle \phi(s_1, a, \mathcal{P}), e_{\text{end}} \rangle = 0, \quad \mathbb{P}_h(s_1|s_1, a, \mathcal{P}) = \langle \phi(s_1, a, \mathcal{P}), \mathbf{0} \rangle = 0.$$

  Sum: $\frac{1}{2} + \frac{1}{2} + 0 + 0 = 1$. The unreachable state s_1 at $h \geq 2$ transits with equal probability $\frac{1}{2}$ each to $\{s_+, s_-\}$ — a consequence of the linearity constraint that $\phi(s, a, \mathcal{P})$ does not depend on s', so the (s_+, s_-) coordinates of $\phi(s_1, a, \mathcal{P})$ (both equal to $\frac{1}{2}$ from Step 2) fully determine these transitions. Since s_1 is left at $h = 1$ with probability 1 from any trajectory starting at (s_1, μ_0), this case is unreachable and does not affect the regret.

Integrated TV-norm bound (verified via signed combinations). Under the Dirac-at-μ_0 structure (F.5), the integrated TV norm in (C.7) evaluates as

$$\|\omega_h^*\|_{\mathbb{H},\text{TV}} \;=\; \sup_{\|f\|_\infty \leq 1} \left\| \textstyle\sum_{s' \in \mathcal{S}} \int f(s', \mu')\, \omega_h^*(s', \mathrm{d}\mu') \right\|_{\mathbb{H}} \;=\; \sup_{|f_{s'}| \leq 1\ \forall s' \in \mathcal{S}} \left\| \textstyle\sum_{s' \in \mathcal{S}} f_{s'}\, v_{h,s'} \right\|_2.$$

Hence the relevant quantity is the supremum of *signed combinations* of $\{v_{h,s'}\}$, not the individual norms $\|v_{h,s'}\|_2$.

Step 1. The signed combination evaluates as

$$\sum_{s' \in \mathcal{S}} f_{s'}\, v_{1,s'} \;=\; \big(\sqrt{k}(f_{s_+} - f_{s_-})\, \xi^*,\ f_{s_+},\ f_{s_-},\ f_{s_{\text{end}}}\big) \in \mathbb{R}^{d_L},$$

with f_{s_1}-term contributing zero (since $v_{1,s_1} = \mathbf{0}$). Maximizing over $|f_{s'}| \leq 1$ (worst case $f_{s_+} = +1$, $f_{s_-} = -1$, $|f_{s_{\text{end}}}| = 1$):

$$\left\| \sum_{s'} f_{s'}\, v_{1,s'} \right\|_2^2 \;\leq\; 4k\, \|\xi^*\|_2^2 + 1 + 1 + 1 \;=\; \frac{k^2}{T} + 3 \;\leq\; 4 \quad (\text{using } T \geq k^2),$$

so $\|\omega_1^*\|_{\mathbb{H},\text{TV}} \leq 2 \leq \sqrt{d_L}$ for $d_L \geq 4$ (with strict inequality for $d_L \geq 5$, the assumed range).

Steps $h \geq 2$. The signed combination is supported on the three one-hot absorbing coordinates: $\sum_{s'} f_{s'}\, v_{h,s'} = (0, \ldots, 0, f_{s_+}, f_{s_-}, f_{s_{\text{end}}})$. Maximizing: $\|\sum_{s'} f_{s'}\, v_{h,s'}\|_2 \leq \sqrt{3} \leq \sqrt{d_L}$ trivially for $d_L \geq 3$.

Therefore $\|\omega_h^*\|_{\mathbb{H},\text{TV}} \leq \sqrt{d_L}$ uniformly for all $h \in [H]$, matching the signed-measure norm convention of Jin et al. (2020).

Reward linearity. Define $\xi_h^{\text{rwd}} = e_+ \in \mathbb{R}^{d_L}$, i.e., the one-hot vector with value 1 in the "+" coordinate. Then $u_h(s, a, \mathcal{P}) = \langle \phi(s, a, \mathcal{P}), \xi_h^{\text{rwd}} \rangle$ gives:

$$u_h(s_+) = 1, \qquad u_h(s_-) = 0, \qquad u_h(s_{\text{end}}) = 0, \qquad u_h(s_1, a) = \frac{1}{2} \ \text{(constant in } a),$$

exactly matching Step 1's reward specification. Moreover, $\|\xi_h^{\text{rwd}}\|_2 = 1$, so $\xi_h^{\text{rwd}} \in B_{d_L}$. All required norm and range conditions of Definition 2.6 and (H.9) are satisfied: $\|\phi\|_2 \leq 1$, $\|\omega_h^*\|_{\mathbb{H},\text{TV}} \leq \sqrt{d_L}$, $\|\xi_h^{\text{rwd}}\|_2 \leq 1$, and $u_h \in [0, 1]$.

Step 3: Follower-side construction (non-degenerate but decoupled). Set $|\mathcal{X}| = |\mathcal{B}| = 2$, $\mu_0 = (1/2, 1/2)$ fixed, $\varphi(x, b) = \mathbb{1}\{x = x_1, b = b_1\}$ (so $d_F = 1$, π-independent), $\theta^* = 0$. Under $\theta^* = 0$, $\nu^{\theta^*}(\cdot|x)$ is uniform on $\mathcal{B}$, and $\mathcal{P}^* = \mu_0 \otimes \nu^{\theta^*}$ is constant across all episodes and leader policies. The population evolution map is constant, $\Phi_h(s, a, \mu, \nu) \equiv \mu_0$ for every $h \in [H]$ and every (s, a, μ, ν), so $\mu_h \equiv \mu_0$. Consequently, FSO observations carry zero information about ξ^*: the only informative channel about ξ^* is the leader's own observation of her next state $s_2 \in \{s_+, s_-\}$ after step 1.

Step 4: Value decomposition and reduction to linear bandit. Under our construction (with $u_h(s_+) = 1, u_h(s_-) = u_h(s_{\text{end}}) = 0$, and $u_h(s_1) = \frac{1}{2}$), the value function admits the decomposition

$$V_1^\pi(s_1, \mu_0; \xi^*) = \underbrace{u_1(s_1, \cdot, \cdot)}_{=1/2} + \sum_{h=2}^{H} \mathbb{E}^\pi\big[u_h(s_h, \cdot, \cdot)\big] = \frac{1}{2} + (H-1) \cdot \mathbb{P}^\pi(s_2 = s_+)$$
$$= \frac{1}{2} + (H-1) \cdot \mathbb{E}_{a\sim\pi_1}\left[\frac{1 + \langle a, \xi^*\rangle}{2}\right] = \frac{H}{2} + \frac{H-1}{2}\,\mathbb{E}_{a\sim\pi_1}[\langle a, \xi^*\rangle],$$

using (F.4) and the absorbing structure. The optimal per-episode value is $V_1^*(s_1, \mu_0; \xi^*) = \frac{H}{2} + \frac{H-1}{2} \max_{a\in\mathcal{A}} \langle a, \xi^*\rangle$. The action-independent constant $\frac{H}{2}$ cancels in the regret, leaving

$$V_1^*(s_1, \mu_0; \xi^*) - V_1^{\pi^t}(s_1, \mu_0; \xi^*) = \frac{H-1}{2} \cdot \Big(\max_a \langle a, \xi^*\rangle - \mathbb{E}_{a_1^t\sim\pi^t}\langle a_1^t, \xi^*\rangle\Big).$$

For the bandit notation used below, set $A_t := a_1^t$ and let $\mathcal{F}_{t-1}$ be the σ-field generated by the complete algorithm and environment history revealed before A_t is sampled. Then the regret over T episodes is

$$\mathrm{Reg}_{\mathfrak{A}}(T) = \frac{H-1}{2} \cdot \underbrace{\sum_{t=1}^{T}\Big(\max_a \langle a, \xi^*\rangle - \mathbb{E}[\langle A_t, \xi^*\rangle \mid \mathcal{F}_{t-1}]\Big)}_{=:\,\mathrm{Reg}_{\text{bandit}}(T;\xi^*)}. \tag{F.6}$$

The Bernoulli observation channel $\mathbb{1}\{s_2 = s_+\} \sim \mathrm{Bern}((1 + \langle a, \xi^*\rangle)/2)$ used in the transcript KL calculation (F.7) depends only on the stochastic transition (F.4) and is independent of the leader reward specification.

Step 5: A hypercube lower bound for the Bernoulli observations. We apply a hypercube testing argument directly to the transition observations in (F.4); see Lattimore & Szepesvári (2020, Chapter 24) for the general method.

Bernoulli observations under an adaptive algorithm. For a parameter $\xi \in \Xi^*$, let $\mathbb{P}_\xi^T$ denote the joint distribution of the complete algorithm–environment transcript up to episode T. Conditional on the history $\mathcal{F}_{t-1}$, the algorithm selects A_t according to the same history-dependent decision rule under every parameter, while the step-1 next-state indicator is distributed as

$$Y_t \mid (\mathcal{F}_{t-1}, A_t) \sim \mathrm{Bern}\big(p_\xi(A_t)\big), \qquad p_\xi(a) = \frac{1 + \langle a, \xi\rangle}{2} \in [1/4, 3/4].$$

For any $\xi, \xi' \in \Xi^*$,

$$p_\xi(a) - p_{\xi'}(a) = \frac{1}{2}\langle a, \xi - \xi'\rangle.$$

By the chain rule for relative entropy and the fact that the algorithm's conditional action law is the same measurable function of the observed history under both environments,

$$\mathrm{KL}\big(\mathbb{P}_\xi^T \,\big\|\, \mathbb{P}_{\xi'}^T\big) = \sum_{t=1}^{T} \mathbb{E}_\xi[\mathrm{KL}(\mathrm{Bern}(p_\xi(A_t)) \,\|\, \mathrm{Bern}(p_{\xi'}(A_t)))]$$
$$\le \frac{16}{3} \sum_{t=1}^{T} \mathbb{E}_\xi\Big[\big(p_\xi(A_t) - p_{\xi'}(A_t)\big)^2\Big] = \frac{4}{3} \sum_{t=1}^{T} \mathbb{E}_\xi\big[\langle A_t, \xi - \xi'\rangle^2\big]. \tag{F.7}$$

For Hamming-adjacent parameters $\xi = (+\varepsilon, \xi_{-i})$ and $\xi' = (-\varepsilon, \xi_{-i})$,

$$\xi - \xi' = 2\varepsilon e_i, \qquad \langle A_t, \xi - \xi' \rangle^2 = 4\varepsilon^2 A_{t,i}^2 = 4\varepsilon^2.$$

With $\varepsilon = 1/(2\sqrt{T})$,

$$\mathrm{KL}\big(\mathbb{P}_{\xi}^{T} \,\|\, \mathbb{P}_{\xi'}^{T}\big) \;\le\; \frac{4}{3} \cdot T \cdot 4\varepsilon^2 \;=\; \frac{16T\varepsilon^2}{3} \;=\; \frac{16T}{3 \cdot 4T} \;=\; \frac{4}{3}.$$

For clarity, write

$$\mathbb{P}_{+,i|\xi_{-i}}^{T} := \mathbb{P}_{(+\varepsilon,\xi_{-i})}^{T}, \qquad \mathbb{P}_{-,i|\xi_{-i}}^{T} := \mathbb{P}_{(-\varepsilon,\xi_{-i})}^{T}.$$

Bretagnolle–Huber (Lattimore & Szepesvári, 2020, Theorem 14.2) then gives, for every event A measurable with respect to the complete transcript,

$$\mathbb{P}_{+,i|\xi_{-i}}^{T}(A) + \mathbb{P}_{-,i|\xi_{-i}}^{T}(A^c) \ge \frac{1}{2} \exp\Big(-\mathrm{KL}\Big(\mathbb{P}_{+,i|\xi_{-i}}^{T} \,\Big\|\, \mathbb{P}_{-,i|\xi_{-i}}^{T}\Big)\Big) \ge \frac{1}{2} e^{-4/3}, \tag{F.8}$$

uniformly over $i \in [k]$ and $\xi_{-i} \in \{\pm\varepsilon\}^{k-1}$.

Aggregating across the k hypercube coordinates. Write $\xi = \varepsilon\sigma$ with $\sigma \in \{-1,+1\}^k$ and $\varepsilon = 1/(2\sqrt{T})$. For every $A_t \in \{-1,+1\}^k$,

$$\max_{a \in \mathcal{A}} \langle a, \varepsilon\sigma \rangle - \langle A_t, \varepsilon\sigma \rangle = 2\varepsilon \sum_{i=1}^{k} \mathbf{1}\{A_{t,i} \neq \sigma_i\}.$$

Fix $i \in [k]$, $t \in [T]$, and $\sigma_{-i} \in \{-1,+1\}^{k-1}$, and apply (F.8) to the event $\{A_{t,i} = -1\}$ with $\xi_{-i} = \varepsilon\sigma_{-i}$. This gives

$$\mathbb{P}_{+,i|\varepsilon\sigma_{-i}}^{T}(A_{t,i} = -1) + \mathbb{P}_{-,i|\varepsilon\sigma_{-i}}^{T}(A_{t,i} = +1) \ge \frac{1}{2} e^{-4/3}.$$

Averaging over the two values of σ_i and then over σ_{-i} yields

$$\frac{1}{2^k} \sum_{\sigma \in \{-1,+1\}^k} \mathbb{P}_{\varepsilon\sigma}^{T}(A_{t,i} \neq \sigma_i) \ge \frac{1}{4} e^{-4/3}.$$

Therefore,

$$\begin{aligned} \frac{1}{2^k} \sum_{\sigma \in \{-1,+1\}^k} \mathbb{E}_{\varepsilon\sigma}[\mathrm{Reg}_{\mathrm{bandit}}(T; \varepsilon\sigma)] &= 2\varepsilon \sum_{t=1}^{T} \sum_{i=1}^{k} \frac{1}{2^k} \sum_{\sigma \in \{-1,+1\}^k} \mathbb{P}_{\varepsilon\sigma}^{T}(A_{t,i} \neq \sigma_i) \\ &\ge \frac{1}{2} e^{-4/3} \varepsilon k T = \frac{1}{4} e^{-4/3} k\sqrt{T}. \end{aligned}$$

Since the supremum is at least the uniform-prior average,

$$\sup_{\xi^* \in \Xi^*} \mathbb{E}_{\xi^*}[\mathrm{Reg}_{\mathrm{bandit}}(T; \xi^*)] \ge \frac{1}{4} e^{-4/3} k\sqrt{T}. \tag{F.9}$$

Thus (F.9) holds with $c_{\mathrm{bandit}} = e^{-4/3}/4 \ge 1/64$.

Assembling the MFG-level lower bound. Substituting (F.9) into (F.6) with $k = d_L - 3$, using $H \ge 2$ (so $H - 1 \ge H/2$) and $d_L \ge 5$ (so $d_L - 3 \ge (2/5)d_L$):

$$\sup_{\mathcal{M}} \mathbb{E}[\mathrm{Reg}_{\mathfrak{A}}(T)] \ge \frac{H-1}{2} \cdot c_{\mathrm{bandit}} \cdot (d_L - 3)\sqrt{T} \ge \frac{H}{4} \cdot \frac{1}{64} \cdot \frac{2d_L}{5} \sqrt{T} = \frac{d_L H \sqrt{T}}{640}.$$

This establishes (C.44) with $c_{\mathrm{LB}} = 1/640$. Under the transition parameterization (F.4), the coefficient multiplying $\mathrm{Reg}_{\mathrm{bandit}}(T; \xi^*)$ in (F.6) is $\frac{1}{2}\sum_{h=2}^{H}\big(u_h(s_+) - u_h(s_-)\big) = (H-1)/2$.

Step 6: Verification of the model assumptions of Theorems C.10 and 4.1.

- Definition 2.6: transition and reward linearity were verified in Step 2 with leader-side dimension d_L and follower-side dimension $d_F = 1$.
- Assumption 2.2: take $\Phi_h(s, a, \mu, \nu) \equiv \mu_0$ and let $\mathcal{T}_{L,h}$ be the leader-state kernel constructed in Step 2; then (2.1) holds exactly.
- Assumption 2.4: ϕ does not depend on $\mathcal{P}$, so the required Lipschitz inequality holds with any fixed positive constant; to preserve the standing convention, take $L_\phi = 1 > 0$.
- Assumptions 2.1 and C.42: verified via the non-degenerate follower side. Specifically, $\varphi(x, b) = \mathbb{1}\{x = x_1, b = b_1\}$ (scalar, $d_F = 1$): (1) $\|\varphi\|_2 \le 1$; (2) $d_F = 1 \le |\mathcal{X}|(|\mathcal{B}| - 1) = 2$; (3) the static feature-difference covariance is $\Lambda_{\text{static}} = \sum_{x\in\mathcal{X}} \sum_{b\neq b'} (\varphi(x, b) - \varphi(x, b'))^2 = 2 \cdot (1 - 0)^2 = 2 > 0$; (4) under $\theta^* = 0$, the follower response $\nu^{p,0}$ is uniform on $\mathcal{B}$ for every $p \in \Delta(\mathcal{A})$, so the centered feature covariance is $\bar{\Sigma}^{p,0}_{s,\mu_0} = \mu_0(x_1)\ \mathbb{V}_{b\sim\text{Unif}(\mathcal{B})}[\varphi(x_1, b)] = (1/2)(1/2 \cdot (1/2)^2 + 1/2 \cdot (1/2)^2) = 1/8$.
- Assumption C.43: $\Phi_h(s, a, \mu, \nu) \equiv \mu_0$ is constant, and $\mathcal{T}_{L,h}$ in the hard instance does not depend on (μ, ν), so the condition holds with $L_\Phi = L_\mathcal{T} = 0$.
- Assumption 2.5: the finite-rank feature construction and the norm bounds verified in Step 2 satisfy the sufficient conditions (H.9); hence Proposition H.2 gives the required closure uniformly over the feasible sets used by the algorithms.
- Assumption 2.3: at each (t, h), the N states $x_h^{t,n}$ are i.i.d. from μ_0, and conditional on the state batch and on p_h^t, the actions $b_h^{t,n}$ are i.i.d. from $\nu^{p_h^t,\theta^*=0} = \text{Unif}(\mathcal{B})$ (the uniform distribution carries no information about ξ^*).
- Assumption C.14: $\mathcal{A} = \{-1, +1\}^{d_L-3}$ is finite, so $\Pi_{\text{loc}} = \Delta(\mathcal{A})$ is a compact simplex of dimension $2^{d_L-3} - 1$. For every nonempty compact feasible set $\Theta' \subseteq \Theta$, the pointwise objective $G_h^t(s, \mu; p, \theta)$ is continuous in (p, θ) on the compact set $\Delta(\mathcal{A}) \times \Theta'$, hence attains its supremum by Weierstrass's theorem. The associated argmax correspondence is nonempty compact-valued and Borel-measurable in (s, μ), so a Borel-measurable selector exists by the finite-action argument in Remark C.17.
- Benchmark attainment for the lower-bound construction: both $\mathcal{S}$ and $\mathcal{A}$ are finite, and the construction absorbs after the first informative decision. Hence the Stackelberg optimum is attained by a deterministic action $a^* \in \text{argmax}_{a\in\mathcal{A}} \langle a, \xi^* \rangle$, so the benchmark policy π^* exists within the admissible class.

Thus the constructed family satisfies the model assumptions of Theorems C.10 and 4.1, and for every admissible learning algorithm under either CO or FSO there is a member of this family for which the expected regret satisfies the lower bound in (C.44). □

Proof of Theorem C.46. Fix $N \ge 1$, and set

$$c := \frac{1}{4}, \qquad \delta := \frac{1}{8\sqrt{N}}, \qquad \bar{\mu} := \text{Ber}(1/2), \qquad \mu_\sigma := \text{Ber}(1/2 + \sigma\delta), \quad \sigma \in \{-1, +1\}.$$

Take $\mathcal{X} = \{0, 1\}$, $\mathcal{A} = \mathcal{B} = \{-1, +1\}$, and $\mathcal{S} = \{s_1, s_2^+, s_2^-, s_3\}$. Set the deterministic initial system state to $(s_0, \mu_0) = (s_1, \bar{\mu})$.

Reachable dynamics and rewards. At step 1, independently of the leader action, draw a fresh $\sigma \sim \text{Unif}\{-1, +1\}$ and set $(s_2, \mu_2) = (s_2^\sigma, \bar{\mu})$. At step 2, independently of the leader action, set

$$(s_3, \mu_3) = (s_3, \mu_\sigma).$$

The step-3 leader reward is

$$u_3(s_3, a, \mathcal{P}) = \frac{1}{2} + c\, a \left\{2 \,\text{marg}_\mathcal{X}\, \mathcal{P}(1) - 1\right\},$$

and all other reachable rewards are zero. Since $c = 1/4$, all leader rewards lie in $[0, 1]$.

Let the scalar follower feature and true parameters be $\varphi(s, a, x, b) = b$ and $\theta^*_h = 0$ for every $h \in [3]$. The true follower reward is therefore zero and the quantal response is uniform on $\mathcal{B}$.

CO benchmark. Define a Borel policy by choosing arbitrary local policies before step 3 and setting

$$\pi^*_3(\cdot \mid s_3, \mu) = \begin{cases} \delta_{+1}, & \mu(1) \geq 1/2, \\ \delta_{-1}, & \mu(1) < 1/2. \end{cases}$$

The transition and the earlier rewards are action independent, so this policy is CO optimal. At the reachable population μ_σ, it selects $a_3 = \sigma$ and obtains $J(\pi^*) = 1/2 + 2c\delta$. This policy is a single episode-start-announced state/population-contingent member of $\Pi(\Delta(\mathcal{A}))$; it does not require knowing the realized σ when the rule is announced.

Bayes lower bound for an FSO commitment rule. Fix an episode and condition on the completed observable episodes and on the episode-start rule g^t selected by the algorithm. The new σ remains uniform and independent of this conditioning. At step 3, the leader state is the common state s_3, while the observed state batch $Y = X_3^{1:N}$ satisfies

$$Y \mid \{\sigma = +1\} \sim P_+^N, \qquad Y \mid \{\sigma = -1\} \sim P_-^N, \qquad P_\pm = \mathrm{Ber}(1/2 \pm \delta).$$

A rule in $\mathfrak{G}_{\mathrm{FSO}}$ cannot use the previously observed state s_2^σ at step 3. To strengthen the learner, allow its conditional mean action to be any measurable function $d(Y) \in [-1, 1]$ of the full ordered batch, rather than only of its empirical distribution. Conditional on σ and Y, its expected step-3 reward is $1/2 + 2c\sigma\delta\, d(Y)$. Hence its equal-prior Bayes gap from the CO benchmark is at least

$$2c\delta \left\{1 - \mathrm{TV}(P_+^N, P_-^N)\right\}. \tag{F.10}$$

Indeed, the variational characterization of total variation gives

$$\sup_{|d| \leq 1} \frac{1}{2} \left(\mathbb{E}_{P_+^N} d - \mathbb{E}_{P_-^N} d\right) = \mathrm{TV}(P_+^N, P_-^N).$$

For $\delta \leq 1/4$,

$$\mathrm{kl}(P_+, P_-) = 2\delta \log \frac{1 + 2\delta}{1 - 2\delta} \leq 16\delta^2.$$

Therefore $\mathrm{KL}(P_+^N, P_-^N) \leq 16N\delta^2 = 1/4$. Pinsker's inequality yields

$$\mathrm{TV}(P_+^N, P_-^N) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(P_+^N, P_-^N)} \leq \frac{1}{\sqrt{8}}.$$

Substituting this bound and $c = 1/4$, $\delta = 1/(8\sqrt{N})$ into (F.10) gives a conditional expected regret of at least

$$\frac{1}{16}\left(1 - \frac{1}{\sqrt{8}}\right)\frac{1}{\sqrt{N}}$$

in every episode. The transition draws a fresh independent σ in each episode. Conditioning successively on completed episodes and summing proves (C.45).

Global linear-model completion. It remains to verify the model assumptions for every input, not only on the reachable path. Let $m = |\mathcal{S}||\mathcal{A}| = 8$, let $\alpha := \sqrt{1 - (1/2 + c)^2} = \sqrt{7}/4$, let $\{e_{(s,a)}\}_{(s,a) \in \mathcal{S} \times \mathcal{A}}$ be the first eight coordinate vectors in $\mathbb{R}^{19}$, and let e_r be the nineteenth coordinate vector. Define

$$\rho(s, a, \mathcal{P}) := \mathbf{1}\{s = s_3\}\left[\frac{1}{2} + c\, a\, \{2\,\mathrm{marg}_{\mathcal{X}}\, \mathcal{P}(1) - 1\}\right],$$
$$\phi(s, a, \mathcal{P}) := \alpha e_{(s,a)} + \rho(s, a, \mathcal{P}) e_r.$$

Since $0 \leq \rho \leq 1/2 + c = 3/4$, one has $\|\phi(s, a, \mathcal{P})\|_2^2 \leq \alpha^2 + (3/4)^2 = 1$. Set $\xi_1^* = \xi_2^* = 0$ and $\xi_3^* = e_r$. Then $u_h = \langle \phi, \xi_h^* \rangle$ gives exactly the reward above globally.

For every h and every state-action index $i = (s, a)$, define a probability kernel $Q_{h,i}$ on $\mathcal{S} \times \Delta(\mathcal{X})$ by

$$Q_{1,(s_1,a)} = \tfrac{1}{2}\delta_{(s_2^+,\bar{\mu})} + \tfrac{1}{2}\delta_{(s_2^-,\bar{\mu})}, \qquad Q_{1,(s,a)} = \delta_{(s,\bar{\mu})} \quad (s \neq s_1),$$
$$Q_{2,(s_2^\sigma,a)} = \delta_{(s_3,\mu_\sigma)}, \qquad Q_{2,(s,a)} = \delta_{(s_3,\bar{\mu})} \quad (s \notin \{s_2^+, s_2^-\}), \qquad Q_{3,(s,a)} = \delta_{(s_3,\bar{\mu})},$$

for every (s, a). Define the first eight coordinates of the vector measure ω_h^* by

$$[\omega_h^*(E)]_{(s,a)} = \alpha^{-1} Q_{h,(s,a)}(E), \qquad E \in \mathcal{B}(\mathcal{S} \times \Delta(\mathcal{X})),$$

and set its remaining coordinates to zero. Then, for every input,

$$\langle \phi(s, a, \mathcal{P}), \omega_h^*(\mathrm{d}s', \mathrm{d}\mu') \rangle = Q_{h,(s,a)}(\mathrm{d}s', \mathrm{d}\mu'),$$

which is a Borel probability kernel and reproduces the reachable dynamics. Moreover,

$$\|\omega_h^*\|_{2,\mathrm{TV}} \leq \frac{\sqrt{8}}{\alpha} = \sqrt{\frac{128}{7}} < \sqrt{19}.$$

Thus (H.9) holds and Proposition H.2 verifies Assumption 2.5.

At each fixed (h, s, a), the kernel $Q_{h,(s,a)}$ is independent of the current (μ, ν). Its population marginal is deterministic, and at $(h, s) = (1, s_1)$ the two stochastic leader-state branches share the same population $\bar{\mu}$. Consequently, it admits the factorization in Assumption 2.2. Both the population evolution map and leader-state kernel are constant in the current population arguments, so Assumption C.43 holds with $L_\Phi = L_\mathcal{T} = 0$.

Because $\mathcal{X} = \{0, 1\}$,

$$|\rho(s, a, \mathcal{P}) - \rho(s, a, \widetilde{\mathcal{P}})| \leq c\|\mathcal{P} - \widetilde{\mathcal{P}}\|_1,$$

so Assumption 2.4 holds with $L_\phi = c$. The follower feature is bounded and has strictly positive static feature-difference covariance because $b - b' \neq 0$ whenever $b \neq b'$. Also $d_F = 1 \leq |\mathcal{X}|(|\mathcal{B}| - 1) = 2$, verifying Assumptions 2.1 and C.42.

Under $\theta_h^* = 0$, the follower actions are conditionally independent uniform draws, so Assumption 2.3 holds. Since all action and state spaces are finite, $\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$ is compact and the objectives are continuous; the measurable planning and exact benchmark-attainment conditions hold. All cardinalities, dimensions, reward bounds, norm bounds, Lipschitz constants, and identifiability constants are independent of N and T. Only the pair of minimax local alternatives $\mu_\pm$ depends on N. □

Proof of Corollary 5.1. Fix integers $d_L \geq 19$, $H \geq 3$, $T \geq (d_L - 3)^2$, and $N \geq 1$, and set $k = d_L - 3$. We embed the two hard constructions into $\mathfrak{M}_{\mathrm{j}}(d_L, H)$ and use the same CO benchmark in both.

Leader-side hard family. The construction for Theorem C.44 already has horizon H, feature dimension d_L, and action set $\{-1, +1\}^k$. Relabel its four leader states and its two follower states and actions to use the common spaces specified above, with its initial leader state labeled s_1. The verification in its proof gives all the stated structural and planning assumptions, Assumption 2.3, the norm bounds (H.9), and an attained CO benchmark. Its feature maps are fixed across the hard family and do not reveal the unknown leader transition parameter. The population distribution is fixed at $\bar{\mu} = \mathrm{Ber}(1/2)$, and the true follower response is uniform, so the follower samples contain no information about that parameter. For any FSO learner, the first-step action is chosen from past observations and parameter-independent current samples, so it is an admissible randomized action in the bandit reduction, and later actions do not affect the episode return. Consequently, restricting any $\mathfrak{A} \in \mathfrak{A}_{\mathrm{FSO}}$ to this family yields

$$\sup_{M \in \mathfrak{M}_{\mathrm{j}}(d_L, H)} \mathbb{E}_M\left[\mathrm{Reg}^M_{\mathfrak{A},\mathrm{FSO}}(T)\right] \geq c_{\mathrm{LB}} d_L H\sqrt{T}. \tag{F.11}$$

Embedding the FSO observation construction. Let M_N^0 be the three-step model from the proof of Theorem C.46, and denote its globally defined features, reward parameters, and transition measures by ϕ^0, φ^0, ξ_h^0, and ω_h^0. Write $\mathbb{P}_h^0$ for its transition kernel. Set $q(a) = a_1$ for $a \in \{-1, +1\}^k$, and let

$I : \mathbb{R}^{19} \to \mathbb{R}^{d_L}$ be the isometric embedding obtained by appending zeros. On every input $(s, a, \mathcal{P})$, define the new leader feature and reward parameters by

$$\phi^{\mathrm{S}}(s, a, \mathcal{P}) = I\phi^0(s, q(a), \mathcal{P}), \qquad \xi_h^{\mathrm{S}} = \begin{cases} I\xi_h^0, & 1 \le h \le 3, \\ 0, & 4 \le h \le H. \end{cases}$$

The original construction has $\xi_1^0 = \xi_2^0 = 0$, so only step 3 has a nonzero reward. Keep the scalar follower feature $\varphi^{\mathrm{S}}(s, a, x, b) = \varphi^0(s, q(a), x, b) = b$ and set $\theta_h^* = 0$ for every $h \in [H]$. For each Borel set $E \subseteq \mathcal{S} \times \Delta(\mathcal{X})$, set

$$\omega_h^{\mathrm{S}}(E) = \begin{cases} I\omega_h^0(E), & 1 \le h \le 3, \\ I\omega_3^0(E), & 4 \le h \le H. \end{cases}$$

Thus the first three steps reproduce the original kernels with binary action $q(a)$, and every subsequent step repeats the original step-3 transition to $(s_3, \bar{\mu})$. All added rewards are zero, and the initial system state remains $(s_1, \bar{\mu})$.

These definitions preserve the linear representation globally because, for $1 \le h \le 3$ and every Borel E,

$$\left\langle \phi^{\mathrm{S}}(s, a, \mathcal{P}), \omega_h^{\mathrm{S}}(E) \right\rangle = \left\langle \phi^0(s, q(a), \mathcal{P}), \omega_h^0(E) \right\rangle = \mathbb{P}_h^0(E \mid s, q(a), \mathcal{P}),$$

and the same identity with $\mathbb{P}_3^0$ holds for $h > 3$. These are valid Borel probability kernels for every input, including unreachable states and population distributions. The isometry and the original global norm calculation give, uniformly in h,

$$\|\phi^{\mathrm{S}}(s, a, \mathcal{P})\|_2 \le 1, \qquad \|\xi_h^{\mathrm{S}}\|_2 \le 1, \qquad \|\omega_h^{\mathrm{S}}\|_{2,\mathrm{TV}} \le \sqrt{128/7} < \sqrt{19} \le \sqrt{d_L}.$$

All leader rewards remain in $[0, 1]$, and the leader feature remains Lipschitz in $\mathcal{P}$ with constant $1/4$. The population evolution map and leader-state transition kernel are independent of the current population arguments, with deterministic next population at each fixed (h, s, a), so Assumptions 2.2 and C.43 hold with $L_\Phi = L_\mathcal{T} = 0$. The follower feature remains bounded, and its static feature-difference covariance is strictly positive uniformly over (s, a). The i.i.d. state sampling and uniform follower response satisfy Assumption 2.3. The norm bounds imply Assumption 2.5 by Proposition H.2. The action simplex is finite-dimensional and compact, and the pointwise planning objectives remain Borel in the state and continuous in the local policy and follower parameter, so Assumption C.14 is preserved.

Preservation of values and FSO commitment rules. For $p \in \Delta(\{-1, +1\}^k)$, define its binary pushforward by

$$(q_\sharp p)(\tau) := \sum_{a : q(a) = \tau} p(a), \qquad \tau \in \{-1, +1\}.$$

Every binary distribution can be obtained in this way, for example by using the two representatives $a^{(\tau)} = (\tau, 1, \dots, 1)$. The rewards and transitions of the first three steps depend on a local policy only through this pushforward, since the true follower response is uniform and independent of the commitment. Thus every binary CO policy has a Borel lift, and every CO policy in the enlarged action space induces a binary policy with the same value. The optimal CO value remains $1/2 + 2c\delta$, with $c = 1/4$ and $\delta = 1/(8\sqrt{N})$, and it is attained by choosing $a^{(\sigma)}$ at population $\mu_\sigma = \mathrm{Ber}(1/2 + \sigma\delta)$.

For every fixed episode-start rule g^t, the tuple $\{q_\sharp g_h^t\}_{h=1}^3$ is a binary FSO commitment rule. Condition on the completed observable episodes and on all algorithmic randomization used to select g^t before the episode. The fresh sign σ remains independent and uniform under this conditioning. Let $P_\pm = \mathrm{Ber}(1/2 \pm \delta)$ and $Y = X_3^{1:N}$, so the conditional law of Y given $\sigma = \pm 1$ is $P_\pm^N$. For $y = (y^1, \dots, y^N) \in \mathcal{X}^N$, define

$$\widehat{\mu}_y := \frac{1}{N} \sum_{n=1}^N \delta_{y^n}, \qquad d_t(y) := \sum_{a \in \{-1,+1\}^k} a_1 \, g_3^t(s_3, \widehat{\mu}_y)(a) \in [-1, 1].$$

By the definition of the FSO commitment rules, the conditional mean of the first action coordinate given σ and Y is $d_t(Y)$. Since step 3 supplies the only nonzero reward, with both values evaluated in the embedded model, the per-episode gap in (C.34) satisfies

$$J(\pi^*) - J_{\mathrm{FSO}}(g^t) = 2c\delta\left[1 - \frac{1}{2}\left(\mathbb{E}_{P_+^N}[d_t(Y)] - \mathbb{E}_{P_-^N}[d_t(Y)]\right)\right]$$
$$\geq 2c\delta\left[1 - \mathrm{TV}(P_+^N, P_-^N)\right] \geq \frac{c_{\mathrm{samp}}}{\sqrt{N}}.$$

The first inequality is the variational bound used in (F.10), and the second follows from the original calculation $\mathrm{KL}(P_+^N, P_-^N) \leq 1/4$ and Pinsker's inequality.

Observations in the added zero-reward steps occur after the informative decision, and subsequent episodes draw fresh independent signs. Summing the conditional bounds yields, for the embedded model M_N^{S},

$$\mathbb{E}_{M_N^{\mathrm{S}}}\left[\mathrm{Reg}^{M_N^{\mathrm{S}}}_{\mathfrak{A},\mathrm{FSO}}(T)\right] \geq c_{\mathrm{samp}}\frac{T}{\sqrt{N}} \qquad \text{for every } \mathfrak{A} \in \mathfrak{A}_{\mathrm{FSO}}. \tag{F.12}$$

This bound continues to hold with the model primitives revealed, so supplying the known feature maps does not weaken it. The preceding verification and benchmark attainment show that $M_N^{\mathrm{S}} \in \mathfrak{M}_{\mathrm{j}}(d_L, H)$.

Combining the two components. Both component inequalities hold for each fixed algorithm in the same class $\mathfrak{A}_{\mathrm{FSO}}$ and use the regret in (C.34) against the optimal CO value of the model being evaluated. The model class $\mathfrak{M}_{\mathrm{j}}(d_L, H)$ is independent of T and N; it contains the leader hard family selected for each admissible T and the embedded observation model selected for each N. By (F.11) and (F.12), for every such algorithm,

$$\sup_{M \in \mathfrak{M}_{\mathrm{j}}(d_L,H)} \mathbb{E}_M\left[\mathrm{Reg}^M_{\mathfrak{A},\mathrm{FSO}}(T)\right] \geq \max\left\{c_{\mathrm{LB}} d_L H\sqrt{T},\ c_{\mathrm{samp}}\frac{T}{\sqrt{N}}\right\}$$
$$\geq \frac{1}{2}\min\{c_{\mathrm{LB}}, c_{\mathrm{samp}}\}\left(d_L H\sqrt{T} + \frac{T}{\sqrt{N}}\right).$$

Taking the infimum over $\mathfrak{A} \in \mathfrak{A}_{\mathrm{FSO}}$ proves (5.1). □

F.3.1 Marginal-observation obstruction

We examine whether the exact joint state-action feedback in the CO setting can be replaced by its two exact state and action marginal distributions while preserving a uniform $\widetilde{O}(\sqrt{T})$ regret guarantee. In the *marginal complete observation* (MCO) setting, the leader observes (s_h, μ_h) and executes the same episode-start commitment policy as under CO, but the follower feedback after disclosure of the realized local policy p_h consists only of the action marginal distribution

$$\alpha_h(b) = \sum_{x \in \mathcal{X}} \mu_h(x)\nu_h^{p_h,\theta_h^*}(b \mid x; s_h), \qquad b \in \mathcal{B}.$$

The joint distribution $\mathcal{P}_h = \mu_h \otimes \nu_h^{p_h,\theta_h^*}$ and individual follower observations are unavailable to the leader. The leader still observes its own action, realized reward, and subsequent system state, and receives no counterfactual response queries. Thus MCO changes the follower observation channel while retaining the CO commitment protocol and the original optimal CO benchmark. Let $\mathfrak{A}_{\mathrm{MCO}}$ denote the possibly randomized, model-uniform algorithms that use known inputs and completed MCO episodes to select and freeze $\pi^t \in \Pi$ before episode t. For a model M, define

$$\mathrm{Reg}^M_{\mathfrak{A},\mathrm{MCO}}(T) := \sum_{t=1}^{T}\left\{J_M(\pi_M^*) - J_M(\pi^t)\right\},$$

where both values use the true model and π_M^* maximizes over the full CO policy class Π.

Proposition F.3 (Marginal-observation obstruction)**.** *Fix positive constants η, B_Θ, and L_ϕ. There is a fixed family $\mathfrak{M}_{\mathrm{marg}}$ of $(2,1)$-linear MFGs with $H = 2$, $|\mathcal{S}| = |\mathcal{A}| = |\mathcal{B}| = 2$, $|\mathcal{X}| = 4$, and*

$\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$. *Every model satisfies Assumptions 2.1, 2.2, 2.4, 2.5, C.14, C.42 and C.43 and has an attained CO benchmark. For some* $c_{\mathrm{marg}} > 0$ *and integer* $T_0 \geq 1$*, its minimax expected regret satisfies*

$$\inf_{\mathfrak{A}\in\mathfrak{A}_{\mathrm{MCO}}} \sup_{M\in\mathfrak{M}_{\mathrm{marg}}} \mathbb{E}_M\left[\mathrm{Reg}^M_{\mathfrak{A},\mathrm{MCO}}(T)\right] \geq c_{\mathrm{marg}} \frac{T^{2/3}}{[\log(2+T)]^{2/3}} \quad (T \in \mathbb{N},\ T \geq T_0). \tag{F.13}$$

The positive constants c_{marg} *and* T_0*, the family, its known feature maps, and all regularity and norm budgets are independent of* T*. The lower bound holds even when the common leader reward and transition primitives, the initial system state, and the description of the family are supplied to the learner. Under the same supplied information, exact joint observations permit regret bounded independently of* T *on this family.*

Proof of Proposition F.3. We construct a single family with constant observed marginals and a scalar unknown follower parameter.

A fixed width function. Choose $0 < \tau < \min\{B_\Theta/2, 1/2\}$ and fix

$$c_* = \frac{\tau}{8}, \qquad a_0 = \frac{\tau}{32}, \qquad \epsilon_0 = \frac{\tau}{256}, \qquad \epsilon_n = \epsilon_0 2^{-n}, \qquad \ell_n = \frac{\tau}{8(n+1)^2} \quad (n \geq 1).$$

Set $t_n = \tau + \sum_{m<n}(\ell_m + 4\epsilon_m)$ and $I_n = [t_n, t_n + \ell_n]$. Since $\sum_{n\geq 1}(\ell_n + 4\epsilon_n) \leq \tau/8 + \tau/64 < \tau/2$, these intervals lie in $[\tau, 3\tau/2]$ and the gap following I_n has length $4\epsilon_n$. For $\theta \in [\tau, 2\tau]$, define the single known function

$$w(\theta) := \min\left\{a_0,\ |\theta - 2\tau|,\ \inf_{m\geq 1}\left[\epsilon_m + \mathrm{dist}(\theta, I_m)\right]\right\}.$$

The distance is Euclidean distance on the real line. Every function inside the infimum is one-Lipschitz, so w is one-Lipschitz and takes values in $[0, a_0]$. For $\theta \in I_n$, its own term equals ϵ_n, the earlier terms are at least ϵ_n, and every later interval is at distance at least $4\epsilon_n$. The cap a_0 and the distance to 2τ are also larger than ϵ_n on I_n. Consequently,

$$w|_{I_n} = \epsilon_n, \qquad w(2\tau) = 0.$$

This function and its intervals are fixed once and for all.

Follower responses with identical marginals. Take $\mathcal{S} = \{y_0, y_1\}$, $\mathcal{A} = \mathcal{B} = \{0, 1\}$, $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$, and $\bar{\mu} = \mathrm{Unif}(\mathcal{X})$. Use the following scalar follower feature at *both* leader states:

$$\varphi(s, a, x, b) = \begin{cases} \mathbf{1}\{b = 1\}, & x = x_1, \\ \mathbf{1}\{b = 0\}, & x = x_2, \\ a\mathbf{1}\{b = 1\}, & x = x_3, \\ a\mathbf{1}\{b = 0\}, & x = x_4. \end{cases}$$

The model indexed by $\theta \in [\tau, 2\tau]$ has $\theta^*_1 = \theta$ and $\theta^*_2 = 0$. These parameters lie in the source ball Θ, the feature is bounded by one, and every true follower reward belongs to $[0, 1]$ because $2\tau < 1$. For each (s, a), the scalar static feature-difference covariance in Assumption C.42 is $4 + 4a^2 \geq 4$, and $d_F = 1 \leq |\mathcal{X}|(|\mathcal{B}| - 1)$. For a local policy $p \in \Delta(\{0, 1\})$, write $v = p(1)$ and let $q(u) = e^{\eta u}/(1 + e^{\eta u})$. At population $\bar{\mu}$, the step-1 joint table, with columns indexed by $b = 0, 1$, is

$$\mathcal{P}^v_\theta = \frac{1}{4}\begin{pmatrix} 1 - q(\theta) & q(\theta) \\ q(\theta) & 1 - q(\theta) \\ 1 - q(\theta v) & q(\theta v) \\ q(\theta v) & 1 - q(\theta v) \end{pmatrix}. \tag{F.14}$$

Its state marginal is $\bar{\mu}$ and its action marginal is $(1/2, 1/2)$ for every (θ, v). At step 2 the joint is $\mathcal{P}_\circ = \bar{\mu} \otimes \mathrm{Unif}(\mathcal{B})$ for every commitment and leader state.

A globally defined leader feature. Set $\lambda_0 = [8(1+e^{2\eta\tau})]^{-1}$. For every ambient joint table $\mathcal{P} \in \Delta(\mathcal{X} \times \mathcal{B})$, define the truncated cell logarithms by

$$L_{i,b}(\mathcal{P}) = \log \max\{\mathcal{P}(x_i, b), \lambda_0\}.$$

For an interval $[u_-, u_+]$, let $[z]_{[u_-,u_+]} = \min\{u_+, \max\{u_-, z\}\}$ denote scalar truncation. Define

$$\begin{aligned} \widehat{\theta}(\mathcal{P}) &= \left[\frac{L_{1,1}(\mathcal{P}) - L_{1,0}(\mathcal{P})}{\eta}\right]_{[\tau,2\tau]}, \quad \widehat{r}(\mathcal{P}) = \left[\frac{L_{3,1}(\mathcal{P}) - L_{3,0}(\mathcal{P})}{\eta}\right]_{[0,2\tau]}, \\ \beta(\mathcal{P}) &= \left[w(\widehat{\theta}(\mathcal{P})) - |\widehat{r}(\mathcal{P}) - c_*|\right]_+. \end{aligned} \tag{F.15}$$

Every cell of (F.14) is at least $[4(1+e^{2\eta\tau})]^{-1}$, so the logarithmic truncation is inactive on these physical tables. The two decoders therefore equal θ and θv, respectively, and

$$\beta(\mathcal{P}^v_\theta) = [w(\theta) - |\theta v - c_*|]_+.$$

Globally, $0 \le \beta \le a_0$, and the Lipschitz properties of truncated logarithms, scalar truncation, and w give the bound $2/(\eta\lambda_0)$ for the ℓ_1-Lipschitz constant of β. At $\mathcal{P}_\circ$, the second decoder is zero, so $\beta(\mathcal{P}_\circ) = 0$ because $a_0 < c_*$. Choose a fixed $0 < c_b \le \min\{1, L_\phi \eta \lambda_0/2\}$ and define, for every $(s, a, \mathcal{P})$,

$$\phi(s, a, \mathcal{P}) = \begin{cases} (2^{-1/2}, c_b\beta(\mathcal{P})), & s = y_0, \\ (2^{-1/2}, 1/2), & s = y_1. \end{cases}$$

This step-independent feature has Euclidean norm at most one and joint-Lipschitz constant at most L_ϕ.

Transitions, rewards, and model-class membership. Let the deterministic initial system state be $(y_0, \bar{\mu})$ and define the vector measures on the full system state space by

$$\begin{aligned} \xi^*_1 &= 0, & \omega^*_1 &= (2^{-1/2}, 1/4)\delta_{(y_1,\bar{\mu})} + (2^{-1/2}, -1/4)\delta_{(y_0,\bar{\mu})}, \\ \xi^*_2 &= (0, 1), & \omega^*_2 &= (\sqrt{2}, 0)\delta_{(y_0,\bar{\mu})}. \end{aligned}$$

Use these measures and coefficients in the linear representation of Definition 2.6. For every ambient input, the first transition assigns probabilities $1/2 + \phi_2(s, a, \mathcal{P})/4$ and $1/2 - \phi_2(s, a, \mathcal{P})/4$ to $(y_1, \bar{\mu})$ and $(y_0, \bar{\mu})$, respectively. Both are nonnegative and sum to one, and the second transition resets to $(y_0, \bar{\mu})$ with probability one. These formulas specify valid Borel kernels even at unreachable inputs. The population evolution map is constantly $\bar{\mu}$, establishing Assumption 2.2 with $L_\Phi = 0$. The leader kernel is Lipschitz in (μ, ν) because $\|\mu \otimes \nu - \mu' \otimes \nu'\|_1 \le \|\mu - \mu'\|_1 + \|\nu - \nu'\|_1$ and its probability vector is Lipschitz in $\mathcal{P}$. In particular, $L_\mathcal{T} = L_\phi/2$ is valid in Assumption C.43. The leader rewards are globally in $[0, 1]$, and $\|\xi^*_h\|_2 \le 1$.

For a bounded Borel test function taking values $f_1, f_0 \in [-1, 1]$ on the two support points,

$$\int f \, \mathrm{d}\omega^*_1 = \left(\frac{f_1 + f_0}{\sqrt{2}}, \frac{f_1 - f_0}{4}\right), \qquad \int f \, \mathrm{d}\omega^*_2 = (\sqrt{2} f_0, 0).$$

The first norm is maximized at a vertex of $[-1, 1]^2$ and is at most $\sqrt{2}$, as is the second norm. Thus (H.9) holds with $d_L = 2$, and Proposition H.2 give Assumption 2.5 with $R_U = 1 + \sqrt{2}$ uniformly over the required compact feasible sets.

The optimistic planning objectives are Borel in the system state and continuous in the local policy and candidate follower parameter, since the feature is continuous in $\mathcal{P}$ and the softmax response is continuous in those arguments. The finite-action simplex and compact parameter sets therefore satisfy the finite-action planning conditions in Assumption C.14.

Write $Y = \mathbf{1}\{s_2 = y_1\}$. Under the true step-2 parameter $\theta^*_2 = 0$, the only nonzero realized reward is $Y/2$, independently of the second-step commitment, and the terminal system state is fixed. Consequently, writing $c_{\mathrm{sig}} = c_b/8$, the value of a policy with first-step commitment mass v is

$$J_\theta(v) = \frac{1}{4} + c_{\mathrm{sig}}[w(\theta) - |\theta v - c_*|]_+. \tag{F.16}$$

The choice $v_\theta^* = c_*/\theta \in [1/16, 1/8]$ is feasible and maximizes this value over the entire simplex, giving $J_\theta^* = 1/4 + c_{\mathrm{sig}} w(\theta)$. Extending this initial-state choice by a constant Borel commitment at other inputs produces an attained CO benchmark in Π. Pointwise CO Bellman maximizers at all other system states also exist by continuity in the commitment and compactness of the action simplex. This completes the verification of the stated structural assumptions for the fixed family $\mathfrak{M}_{\mathrm{marg}} = \{M_\theta : \theta \in [\tau, 2\tau]\}$. All its primitives except θ_1^* are common and may be revealed to the learner.

Packing and the complete observable likelihood. Fix a sufficiently large n and select points $\theta_1, \dots, \theta_{K_n} \in I_n$ separated by $128\epsilon_n$. An equally spaced selection gives

$$K_n \geq \frac{\ell_n}{128\epsilon_n} = \frac{c_{\mathrm{pack}}}{(n+1)^2\epsilon_n}, \qquad c_{\mathrm{pack}} = \frac{\tau}{1024}. \tag{F.17}$$

The centers $v_j^* = c_*/\theta_j$ satisfy

$$|v_i^* - v_j^*| = \frac{c_*|\theta_i - \theta_j|}{\theta_i\theta_j} \geq \frac{|\theta_i - \theta_j|}{32\tau}.$$

Hence the open sets $U_j = \{v \in [0,1] : |\theta_j v - c_*| < \epsilon_n\}$ are pairwise disjoint, since each has radius at most ϵ_n/τ around its center. In model M_{θ_j}, every episode whose first commitment lies outside U_j incurs value gap $c_{\mathrm{sig}}\epsilon_n$. Use $M_{2\tau}$ as the null model; its value is identically $1/4$ because $w(2\tau) = 0$.

Fix any algorithm and let v_t be its first-step commitment mass in episode t. Let $N_j = \sum_{t=1}^T \mathbf{1}\{v_t \in U_j\}$, and let $\mathbb{P}_0$ and $\mathbb{P}_j$ denote the laws of its complete observable transcripts under the null model and M_{θ_j}. The learner's randomization can be included as a model-independent seed in these transcripts. Conditionally on the past and v_t, the only model-dependent observation is

$$Y_t \sim \mathrm{Ber}\left(\frac{1}{2} + \frac{c_b}{4}[\epsilon_n - |\theta_j v_t - c_*|]_+\right)$$

under M_{θ_j}, compared with $\mathrm{Ber}(1/2)$ under the null. The exact state and action marginals are constant, the first leader action is drawn from the known commitment, and the second-step reward is a deterministic function of Y_t. Subsequent actions and the terminal state supply no further model-dependent factor in the conditional likelihood. Thus this calculation retains all the leader feedback allowed by the MCO protocol.

For $0 \leq x \leq \epsilon_n/4$, the Bernoulli relative entropy obeys

$$\mathrm{KL}(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2 + x)) = -\frac{1}{2}\log(1 - 4x^2) \leq 4x^2 \leq \epsilon_n^2.$$

The chain rule, including adaptive commitment choices, therefore gives

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_j) \leq \epsilon_n^2 \mathbb{E}_0 N_j.$$

Applying Pinsker's inequality to $N_j/T \in [0,1]$, averaging over j, and using Cauchy–Schwarz and $\sum_j N_j \leq T$, we obtain

$$\frac{1}{K_n}\sum_{j=1}^{K_n}\frac{\mathbb{E}_j N_j}{T} \leq \frac{1}{K_n}\sum_{j=1}^{K_n}\frac{\mathbb{E}_0 N_j}{T} + \frac{\epsilon_n}{K_n}\sum_{j=1}^{K_n}\sqrt{\frac{\mathbb{E}_0 N_j}{2}} \leq \frac{1}{K_n} + \epsilon_n\sqrt{\frac{T}{2K_n}}. \tag{F.18}$$

The expected regret in model M_{θ_j} is at least $c_{\mathrm{sig}}\epsilon_n(T - \mathbb{E}_j N_j)$ by (F.16). Only the first-step commitment affects the return, so the same argument also permits arbitrary causal use of observations in subsequent decisions.

Choice of scale and regret bound. Choose a fixed n_0 sufficiently large that the packing above has $K_n \geq 4$ for every $n \geq n_0$. For each sufficiently large T, let $n \geq n_0$ be the smallest index satisfying

$$T(n+1)^2\epsilon_n^3 \leq \frac{c_{\mathrm{pack}}}{8}. \tag{F.19}$$

Such an index exists because ϵ_n decays geometrically. Equations (F.17) and (F.19) make the right-hand side of (F.18) at most $1/2$. For all sufficiently large T, this minimal index is larger than n_0, and failure of (F.19) at $n-1$ gives

$$8Tn^2\epsilon_n^3 > \frac{c_{\mathrm{pack}}}{8}, \qquad \epsilon_n > \left(\frac{c_{\mathrm{pack}}}{64Tn^2}\right)^{1/3}.$$

Moreover, testing the index $n_0 + \lceil \log_2(2+T) \rceil$ in (F.19) shows that $n \le C \log(2+T)$ for a fixed C and all sufficiently large T. It follows that

$$\epsilon_n \ge c\, T^{-1/3} [\log(2+T)]^{-2/3}.$$

Averaging regret over the packed models now yields

$$\sup_{\theta \in [\tau, 2\tau]} \mathbb{E}_\theta \left[\mathrm{Reg}^{M_\theta}_{\mathfrak{A}, \mathrm{MCO}}(T) \right] \ge \frac{c_{\mathrm{sig}}}{2} T \epsilon_n \ge c_{\mathrm{marg}} \frac{T^{2/3}}{[\log(2+T)]^{2/3}}.$$

The constants are uniform over algorithms, so taking the infimum proves (F.13). Only the finite subset used in the testing argument depends on T; the family and its feature maps were already fixed.

Comparison with exact joint observation. With exact joint feedback, (F.15) identifies θ_1^* after the first episode, whatever first-step commitment was used. Given the common primitives, every subsequent episode can choose $v = c_*/\theta_1^*$ and attain the optimal CO value. The resulting total regret is at most the first episode's value gap, hence at most $H = 2$. □

The fixed-dimensional family also lies in the model class of the CO upper bound (4.3). Since $T^{2/3}/[\log(2+T)]^{2/3}$ eventually exceeds $\sqrt{T}$ times any fixed power of $\log(2+T)$, Proposition F.3 rules out replacing joint observation by exact marginal observation in a uniform $\widetilde{O}(\sqrt{T})$ guarantee over the stated model class. It establishes the statistical role of state–action association information while allowing that additional model structure can make marginals sufficient.

G Proofs for the finite-sample observation setting

This appendix gives the FSO proof architecture, concentration and stability results, and proofs of the FSO regret bounds.

G.1 Proof architecture for Theorem 4.1

G.1.1 The sample-size question and the role of the assumptions

Compared with CO, the FSO setting introduces an empirical plug-in error because the leader plans and performs Bellman regression using $(\widehat{\mu}_h^t, \widehat{\mathcal{P}}_h^t)$ rather than $(\mu_h^t, \mathcal{P}_h^t)$. Since the CO analysis yields a $\sqrt{T}$-type learning rate in T, a natural question arises:

How many follower samples per step are needed for finite-sample observation to recover the CO $\sqrt{T}$-type regret rate?

The main guarantees answer this question through Theorem 4.1 and Corollary 4.3. Under the explicit sufficient condition in (G.37), the displayed finite-sample contribution is controlled with $N = \widetilde{\Omega}(T|\mathcal{Z}|)$ when the primitive constants are fixed and the stated radius-absorption conditions hold. In the finite-rank specialization, recovery of the $\sqrt{T}$ temporal order further uses the dimension and regularity conditions stated in Corollary 4.3. The generic complete-cover argument yields the leader scale $\widetilde{O}(d_L^{3/2} H^2 \sqrt{T})$ when $d_F^2 = O(d_L^2)$. A tabular comparison with prior regret guarantees appears in Table 8.

The FSO analysis has two statistical departures from the CO setting. First, the leader forms empirical joint-distribution queries, and the resulting historical target perturbations are absorbed into the actual predictor radius while the current-query perturbations remain in the finite-sample coefficient. Second, the follower parameter is estimated through a well-specified maximum-likelihood confidence set.

Four ingredients close the proof. The batch-adaptive performance-difference identity in Lemma G.8 yields the regret decomposition in Lemma G.9, which separates the leader Bellman error, the follower joint-distribution error, and the two-shift batch mismatch. The actual-predictor Bellman-error lemma evaluates uncertainty at true queries and converts only the executed bonus to the empirical design in Lemma G.16. The well-specified maximum-likelihood result gives $\theta_h^* \in \Theta_h^t$ with radius β_{F}^t in Lemma G.5. The direct joint-distribution-error lemma retains the exact factor $L_{T,N}$ and combines

the TV perturbation bound with matrix Cauchy–Schwarz in Lemma G.17. After these ingredients are combined, every term that is linear in $HT\bar{\epsilon}_N$ is collected as $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})HT\bar{\epsilon}_N$. The constants and the sufficient sample-size condition for radius absorption are specified in Remark C.35. The following proof architecture organizes these statements and their proofs. Combining them yields the main guarantee for Algorithm 3.

The follower model and paired-sampling assumptions support response estimation, while RKHS Bellman optimality completeness and feature regularity support the leader's value estimation. Measurable planning and benchmark attainment serve different purposes: Assumption C.14 defines the algorithm's optimizer, whereas Proposition C.3 supplies the CO policy used for comparison. The confidence-calibration and radius-absorption conditions control the estimation and plug-in errors in the FSO updates.

The FSO proof follows this sequence of steps:

$$\text{regret decomposition} \;\to\; \text{leader Bellman error} \;+\; \text{follower response error} \;+\; \text{batch mismatch},$$

but each component requires a mean-field-specific modification. The full proof is organized around the following four ingredients.

(I) FSO regret decomposition. Lemma G.9 decomposes regret into

$$\mathrm{Reg}_{\mathrm{FSO}}(T) \le \mathrm{LBE}' + L_U^{\mathrm{FSO}}\mathrm{JDE}' + \mathrm{Bias}_{\mathrm{batch}}.$$

The batch mismatch satisfies $\mathrm{Bias}_{\mathrm{batch}} \le 2HT\bar{\epsilon}_N C_{\mathrm{pair}}^{\mathrm{env}}(B_T^{\mathrm{FSO}})$ by (G.29). The current state batch affects the executed local policy, so the performance-difference identity keeps that policy inside the batch expectation. To control the mismatch, first move the optimistic value envelope from the true to the empirical population, then move the fixed empirical optimizer's objective back to the true population. Both shifts use value-level Lipschitz bounds; no comparison between two optimizer selectors is needed.

(II) Leader-side LBE control. Lemma G.16 controls LBE' using RKHS self-normalized concentration. KRR uses empirical designs and empirical next-population targets. A simultaneous oracle uncertainty event covers the complete deterministic class of optimistic continuations. On a predictable historical event, a same-design KRR comparison absorbs target perturbations into the actual radius, yielding uncertainty at true queries. The historical target perturbation is incorporated through $B_T^{\mathrm{FSO}} = B_T^{\mathrm{base}} + \beta_{\mathrm{tar}}^{\max}$ in (G.34). The non-circularity and cap of this radius are established in Lemma G.12.

Only the executed bonus is subsequently converted to the empirical design for the potential bound. Current-batch errors remain in conditional expectation and are controlled by a first-moment bound, separately from historical target inflation. The current-query input perturbation is handled by the explicit bonus $\Gamma_{\mathrm{FS}} = L_U^{\mathrm{FSO}}\bar{\epsilon}_N$. Together these steps produce the finite-sample residual. The complete value-class cover makes uncertainty valid for adaptive regression targets, while information gain bounds the accumulated executed bonuses.

(III) Follower-side MLE confidence. Lemma G.5 gives simultaneous coverage $\theta_h^* \in \Theta_h^t$ for the step-specific softmax MLE through a Hellinger-affinity drift-normalized likelihood-ratio argument under the conditional paired-response law. This coverage-free result adapts the MLE concentration step in Chen et al. (2023a) to batched FSO observations.

(IV) JDE control by TV perturbation bound and matrix Cauchy–Schwarz. Lemma G.17 first applies a deterministic TV perturbation bound to move from μ_h^t to $\widehat{\mu}_h^t$ before any square-root or Cauchy–Schwarz step. This extracts the empirical mean-field error as an additive $2HT\bar{\epsilon}_N$ residual. The remaining response error is controlled in Fisher geometry by matrix Cauchy–Schwarz and a comparison with the auxiliary matrix V_h^t, whose recurrence is additive. Population-weighted aggregation places the state sum inside the trace, so a single elliptical-potential argument avoids the spurious statewise $|\mathcal{X}|$ multiplier. This ordering prevents the $T^{3/4}$ contribution of the naive perturbation route while retaining the genuine $|\mathcal{Z}|$ dependence in population estimation.

Combining (I)–(IV) on the five-event intersection in the proof of Theorem 4.1 gives (4.1); the finite-rank simplification and the data-abundant regime are summarized in Corollary 4.3 and remark C.33.

Table 9: Dependency map for the principal theoretical results.

Result	Required structure
Benchmark existence	Compact local-policy class and generated-value attainment, or an explicit approximation slack
Oracle uncertainty	RKHS closure, complete deterministic cover, and predictable empirical design
Actual uncertainty	Predictable historical event, same-design KRR push-through, and radius absorption
Batch-adaptive performance difference	Causal factorization and joint measurability
Leader Bellman error	Actual uncertainty, true-query evaluation, bonus martingale, and elliptical potential
Main follower-response error	Well-specified MLE, TV perturbation bound, and Fisher geometry
Main regret theorem	Benchmark optimism, batch-adaptive performance difference, leader Bellman error, follower-response error, and the intersection of the required events

G.2 Lipschitz, stability, and continuity lemmas

Proof of Lemma C.23. Fix h, t, s and the episode-prefix-measurable feasible set Θ_h^t. For fixed $(p, \theta) \in \Pi_{\mathrm{loc}} \times \Theta_h^t$, define

$$G_h^t(s, \mu; p, \theta) := \left\langle \widehat{U}_h^t\left(s, \cdot, \mu \otimes \nu^{p,\theta}(\cdot \mid \cdot; s)\right), p \right\rangle_{\mathcal{A}}.$$

Using the response in (C.2) at the fixed pair (p, θ) and summing over follower actions gives

$$\left\|\mu \otimes \nu^{p,\theta} - \mu' \otimes \nu^{p,\theta}\right\|_1 = \|\mu - \mu'\|_1.$$

By the class-membership premise and Assumption 2.4, the RKHS component and the posterior-standard-deviation bonus each have joint-distribution Lipschitz constant at most $B_T^{\mathrm{FSO}} L_\phi \lambda^{-1/2}$. The scalar offset is independent of the query and truncation is nonexpansive, giving the common Lipschitz constant $L_U^{\mathrm{FSO}} = 2L_\phi B_T^{\mathrm{FSO}} \lambda^{-1/2}$ in (G.38). Therefore, uniformly over (p, θ),

$$\left|G_h^t(s, \mu; p, \theta) - G_h^t(s, \mu'; p, \theta)\right| \le C_G^{\mathrm{act}}\left(B_T^{\mathrm{FSO}}\right) \|\mu - \mu'\|_1 \le L_U^{\mathrm{FSO}} \|\mu - \mu'\|_1.$$

The feasible domain $\Pi_{\mathrm{loc}} \times \Theta_h^t$ is the same at the two hypothetical arguments. The elementary inequality

$$\left|\sup_x f(x) - \sup_x g(x)\right| \le \sup_x |f(x) - g(x)|$$

therefore gives (C.35). By the definition of C_W^{act} and the common envelope in (G.38), the tight constant satisfies (C.36). No Lipschitz property of an argmax selector and no horizon relaxation are used. □

Proof of Lemma C.25. The lower bound follows because $\widehat{W}_h^t(s, \mu)$ is the supremum of the fixed-pair objective at μ. For the upper bound, write

$$\begin{aligned}\widehat{W}_h^t(s, \mu) - G_h^t(s, \mu; p_{\widehat{\mu}}, \bar{\theta}_{\widehat{\mu}}) = [\widehat{W}_h^t(s, \mu) - \widehat{W}_h^t(s, \widehat{\mu})] + [\widehat{W}_h^t(s, \widehat{\mu}) - G_h^t(s, \widehat{\mu}; p_{\widehat{\mu}}, \bar{\theta}_{\widehat{\mu}})] \\ + [G_h^t(s, \widehat{\mu}; p_{\widehat{\mu}}, \bar{\theta}_{\widehat{\mu}}) - G_h^t(s, \mu; p_{\widehat{\mu}}, \bar{\theta}_{\widehat{\mu}})].\end{aligned}$$

The middle term is zero by optimality at $\widehat{\mu}$. The first and last terms are at most $C_W^{\mathrm{act}}(B_T^{\mathrm{FSO}})\|\mu - \widehat{\mu}\|_1$ and $C_G^{\mathrm{act}}(B_T^{\mathrm{FSO}})\|\mu - \widehat{\mu}\|_1$, respectively. Adding them and applying (G.38) proves (C.37). □

Auxiliary: self-concordance of softmax covariance.

Lemma G.1 (Self-concordance of the softmax covariance)**.** *For any two parameters $\theta_1, \theta_2 \in \mathbb{R}^{d_F}$ and any fixed (s, p, x) with $p \in \Delta(\mathcal{A})$,*

$$\exp\left(-2\eta \, \|\theta_1 - \theta_2\|_\infty^{\varphi^p}\right) \Sigma_{x,s}^{p,\theta_1} \preceq \Sigma_{x,s}^{p,\theta_2} \preceq \exp\left(2\eta \, \|\theta_1 - \theta_2\|_\infty^{\varphi^p}\right) \Sigma_{x,s}^{p,\theta_1},$$

where $\|\theta_1-\theta_2\|_\infty^{\varphi^p} := \max_b |\langle \varphi^p(s,x,b), \theta_1 - \theta_2\rangle| \le B_\varphi \|\theta_1-\theta_2\|_2$ *and* $\Sigma^{p,\theta}_{x,s} := \mathbb{C}_{b\sim\nu^{p,\theta}(\cdot|x;s)}[\varphi^p(s,x,b)]$ *is the per-state response feature covariance under the local policy* p.

Proof of Lemma G.1. We prove the covariance comparison from a likelihood-ratio bound. Let $\Delta := \|\theta_1-\theta_2\|_\infty^{\varphi^p}$ and $\nu_i(b) := \nu^{p,\theta_i}(b\,|\,x;s)$ for $i\in\{1,2\}$, with feature shorthand $\varphi_b := \varphi^p(s,x,b)$.

Step 1 (Likelihood-ratio sandwich). The softmax form gives

$$\frac{\nu_1(b)}{\nu_2(b)} = \frac{\exp(\eta\langle\varphi_b,\theta_1-\theta_2\rangle)}{Z(\theta_1)/Z(\theta_2)}, \qquad Z(\theta) := \sum_{b'}\exp(\eta\langle\varphi_{b'},\theta\rangle).$$

The numerator lies in $[e^{-\eta\Delta}, e^{\eta\Delta}]$, and $Z(\theta_1)/Z(\theta_2)\in[e^{-\eta\Delta},e^{\eta\Delta}]$ by log-sum-exp Lipschitzness. Hence

$$e^{-2\eta\Delta} \le \frac{\nu_1(b)}{\nu_2(b)} \le e^{2\eta\Delta}, \qquad \forall\, b\in\mathcal{B}. \tag{G.1}$$

Step 2 (From pointwise sandwich to covariance sandwich, with PSD centering correction). Let $\bar\varphi_i := \mathbb{E}_{\nu_i}[\varphi_b]$. From (G.1),

$$\Sigma^{p,\theta_1}_{x,s} = \sum_b \nu_1(b)\,(\varphi_b-\bar\varphi_1)(\varphi_b-\bar\varphi_1)^\top \succeq e^{-2\eta\Delta}\sum_b \nu_2(b)\,(\varphi_b-\bar\varphi_1)(\varphi_b-\bar\varphi_1)^\top.$$

The remaining sum is centered at $\bar\varphi_1$, not at the natural centroid $\bar\varphi_2$ of ν_2. By the bias-variance decomposition,

$$\sum_b \nu_2(b)\,(\varphi_b-\bar\varphi_1)(\varphi_b-\bar\varphi_1)^\top = \Sigma^{p,\theta_2}_{x,s} + (\bar\varphi_2-\bar\varphi_1)(\bar\varphi_2-\bar\varphi_1)^\top \succeq \Sigma^{p,\theta_2}_{x,s},$$

where the inequality uses that the rank-one centering correction $(\bar\varphi_2-\bar\varphi_1)(\bar\varphi_2-\bar\varphi_1)^\top$ is PSD. Combining,

$$\Sigma^{p,\theta_1}_{x,s} \succeq e^{-2\eta\Delta}\,\Sigma^{p,\theta_2}_{x,s}.$$

The reverse inequality follows by exchanging θ_1 and θ_2. □

G.3 FSO FILTRATIONS AND EMPIRICAL CONCENTRATION

For the FSO analysis, let the complete leader-observed record from episode τ be

$$\mathsf{O}^\tau_{\mathrm{FSO}} := \left(s^\tau_h, a^\tau_h, u^\tau_h, x^{\tau,1:N}_h, b^{\tau,1:N}_h, \widehat\mu^\tau_h, \widehat{\mathcal{P}}^\tau_h, s^\tau_{h+1}\right)_{h\in[H]}.$$

Define the observable episode-start filtration by

$$\mathcal{F}^{t-1}_{\mathrm{obs}} := \sigma\left(\mathsf{O}^1_{\mathrm{FSO}},\dots,\mathsf{O}^{t-1}_{\mathrm{FSO}}\right). \tag{G.2}$$

Every episode-start FSO algorithmic object is constructed from $\mathcal{F}^{t-1}_{\mathrm{obs}}$ and known model-class inputs.

The empirical distributions and all other algorithmic quantities in $\mathsf{O}^\tau_{\mathrm{FSO}}$ are functions of the observed record and do not reveal any true population distribution.

The FSO algorithm is required to be a model-uniform function of this observable history and the known model-class inputs; in particular, it is not parameterized by the numerical value of any true population distribution. This restriction is substantive even when μ_0 is deterministic, because probabilistic measurability alone cannot encode whether a model-fixed constant has been revealed to the learner. In the CO analysis, the same notation is used with the natural CO observable record, which includes the exact population variables supplied by the CO protocol.

For analysis, enlarge the observable episode-start filtration and define the associated conditional expectation by

$$\mathcal{F}^{t-1} := \mathcal{F}^{t-1}_{\mathrm{obs}} \vee \sigma\left(\{\mu^\tau_h, \mathcal{P}^\tau_h : \tau<t,\ h\in[H]\}, \{\mu^\tau_{H+1}:\tau<t\}\right), \tag{G.3}$$
$$\mathbb{E}^t[X] := \mathbb{E}\left[X\mid\mathcal{F}^{t-1}\right],$$

for every integrable episode-t random variable X.

Thus $\mathcal{F}^{t-1}_{\mathrm{obs}}\subseteq\mathcal{F}^{t-1}$. Every episode-start FSO algorithmic object is constructed from $\mathcal{F}^{t-1}_{\mathrm{obs}}$ and known model-class inputs, whereas $\mathcal{F}^{t-1}$ is used only for proof events and conditional expectations.

Composition with the empirical-measure map. For every (h,t), denote the FSO maximizing selector fixed in Assumption C.14 by $(g_h^t(\omega;s,m),\bar\theta_h^t(\omega;s,m))$. For a state batch $\mathbf{x}=(x^1,\dots,x^N)\in\mathcal{X}^N$, write $\widehat\mu_{\mathbf{x}}:=N^{-1}\sum_{n=1}^N\delta_{x^n}$. Composing this fixed selector with the Borel empirical-measure map gives

$$(\omega,s,\mathbf{x})\longmapsto\left(p_{h,\mathbf{x}}^t(\omega;s),\bar\theta_{h,\mathbf{x}}^t(\omega;s)\right), \tag{G.4}$$

where $p_{h,\mathbf{x}}^t(\omega;s):=g_h^t(\omega;s,\widehat\mu_{\mathbf{x}})$ and $\bar\theta_{h,\mathbf{x}}^t(\omega;s):=\bar\theta_h^t(\omega;s,\widehat\mu_{\mathbf{x}})$. This map is measurable with respect to $\mathcal{F}_{\mathrm{obs}}^{t-1}\otimes\mathcal{B}(\mathcal{S})\otimes\mathcal{B}(\mathcal{X}^N)$. The corresponding candidate response is

$$\widetilde\nu_{h,\mathbf{x}}^t(\cdot\mid x;s):=\nu^{p_{h,\mathbf{x}}^t(\omega;s),\bar\theta_{h,\mathbf{x}}^t(\omega;s)}(\cdot\mid x;s). \tag{G.5}$$

Conditionally on $\mathcal{F}_{\mathrm{obs}}^{t-1}$, this is a Borel stochastic kernel in $(s,\mathbf{x},x)$. After conditioning, suppress the observed-history argument ω. On the realized batch $x_h^{t,1:N}$, abbreviate $p_h^t:=p_{h,x_h^{t,1:N}}^t(s_h^t)$ and $\bar\theta_h^t:=\bar\theta_{h,x_h^{t,1:N}}^t(s_h^t)$.

Within-step filtrations for FSO. For each episode t and step h, let $\mathcal{F}_{t,h}^-$ be the σ-field used in the analysis, generated by the complete history immediately before the state batch $x_h^{t,1:N}$ is drawn and including the current system state (s_h^t,μ_h^t). These filtrations are used only in the analysis and do not enlarge the information available to the algorithm. Define

$$\mathcal{F}_{t,h}^X:=\sigma\left(\mathcal{F}_{t,h}^-,x_h^{t,1:N},\widehat\mu_h^t,p_h^t,\bar\theta_h^t\right),\qquad \mathcal{F}_{t,h}^B:=\sigma\left(\mathcal{F}_{t,h}^X,b_h^{t,1:N},\widehat{\mathcal{P}}_h^t,a_h^t,u_h^t\right), \tag{G.6}$$

where $(p_h^t,\bar\theta_h^t)$ is the realized maximizing pair obtained by evaluating (G.4) at $(s_h^t,x_h^{t,1:N})$. Thus $\mathcal{F}_{t,h}^X$ is the information available after the current state batch and its batch-adaptive maximizing pair have been formed, whereas $\mathcal{F}_{t,h}^B$ is the information available immediately before the next system-state innovation. Under Assumption 2.2, conditionally on $\mathcal{F}_{t,h}^B$, the next system state is generated from the true transition kernel at the true joint distribution $\mathcal{P}_h^t=\mu_h^t\otimes\nu^{p_h^t,\theta_h^*}(\cdot\mid\cdot;s_h^t)$.

Let $\epsilon_{\mu,N}$ and ϵ_N denote the deterministic marginal and joint error scales defined in Lemma G.3. For each $t\in[T+1]$, define the predictable historical event

$$\mathcal{E}_{\mathrm{emp}}^{<t}:=\bigcap_{\tau<t}\bigcap_{h\in[H]}\left\{\|\widehat\mu_h^\tau-\mu_h^\tau\|_1\le\epsilon_{\mu,N},\quad\|\widehat{\mathcal{P}}_h^\tau-\mathcal{P}_h^\tau\|_1\le\epsilon_N\right\}.$$

The event $\mathcal{E}_{\mathrm{emp}}^{<t}$ belongs to the episode-start history $\mathcal{F}^{t-1}$ and may therefore be used inside $\mathbb{E}^t[\cdot]$. By contrast, the current-batch inequalities at episode t are not $\mathcal{F}^{t-1}$-measurable and are handled through the following Lemma G.2.

Lemma G.2 (Conditional first moments for the current FSO batch)**.** *Under Assumption 2.3, conditionally on $\mathcal{F}_{t,h}^-$,*

$$\mathbb{E}\left[\|\widehat\mu_h^t-\mu_h^t\|_1\mid\mathcal{F}_{t,h}^-\right]\le\sqrt{\frac{|\mathcal{X}|}{N}}. \tag{G.7}$$

Define the conditional mean of the empirical joint distribution after the state batch by

$$\bar{\mathcal{P}}_h^t:=\frac{1}{N}\sum_{n=1}^N\delta_{x_h^{t,n}}\otimes\nu^{p_h^t,\theta_h^*}(\cdot\mid x_h^{t,n};s_h^t).$$

Conditionally on $\mathcal{F}_{t,h}^X$,

$$\mathbb{E}\left[\|\widehat{\mathcal{P}}_h^t-\bar{\mathcal{P}}_h^t\|_1\mid\mathcal{F}_{t,h}^X\right]\le\sqrt{\frac{|\mathcal{Z}|}{N}}, \tag{G.8}$$

and hence, conditionally on $\mathcal{F}_{t,h}^-$,

$$\mathbb{E}\left[\|\widehat{\mathcal{P}}_h^t-\mathcal{P}_h^t\|_1\mid\mathcal{F}_{t,h}^-\right]\le\sqrt{\frac{|\mathcal{Z}|}{N}}+\sqrt{\frac{|\mathcal{X}|}{N}}. \tag{G.9}$$

Proof of Lemma G.2. For the marginal empirical distribution, Cauchy–Schwarz and the multinomial variance identity give

$$\mathbb{E}\Big[\|\widehat{\mu}_h^t - \mu_h^t\|_1 \mid \mathcal{F}_{t,h}^-\Big] \le \sqrt{|\mathcal{X}| \sum_{x\in\mathcal{X}} \mathbb{V}(\widehat{\mu}_h^t(x) \mid \mathcal{F}_{t,h}^-)} \le \sqrt{|\mathcal{X}|/N}.$$

Conditionally on $\mathcal{F}_{t,h}^X$, the random vectors $\delta_{(x_h^{t,n}, b_h^{t,n})}$ are independent and have conditional means $\delta_{x_h^{t,n}} \otimes \nu^{p_h^t,\theta_h^*}(\cdot \mid x_h^{t,n}; s_h^t)$. The same Cauchy–Schwarz/variance argument on the finite set $\mathcal{Z}$ proves (G.8). Moreover, for the fixed realized policy p_h^t,

$$\|\bar{\mathcal{P}}_h^t - \mathcal{P}_h^t\|_1 = \|\widehat{\mu}_h^t - \mu_h^t\|_1,$$

because both joint distributions use the same conditional response policy and differ only in their $\mathcal{X}$-marginal. The triangle inequality, the tower property, and (G.7)–(G.8) yield (G.9). □

Truncation using the simplex diameter bound, together with the tower property, gives the current-batch conditional first-moment bounds

$$\mathbb{E}^t\big[\|\widehat{\mu}_h^t - \mu_h^t\|_1\big] \le \min\left\{2, \sqrt{\frac{|\mathcal{X}|}{N}}\right\} \le \bar{\epsilon}_N,$$

$$\mathbb{E}^t\Big[\big\|\widehat{\mathcal{P}}_h^t - \mathcal{P}_h^t\big\|_1\Big] \le \min\left\{2, \sqrt{\frac{|\mathcal{Z}|}{N}} + \sqrt{\frac{|\mathcal{X}|}{N}}\right\} \le \bar{\epsilon}_N.$$

Lemma G.3 (Empirical concentration)**.** *Suppose Assumption 2.3 holds. Define the errors*

$$\epsilon_{\mu,N} := \sqrt{\frac{2|\mathcal{X}|\log 2 + 2\log(4HT/\zeta)}{N}}, \qquad \epsilon_{\mathrm{act},N} := \sqrt{\frac{|\mathcal{Z}|}{N}} + \sqrt{\frac{2\log(4HT/\zeta)}{N}}, \tag{G.10}$$

$$\epsilon_N := \epsilon_{\mu,N} + \epsilon_{\mathrm{act},N}. \tag{G.11}$$

Let $\mathcal{E}_{\mathrm{emp}}$ denote the event on which $\|\widehat{\mu}_h^t - \mu_h^t\|_1 \le \epsilon_{\mu,N}$ and $\big\|\widehat{\mathcal{P}}_h^t - \mathcal{P}_h^t\big\|_1 \le \epsilon_N$ hold for all $(h,t) \in [H]\times[T]$. Then $\mathbb{P}(\mathcal{E}_{\mathrm{emp}}) \ge 1-\zeta$. For notational economy, ϵ_N serves as a common untruncated empirical scale because $\|\widehat{\mu}_h^t - \mu_h^t\|_1 \le \epsilon_{\mu,N} \le \epsilon_N$, whereas downstream finite-sample residuals use the truncated scale $\bar{\epsilon}_N$ defined immediately below.

To truncate the empirical scale at the universal ℓ_1-diameter of a probability simplex, set $\bar{\epsilon}_N := \min\{\epsilon_N, 2\}$ and fix the universal numerical constant $C_\epsilon := 5$. Since $|\mathcal{X}| \le |\mathcal{Z}|$, $\zeta \le 1/5$, and $H, T \ge 1$, (G.10) and (G.11) give

$$\bar{\epsilon}_N \le \epsilon_N \le C_\epsilon \sqrt{\frac{|\mathcal{Z}| + \log(HT/\zeta)}{N}}. \tag{G.12}$$

Proof of Lemma G.3. Fix (t,h) and condition on the past trajectory and the current (s_h^t, μ_h^t).

State marginal. The variables $x_h^{t,n}$ are i.i.d. from μ_h^t. The standard ℓ_1 empirical-multinomial bound (Weissman et al., 2003) gives

$$\mathbb{P}\big(\big\|\widehat{\mu}_h^t - \mu_h^t\big\|_1 > \epsilon_{\mu,N}\big) \le 2^{|\mathcal{X}|}\exp(-N\epsilon_{\mu,N}^2/2) \le \zeta/(4HT).$$

Conditional action noise. Define

$$\bar{\mathcal{P}}_h^t(x,b) := \widehat{\mu}_h^t(x)\,\nu^{p_h^t,\theta_h^*}(b \mid x; s_h^t).$$

Conditional on $\mathcal{F}_{t,h}^X$, the empirical mean field $\widehat{\mu}_h^t$ and the selected local policy $p_h^t = g_h^t(s_h^t, \widehat{\mu}_h^t)$ are measurable, while the action variables $\{b_h^{t,n}\}_{n=1}^N$ are mutually conditionally independent by Assumption 2.3. More precisely, for every $n \in [N]$,

$$b_h^{t,n} \mid \mathcal{F}_{t,h}^X \sim \nu^{p_h^t,\theta_h^*}(\cdot \mid x_h^{t,n}; s_h^t).$$

Define the centered vectors

$$Y_n(x,b) := \mathbf{1}\{x_h^{t,n} = x\}\Big(\mathbf{1}\{b_h^{t,n} = b\} - \nu^{p_h^t,\theta_h^*}(b \mid x_h^{t,n}; s_h^t)\Big).$$

By construction $\mathbb{E}[Y_n \mid \mathcal{F}_{t,h}^X] = 0$, and $\widehat{\mathcal{P}}_h^t - \bar{\mathcal{P}}_h^t = N^{-1}\sum_{n=1}^N Y_n$. Since the $b_h^{t,n}$ are conditionally independent given $\mathcal{F}_{t,h}^X$, the Y_n are conditionally pairwise uncorrelated: $\mathbb{E}[Y_n Y_m^\top \mid \mathcal{F}_{t,h}^X] = 0$ for $n \neq m$. The conditional pairwise uncorrelatedness yields the second-moment bound below, while the full conditional independence is used for the McDiarmid tail bound.

By Jensen's and Cauchy–Schwarz inequalities applied conditionally,

$$\mathbb{E}\big[\big\|N^{-1}\textstyle\sum_n Y_n\big\|_1 \mid \mathcal{F}_{t,h}^X\big] \;\le\; \sqrt{|\mathcal{Z}|}\,\mathbb{E}\big[\big\|N^{-1}\textstyle\sum_n Y_n\big\|_2 \mid \mathcal{F}_{t,h}^X\big] \;\le\; \sqrt{|\mathcal{Z}|/N},$$

where the second inequality uses the conditional pairwise uncorrelatedness and the conditional second-moment bound

$$\mathbb{E}\big[\|Y_n\|_2^2 \mid \mathcal{F}_{t,h}^X\big] = \mathbb{E}_{b\sim q_n}\big[\|e_b - q_n\|_2^2\big] = 1 - \|q_n\|_2^2 \le 1,$$

where $q_n := \nu^{p_h^t,\theta_h^*}(\cdot \mid x_h^{t,n}; s_h^t)$. Hence

$$\mathbb{E}\big[\|N^{-1}\textstyle\sum_n Y_n\|_2^2 \mid \mathcal{F}_{t,h}^X\big] = N^{-2}\textstyle\sum_{n=1}^N \mathbb{E}\big[\|Y_n\|_2^2 \mid \mathcal{F}_{t,h}^X\big] \le 1/N.$$

For the tail bound, we apply McDiarmid's inequality conditionally on $\mathcal{F}_{t,h}^X$. Viewing $\big\|\widehat{\mathcal{P}}_h^t - \bar{\mathcal{P}}_h^t\big\|_1 = \big\|N^{-1}\sum_n Y_n\big\|_1$ as a function of the conditionally independent variables $(b_h^{t,1},\dots,b_h^{t,N})$, the bounded-difference constant per coordinate is at most $2/N$, because changing one action shifts $\widehat{\mathcal{P}}_h^t$ by at most $2/N$ in ℓ_1. McDiarmid's inequality yields

$$\mathbb{P}\left(\left\|\widehat{\mathcal{P}}_h^t - \bar{\mathcal{P}}_h^t\right\|_1 > \epsilon_{\mathrm{act},N} \mid \mathcal{F}_{t,h}^X\right) \;\le\; \exp\left(-N(\epsilon_{\mathrm{act},N} - \sqrt{|\mathcal{Z}|/N})^2/2\right) \;\le\; \zeta/(4HT)$$

by the choice (G.10). Removing the conditioning by the tower property preserves the bound.

Adaptive target. Finally,

$$\big\|\bar{\mathcal{P}}_h^t - \mathcal{P}_h^t\big\|_1 = \sum_{x\in\mathcal{X}}\big|\widehat{\mu}_h^t(x) - \mu_h^t(x)\big|\sum_{b\in\mathcal{B}}\nu^{p_h^t,\theta_h^*}(b \mid x; s_h^t) = \big\|\widehat{\mu}_h^t - \mu_h^t\big\|_1.$$

Thus $\left\|\widehat{\mathcal{P}}_h^t - \mathcal{P}_h^t\right\|_1 \le \epsilon_{\mathrm{act},N} + \epsilon_{\mu,N} = \epsilon_N$ on the intersection of the two events. A union bound over (t,h) completes the proof. □

G.4 MLE CONCENTRATION AND FISHER–TRAJECTORY EQUIVALENCE

Remark G.4 (Unregularized MLE, regularized confidence geometry). The negative log-likelihood minimized in (3.3) is not ridge-regularized; the MLE is constrained only by the compact parameter domain Θ. The ridge term $\lambda_\theta I_{d_F}$ is added solely to the empirical Fisher matrix used in the confidence ellipsoid. This guarantees invertibility from the first episode and enables the elliptical-potential argument. Its only statistical cost is the additive radius term $O(\lambda_\theta B_\Theta^2)$, since both $\widehat{\theta}_h^t$ and θ_h^* lie in $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$. In particular, the confidence ellipsoid Θ_h^t is deterministically nonempty because its center $\widehat{\theta}_h^t \in \Theta$ satisfies $\|\widehat{\theta}_h^t - \widehat{\theta}_h^t\|_{\widehat{F}_h^t} = 0 \le \sqrt{\beta_{\mathrm{F}}^t}$.

Lemma G.5 (MLE concentration)**.** *For the response model in (C.2) and the observation protocol in Assumption 2.3, there exists an event $\mathcal{E}_{\mathrm{MLE,main}}$ with probability at least $1-\zeta$ such that, on this event, for all (h,t),*

$$\theta_h^* \in \Theta_h^t,$$

and

$$\beta_{\mathrm{F}}^t = 8C_{\mathrm{sc}}(1+\eta B_A^\Theta)^2\,[1 + \log(H\mathcal{N}_\Theta(\varepsilon_{\mathrm{net}})/\zeta)] + 4C_{\mathrm{sc}}\lambda_\theta B_\Theta^2 = \widetilde{O}(d_F), \tag{G.13}$$

$$\big\|\theta_h^* - \widehat{\theta}_h^t\big\|_{\widehat{F}_h^t}^2 \le \beta_{\mathrm{F}}^t, \tag{G.14}$$

$$V_h^t := \lambda_\theta I_{d_F} + \eta^2\textstyle\sum_{\tau<t,n}\Sigma_{x_h^{\tau,n},s_h^\tau}^{p_h^\tau,\widehat{\theta}_h^\tau}, \tag{G.15}$$

$$\big\|\theta_h^* - \widehat{\theta}_h^t\big\|_{V_h^t}^2 \le C_{\mathrm{sc}}\,\beta_{\mathrm{F}}^t. \tag{G.16}$$

Here $C_{\mathrm{sc}} = e^{4\eta B_\Theta}$ is the deterministic two-sided softmax Fisher comparison constant in (C.18), B_A^Θ is the bound uniform over candidate parameters in (C.17), and $\varepsilon_{\mathrm{net}}$ is the exact covering scale in (G.19):

$$\varepsilon_{\mathrm{net}} = \frac{1}{1 + TN[\eta(1 + B_A^\Theta)B_\varphi + \bar{L}_H]}, \qquad \bar{L}_H := 2\eta B_\varphi + 8\eta^2(1 + \eta B_A^\Theta)^2 B_\varphi^2.$$

This scale controls the combined likelihood and Hellinger discretization error by one. The two-sided PSD comparison between the empirical Fisher matrix $\widehat{F}_h^t$ and the auxiliary additive Fisher matrix V_h^t follows from Lemma C.31. The event $\mathcal{E}_{\mathrm{emp}}$ is not needed for this well-specified MLE result and is combined with $\mathcal{E}_{\mathrm{MLE,main}}$ only in the FSO theorem.

Uniformity in t. The right-hand side of (G.13) does not depend on the episode index t. The T dependence enters through the exact scale $\varepsilon_{\mathrm{net}} = \big(1 + TN[\eta(1 + B_A^\Theta)B_\varphi + \bar{L}_H]\big)^{-1}$ and the Euclidean-ball bound $\mathcal{N}_\Theta(\epsilon) \le (1 + 2B_\Theta/\epsilon)^{d_F}$. There is no additional $\log T$ factor from a union bound over t, because Ville's inequality is uniform in time. Hence $\beta_{\mathrm{F}}^t = \beta_{\mathrm{F}}^T$ numerically for every $t \le T$. The two notations are retained to distinguish the radius used at episode t from its common worst-case value in the theorem statement.

Proof of Lemma G.5. The proof follows the adaptive softmax-MLE concentration template of Chen et al. (2023a, Lemma B.5 and Appendix F.4), generalized to multi-sample $N \ge 1$ per episode. We give the full structure with explicit filtration handling.

Filtration setup. We use the global within-step sigma-field $\mathcal{F}_{\tau,h}^X$ defined in (G.6). The pre-batch field $\mathcal{F}_{\tau,h}^-$ is generated by the complete history immediately before the state batch $x_h^{\tau,1:N}$ is drawn and satisfies $\sigma(s_h^\tau, \mu_h^\tau) \subseteq \mathcal{F}_{\tau,h}^-$ for the current system state (s_h^τ, μ_h^τ). The current state batch is $x_h^{\tau,1:N} = (x_h^{\tau,1}, \dots, x_h^{\tau,N})$, and its empirical measure is $\widehat{\mu}_h^\tau = N^{-1}\sum_{n=1}^N \delta_{x_h^{\tau,n}}$. The realized maximizing pair is $(p_h^\tau, \bar{\theta}_h^\tau) = \big(p^\tau_{h,x_h^{\tau,1:N}}(s_h^\tau), \bar{\theta}^\tau_{h,x_h^{\tau,1:N}}(s_h^\tau)\big)$, as defined through (G.4). The generators in (G.6) do not include the follower-action batch $b_h^{\tau,1:N} = (b_h^{\tau,1}, \dots, b_h^{\tau,N})$. Conditional on $\mathcal{F}_{\tau,h}^X$:

(i) $p_h^\tau = g_h^\tau(s_h^\tau, \widehat{\mu}_h^\tau)$ is $\mathcal{F}_{\tau,h}^X$-measurable (the realized policy depends on the observed state batch $x_h^{\tau,1:N}$ only through $\widehat{\mu}_h^\tau = N^{-1}\sum_{n=1}^N \delta_{x_h^{\tau,n}}$).

(ii) Conditional on $\mathcal{F}_{\tau,h}^X$, the action observations $\{b_h^{\tau,n}\}_{n=1}^N$ are mutually independent with marginal distribution

$$b_h^{\tau,n} \mid \mathcal{F}_{\tau,h}^X \;\sim\; \nu^{p_h^\tau,\theta_h^*}\big(\cdot \mid x_h^{\tau,n};\, s_h^\tau\big).$$

This conditional action law is the model used in the likelihood objective (3.3), so the MLE is well-specified.

Step 1: Drift-normalized likelihood-ratio supermartingale. For any fixed $\theta \in \Theta$, define the per-sample half-log-likelihood-ratio increment

$$Z_{\tau,n}^h(\theta) := \frac{1}{2}\log\frac{\nu^{p_h^\tau,\theta}(b_h^{\tau,n}|x_h^{\tau,n};s_h^\tau)}{\nu^{p_h^\tau,\theta_h^*}(b_h^{\tau,n}|x_h^{\tau,n};s_h^\tau)},$$
$$D_{H,\tau,n}^{h,2}(\theta) := D_H^2\big(\nu^{p_h^\tau,\theta}(\cdot|x_h^{\tau,n};s_h^\tau),\, \nu^{p_h^\tau,\theta_h^*}(\cdot|x_h^{\tau,n};s_h^\tau)\big).$$

By Assumption 2.3 and (C.2), conditional on $\mathcal{F}_{\tau,h}^X$ and on the prior actions, $b_h^{\tau,n} \sim \nu^{p_h^\tau,\theta_h^*}(\cdot|x_h^{\tau,n};s_h^\tau)$, giving the *exact one-sample Hellinger affinity identity*

$$\mathbb{E}\Big[e^{Z_{\tau,n}^h(\theta)} \,\Big|\, \mathcal{F}_{\tau,n-1}^h\Big] \;=\; 1 - D_{H,\tau,n}^{h,2}(\theta) \;\le\; e^{-D_{H,\tau,n}^{h,2}(\theta)}, \tag{G.17}$$

where $\mathcal{F}_{\tau,n-1}^h$ is the per-sample filtration defined by

$$\mathcal{F}_{\tau,n-1}^h := \sigma\Big(\bigcup_{\tau'<\tau}\sigma\big(\mathcal{F}_{\tau',h}^X,\, b_h^{\tau',1:N}\big) \cup \mathcal{F}_{\tau,h}^X \cup \sigma\big(b_h^{\tau,1},\dots,b_h^{\tau,n-1}\big)\Big),$$

with samples ordered lexicographically: $(\tau, n) \prec (\tau', n')$ if $\tau < \tau'$, or $\tau = \tau'$ and $n < n'$. Crucially, $\mathcal{F}^h_{\tau,n-1}$ contains the entire current-step state batch (via $\mathcal{F}^X_{\tau,h} \subseteq \mathcal{F}^h_{\tau,n-1}$), so s^τ_h, $x^{\tau,n}_h$, $\widehat{\mu}^\tau_h$, and $p^\tau_h = g^\tau_h(s^\tau_h, \widehat{\mu}^\tau_h)$ are all $\mathcal{F}^h_{\tau,n-1}$-measurable; conditional on $\mathcal{F}^h_{\tau,n-1}$, the next observation satisfies $b^{\tau,n}_h \sim \nu^{p^\tau_h, \theta^*_h}(\cdot|x^{\tau,n}_h; s^\tau_h)$ by Assumption 2.3 and (C.2), which is what makes (G.17) valid. Define the drift-normalized exponential process

$$\widetilde{M}^h_m(\theta) := \exp\Bigg(\sum_{(\tau,n)\preceq m} Z^h_{\tau,n}(\theta) + \sum_{(\tau,n)\preceq m} D^{h,2}_{H,\tau,n}(\theta) \Bigg). \tag{G.18}$$

Then by (G.17), $\widetilde{M}^h_m(\theta)$ is a non-negative supermartingale with $\mathbb{E}[\widetilde{M}^h_m(\theta)] \le 1$:

$$\begin{aligned} \mathbb{E}\Big[\widetilde{M}^h_m(\theta) \,\Big|\, \mathcal{F}^h_{m-1}\Big] &= \widetilde{M}^h_{m-1}(\theta) \cdot e^{D^{h,2}_{H,m}(\theta)} \cdot \mathbb{E}[e^{Z^h_m(\theta)} | \mathcal{F}^h_{m-1}] \\ &\le \widetilde{M}^h_{m-1}(\theta) \cdot e^{D^{h,2}_{H,m}(\theta)} \cdot e^{-D^{h,2}_{H,m}(\theta)} = \widetilde{M}^h_{m-1}(\theta). \end{aligned}$$

This drift-normalization is the key technical step: it explicitly converts the conditional Hellinger affinity into a Hellinger drift inside the supermartingale exponent, so that bounding the drift-normalized process directly bounds the cumulative Hellinger sum.

Step 2: Generalized likelihood-ratio (GLR) inequality and Hellinger-drift extraction. We extend the pointwise drift-normalized inequality uniformly to all $\theta \in \Theta$ by an ε_{net} covering argument with explicit Lipschitz control. Let $\Theta_{\varepsilon_{\text{net}}} \subset \Theta$ be a minimal Euclidean ε_{net} cover of $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$. The all-scale volume bound gives $|\Theta_{\varepsilon_{\text{net}}}| \le \mathcal{N}_\Theta(\varepsilon_{\text{net}}) \le (1 + 2B_\Theta/\varepsilon_{\text{net}})^{d_F}$ for every $\varepsilon_{\text{net}} > 0$.

Lipschitz control of the likelihood and Hellinger components. For any $\theta, \theta' \in \Theta$, the per-sample log-likelihood ratio increment satisfies (using the softmax logit gradient bound $\big\|\nabla_\theta \log \nu^{p,\theta}(b|x;s)\big\|_2 = \eta \, \|\varphi^p(s,x,b) - \mathbb{E}_{B\sim\nu^{p,\theta}} \varphi^p(s,x,B)\|_2 \le 2\eta B_\varphi$ by direct softmax differentiation and the inequality $2\eta B_\varphi \le 2\eta(1 + B^\Theta_A)B_\varphi$):

$$\big|Z^h_{\tau,n}(\theta) - Z^h_{\tau,n}(\theta')\big| \le \eta(1 + B^\Theta_A)\, B_\varphi \, \|\theta - \theta'\|_2 \,,$$

where the slightly looser $\eta(1 + B^\Theta_A)B_\varphi$ rate is retained for consistency with the ε_{net} scale. Hence $|\sum_{(\tau,n)\preceq m}[Z(\theta) - Z(\theta')]| \le m\,\eta(1+B^\Theta_A)B_\varphi \, \|\theta - \theta'\|_2 \le TN\,\eta(1+B^\Theta_A)B_\varphi \, \|\theta - \theta'\|_2$. Similarly, the per-sample squared Hellinger distance is L_H-Lipschitz in θ with $L_H \le 2\eta B_\varphi$ by Lemma G.6. Hence

$$\Big| \sum_{(\tau,n)\preceq m} \big[D^{h,2}_{H,\tau,n}(\theta) - D^{h,2}_{H,\tau,n}(\theta')\big] \Big| \le TN\, L_H \, \|\theta - \theta'\|_2 \,.$$

To control the combined likelihood and Hellinger discretization error, we choose the covering scale

$$\varepsilon_{\text{net}} := \frac{1}{1 + TN[\eta(1 + B^\Theta_A)B_\varphi + \bar{L}_H]}, \qquad \bar{L}_H := 2\eta B_\varphi + 8\eta^2(1 + \eta B^\Theta_A)^2 B^2_\varphi, \tag{G.19}$$

where $\bar{L}_H$ is the envelope upper bound on L_H established in Lemma G.6. The combined likelihood and Hellinger discretization error is at most $TN[\eta(1 + B^\Theta_A)B_\varphi + \bar{L}_H]\varepsilon_{\text{net}} \le 1$.

Pointwise Ville inequality plus union bound over the net. For each fixed $\theta \in \Theta_{\varepsilon_{\text{net}}}$ and each $h \in [H]$, Ville's inequality applied to the supermartingale $(\widetilde{M}^h_m(\theta))_m$ gives, *automatically uniformly over all time indices $m = (\tau, n)$ with $\tau \le T$* (hence uniformly over $t \in [T]$ via the choice $m = (t-1, N)$),

$$\mathbb{P}\Big[\sup_m \log \widetilde{M}^h_m(\theta) > \log\big(H\mathcal{N}_\Theta(\varepsilon_{\text{net}})/\zeta\big)\Big] \le \frac{\zeta}{H\mathcal{N}_\Theta(\varepsilon_{\text{net}})}.$$

A union bound over $h \in [H]$ and $\theta \in \Theta_{\varepsilon_{\text{net}}}$ alone (no further union over t is needed — the uniform-in-time guarantee comes for free from $\sup_m$) gives, with probability at least $1 - \zeta$, that for all $\theta \in \Theta_{\varepsilon_{\text{net}}}$ and all $(h, t) \in [H] \times [T]$,

$$\log \widetilde{M}^h_{(t-1),N}(\theta) \le \log\big(H\mathcal{N}_\Theta(\varepsilon_{\text{net}})/\zeta\big).$$

Extending from the net to every $\theta \in \Theta$ then gives, with probability at least $1 - \zeta$, simultaneously for every (h, t, θ),

$$\log \widetilde{M}^h_{(t-1),N}(\theta) \le 1 + \log(H\mathcal{N}_\Theta(\varepsilon_{\text{net}})/\zeta) \,.$$

Substituting (G.18) and rearranging gives

$$\sum_{\tau=1}^{t-1}\sum_{n=1}^{N} D^{h,2}_{H,\tau,n}(\theta) \le -\sum_{\tau=1}^{t-1}\sum_{n=1}^{N} Z^h_{\tau,n}(\theta) + 1 + \log\left(\frac{H\mathcal{N}_\Theta(\varepsilon_{\text{net}})}{\zeta}\right). \tag{G.20}$$

Now specialize to the data-dependent $\theta = \widehat{\theta}^t_h$, which is permitted because the inequality is uniform on Θ. The half-log-ratio identity and the definition of the sum-form negative log-likelihood give

$$-\sum_{\tau=1}^{t-1}\sum_{n=1}^{N} Z^h_{\tau,n}(\theta) = \frac{1}{2}\left[\mathcal{L}^t_h(\theta) - \mathcal{L}^t_h(\theta^*_h)\right].$$

Because $\widehat{\theta}^t_h$ minimizes $\mathcal{L}^t_h$ over Θ and $\theta^*_h \in \Theta$, this likelihood difference is nonpositive. Consequently,

$$\sum_{\tau=1}^{t-1}\sum_{n=1}^{N} D^2_H\left(\nu^{p^\tau_h,\widehat{\theta}^t_h}(\cdot|x^{\tau,n}_h; s^\tau_h), \nu^{p^\tau_h,\theta^*_h}(\cdot|x^{\tau,n}_h; s^\tau_h)\right) \le 1 + \log\left(\frac{H\mathcal{N}_\Theta(\varepsilon_{\text{net}})}{\zeta}\right). \tag{G.21}$$

This is the *cumulative Hellinger bound at the MLE*, established rigorously via drift-normalized GLR with full uniformization. The drift normalization is essential: without the $+\sum D^2_H$ term in (G.18), Markov's inequality on the raw likelihood ratio would only control $\sum Z_{\tau,n}$, not the Hellinger drift $\sum D^2_H$.

Step 3: Hellinger-to-Fisher conversion (with explicit Hellinger symmetrization). The stated form of Lemma H.8 (equivalently Chen et al. (2023a, Lemma G.2), a softmax-specific Hellinger lower bound) states, for any $\theta_1, \theta_2 \in \Theta$ and any (p, x, s):

$$D^2_H\left(\nu^{p,\theta_1}(\cdot|x; s), \nu^{p,\theta_2}(\cdot|x; s)\right) \;\ge\; \frac{\eta^2}{8(1+\eta B^\Theta_A)^2} \cdot \|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_1}_{x,s}}, \tag{G.22}$$

where the Fisher matrix is at the parameter of the *first* argument θ_1 (cf. the inner-product expectation under ν^{p,θ_1} inside Lemma H.8, mirroring the use of $\Sigma^{p,\theta^*}_{x,s}$ in the softmax comparison (H.27), where θ^* is the first argument). By the symmetry of the Hellinger distance, $D^2_H(P, Q) = D^2_H(Q, P)$; applying (G.22) to the swapped pair $(\nu^{p,\theta_2}, \nu^{p,\theta_1})$ and using $\|\theta_2 - \theta_1\|^2 = \|\theta_1 - \theta_2\|^2$ yields the equivalent symmetrized form (with Σ at the *second* argument θ_2):

$$\begin{aligned} D^2_H\left(\nu^{p,\theta_1}(\cdot|x; s), \nu^{p,\theta_2}(\cdot|x; s)\right) \;&\ge\; \frac{1}{8(\eta^{-1}+B^\Theta_A)^2} \cdot \|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}} \\ &=\; \frac{\eta^2}{8(1+\eta B^\Theta_A)^2} \cdot \|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}}. \end{aligned} \tag{G.23}$$

We adopt the symmetrized form (G.23) because we will set $\theta_2 = \theta^*_h$ below, aligning the cumulative Fisher matrix on the right-hand side with the analytic-design covariance $V^{t,*}_h := \lambda_\theta I_{d_F} + \eta^2 \sum_{\tau,n} \Sigma^{p^\tau_h,\theta^*_h}_{x^{\tau,n}_h,s^\tau_h}$. (This first-vs.-second-argument symmetry swap parallels the analogous step inside Lemma G.17 Stage 2.)

Applying (G.23) with $\theta_1 = \widehat{\theta}^t_h$ and $\theta_2 = \theta^*_h$, summing over (τ, n), and using (G.21) gives

$$\sum_{\tau=1}^{t-1}\sum_{n=1}^{N} \left\|\widehat{\theta}^t_h - \theta^*_h\right\|^2_{\Sigma^{p^\tau_h,\theta^*_h}_{x^{\tau,n}_h,s^\tau_h}} \le \frac{8(1+\eta B^\Theta_A)^2}{\eta^2}\left[1 + \log\left(\frac{H\mathcal{N}_\Theta(\varepsilon_{\text{net}})}{\zeta}\right)\right].$$

Multiplying both sides by η^2 and adding the regularization contribution $\lambda_\theta\|\widehat{\theta}^t_h - \theta^*_h\|^2_2 \le 4\lambda_\theta B^2_\Theta$ yields

$$\left\|\widehat{\theta}^t_h - \theta^*_h\right\|^2_{V^{t,*}_h} \le 8(1+\eta B^\Theta_A)^2\left[1 + \log\left(\frac{H\mathcal{N}_\Theta(\varepsilon_{\text{net}})}{\zeta}\right)\right] + 4\lambda_\theta B^2_\Theta. \tag{G.24}$$

Here $V^{t,*}_h := \lambda_\theta I_{d_F} + \eta^2 \sum_{\tau,n} \Sigma^{p^\tau_h,\theta^*_h}_{x^{\tau,n}_h,s^\tau_h}$. The η^2 in $V^{t,*}_h$ cancels the inverse η^2 in the Hellinger-to-Fisher conversion. The dimension dependence comes from the all-scale cover $\log\mathcal{N}_\Theta(\varepsilon_{\text{net}}) \le d_F \log(1 + 2B_\Theta/\varepsilon_{\text{net}})$.

Step 4: Self-concordance conversion to the algorithmic $\widehat{F}_h^t$. The algorithmic confidence ellipsoid (3.5) uses $\widehat{F}_h^t$ (cumulative empirical Fisher evaluated at the current MLE $\widehat{\theta}_h^t$). By Lemma G.1, for any $\theta', \theta'' \in \Theta$,

$$e^{-2\eta\|\theta'-\theta''\|_2}\,\Sigma_{x,s}^{p,\theta'} \preceq \Sigma_{x,s}^{p,\theta''} \preceq e^{2\eta\|\theta'-\theta''\|_2}\,\Sigma_{x,s}^{p,\theta'}.$$

Since $\|\theta_h^* - \widehat{\theta}_h^t\|_2 \le 2B_\Theta$ and $C_{\mathrm{sc}} = e^{4\eta B_\Theta}$ by (C.18), applying the sandwich to every cumulative term gives

$$C_{\mathrm{sc}}^{-1} V_h^{t,*} \preceq \widehat{F}_h^t \preceq C_{\mathrm{sc}} V_h^{t,*}.$$

Hence $\left\|\widehat{\theta}_h^t - \theta_h^*\right\|_{\widehat{F}_h^t}^2 \le C_{\mathrm{sc}} \left\|\widehat{\theta}_h^t - \theta_h^*\right\|_{V_h^{t,*}}^2$. Combining this inequality with (G.24) gives

$$\left\|\widehat{\theta}_h^t - \theta_h^*\right\|_{\widehat{F}_h^t}^2 \le 8C_{\mathrm{sc}}(1+\eta B_A^\Theta)^2\,[1+\log(H\mathcal{N}_\Theta(\varepsilon_{\mathrm{net}})/\zeta)] + 4C_{\mathrm{sc}}\lambda_\theta B_\Theta^2.$$

This is exactly the displayed radius (G.13), including its numerical factor 8 and additive one. The leading d_F dependence comes from $\log \mathcal{N}_\Theta(\varepsilon_{\mathrm{net}}) \le d_F \log(1 + 2B_\Theta/\varepsilon_{\mathrm{net}})$. The same self-concordance comparison between V_h^t and $V_h^{t,*}$ gives (G.16).

Step 5: Conclusion (no separate union bound over t is needed). By Step 2, with probability at least $1-\zeta$, the inequality (G.20) (hence its downstream consequences in Steps 3–4) holds *simultaneously for all $\theta \in \Theta$ and all $(h,t) \in [H] \times [T]$*. The uniform-in-t guarantee is delivered automatically by Ville's inequality at the supermartingale level ($\sup_m$ in Step 2): the union bound is taken only over $h \in [H]$ and the $\varepsilon_{\mathrm{net}}$-cover $\Theta_{\varepsilon_{\mathrm{net}}}$ (with the Lipschitz extension absorbing the cover-discretization error into the universal constant; cf. Lemma G.6), *not* over $t \in [T]$. Consequently, the displayed validity (G.14) holds with overall failure probability at most ζ. The proof relies on the well-specified conditional likelihood structure (Step 1), the GLR concentration (Steps 2–3), and the standard self-concordance sandwich (Step 4). □

Proof of Lemma C.31. The proof is in three short steps.

Step 1: Empirical aggregation. Define the empirical $\widehat{\mu}$-weighted aggregate $\widehat{\bar{\Sigma}}$ by the following equivalent representations:

$$N \cdot \widehat{\bar{\Sigma}}_{s_h^\tau, \widehat{\mu}_h^\tau}^{p_h^\tau, \widehat{\theta}_h^t} = N \sum_{x \in \mathcal{X}} \widehat{\mu}_h^\tau(x)\, \Sigma_{x,s_h^\tau}^{p_h^\tau, \widehat{\theta}_h^t} = \sum_{n=1}^{N} \Sigma_{x_h^{\tau,n}, s_h^\tau}^{p_h^\tau, \widehat{\theta}_h^t},$$

since $\widehat{\mu}_h^\tau = N^{-1}\sum_n \delta_{x_h^{\tau,n}}$. Hence, the empirical Fisher matrix in (3.4) satisfies

$$\widehat{F}_h^t = \lambda_\theta I_{d_F} + \eta^2 \sum_{\tau=1}^{t-1} \sum_{n=1}^{N} \Sigma_{x_h^{\tau,n}, s_h^\tau}^{p_h^\tau, \widehat{\theta}_h^t}.$$

This expression has the same form as V_h^t in (G.15), except that each historical term is evaluated at the *current* MLE $\widehat{\theta}_h^t$ rather than at the *contemporaneous* MLE $\widehat{\theta}_h^\tau$.

Step 2: Pointwise self-concordance sandwich on each historical term. By Lemma G.1, for any $\theta', \theta'' \in \Theta$ (so $\|\theta' - \theta''\|_2 \le 2B_\Theta$ deterministically by the closed-ball form of Θ) and any (p, x, s),

$$e^{-4\eta B_\Theta}\,\Sigma_{x,s}^{p,\theta'} \preceq \Sigma_{x,s}^{p,\theta''} \preceq e^{4\eta B_\Theta}\,\Sigma_{x,s}^{p,\theta'}.$$

Specializing to $\theta' = \widehat{\theta}_h^t$ and $\theta'' = \widehat{\theta}_h^\tau$ for each $\tau < t$ and each (x,s) along the historical data, and using $C_{\mathrm{sc}} = e^{4\eta B_\Theta}$ from (C.18),

$$C_{\mathrm{sc}}^{-1}\,\Sigma_{x_h^{\tau,n}, s_h^\tau}^{p_h^\tau, \widehat{\theta}_h^\tau} \preceq \Sigma_{x_h^{\tau,n}, s_h^\tau}^{p_h^\tau, \widehat{\theta}_h^t} \preceq C_{\mathrm{sc}}\,\Sigma_{x_h^{\tau,n}, s_h^\tau}^{p_h^\tau, \widehat{\theta}_h^\tau},$$

deterministically.

Step 3: Sum and add regularizer. Summing the Step 2 sandwich over $\tau \in [t-1]$ and $n \in [N]$ (PSD sandwich is preserved under summation), multiplying by η^2, and adding $\lambda_\theta I_{d_F}$ (PSD sandwich is preserved under additive PSD shift, as $\lambda_\theta I_{d_F} \preceq \lambda_\theta I_{d_F} \preceq \lambda_\theta I_{d_F}$ with constant 1):

$$\lambda_\theta I + \eta^2 C_{\mathrm{sc}}^{-1} \sum_{\tau,n} \Sigma^{\widehat{\theta}_h^\tau} \preceq \lambda_\theta I + \eta^2 \sum_{\tau,n} \Sigma^{\widehat{\theta}_h^t} \preceq \lambda_\theta I + \eta^2 C_{\mathrm{sc}} \sum_{\tau,n} \Sigma^{\widehat{\theta}_h^\tau}.$$

Using $C_{\mathrm{sc}}^{-1} \le 1 \le C_{\mathrm{sc}}$ on the regularizer term (i.e., $C_{\mathrm{sc}}^{-1}\lambda_\theta I \preceq \lambda_\theta I \preceq C_{\mathrm{sc}}\lambda_\theta I$), this can be rewritten as

$$C_{\mathrm{sc}}^{-1}\, V_h^t \;\preceq\; \widehat{F}_h^t \;\preceq\; C_{\mathrm{sc}}\, V_h^t,$$

which is (C.39). □

Lemma G.6 (Net stability of the likelihood and Hellinger objective). *Fix $h \in [H]$. For $t \ge 2$ and any fixed history up to episode $t-1$, the negative log-likelihood $\mathcal{L}_h^t$ in (3.3) is $(t-1)N\eta(1+B_A^\Theta)B_\varphi$-Lipschitz over Θ in $\|\cdot\|_2$. The cumulative conditional Hellinger contrast*

$$\sum_{\tau=1}^{t-1}\sum_{n=1}^{N} D_H^2\Big(\nu^{p_h^\tau,\theta}(\cdot \mid x_h^{\tau,n}; s_h^\tau), \nu^{p_h^\tau,\theta_h^*}(\cdot \mid x_h^{\tau,n}; s_h^\tau)\Big)$$

is $(t-1)N\bar{L}_H$-Lipschitz in θ, where

$$\bar{L}_H := 2\eta B_\varphi + 8\eta^2(1+\eta B_A^\Theta)^2 B_\varphi^2.$$

The displayed $\bar{L}_H$ is a conservative envelope satisfying $\bar{L}_H \ge 2\eta B_\varphi$ for every $\eta > 0$. We use the convention $D_H^2(p,q) = 1 - \sum_b \sqrt{p(b)q(b)} = \frac{1}{2}\sum_b(\sqrt{p(b)} - \sqrt{q(b)})^2$ for discrete probability vectors. Consequently, the global covering scale $\varepsilon_{\mathrm{net}}$ in (G.19) makes the sum of the two discretization errors at episode t at most $(t-1)/T \le 1$, while

$$\log \mathcal{N}_\Theta(\varepsilon_{\mathrm{net}}) \le d_F \log\big(1 + 2B_\Theta\big(1 + TN[\eta(1+B_A^\Theta)B_\varphi + \bar{L}_H]\big)\big) = \widetilde{O}(d_F).$$

For $t = 1$, the statement is vacuous.

Proof of Lemma G.6. Likelihood Lipschitz constant. For each observation, the per-sample score

$$-\nabla_\theta \log \nu^{p,\theta}(b \mid x; s) \;=\; \eta(\mathbb{E}_{\nu^{p,\theta}}\varphi - \varphi_b)$$

has Euclidean norm bounded by $2\eta\,\|\varphi\|_2 \le 2\eta B_\varphi \le \eta(1+B_A^\Theta)B_\varphi$, the last inequality from $B_A^\Theta \ge 1$. Summing over the $(t-1)N$ samples yields the cumulative Lipschitz bound.

Hellinger Lipschitz constant. For each observation, use the following direct gradient bound.

Simple gradient bound. Using the identity $\nabla_\theta\sqrt{p_\theta(b)} = \frac{\eta}{2}\sqrt{p_\theta(b)}(\varphi_b - \mathbb{E}_{p_\theta}\varphi)$ (from $\partial_\theta p_\theta(b) = p_\theta(b)\eta(\varphi_b - \mathbb{E}\varphi)$ via chain rule) and Cauchy–Schwarz,

$$\big\|\nabla_\theta D_H^2(p_\theta, q)\big\|_2 \le \sqrt{2}\, D_H(p_\theta, q) \cdot \frac{\eta}{2}\sqrt{\sum_b p_\theta(b)\,\|\varphi_b - \mathbb{E}\varphi\|_2^2} \le 2\eta B_\varphi,$$

using $D_H \le 1$ and $\sum_b p_\theta(b)\,\|\varphi_b - \mathbb{E}\varphi\|_2^2 \le 4B_\varphi^2$. This bound is valid for *all* $\eta > 0$.

Combining. The simple gradient bound gives a per-observation Lipschitz constant at most $2\eta B_\varphi \le \bar{L}_H$. Summing over the $(t-1)N$ observations yields the cumulative Hellinger Lipschitz bound.

The covering-number claim follows from the all-scale Euclidean-ball bound $\mathcal{N}_\Theta(\epsilon) \le (1+2B_\Theta/\epsilon)^{d_F}$ on $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$, valid for every $\epsilon > 0$. □

G.5 Auxiliary covariance bounds for the FSO analysis

Lemma G.7 (Empirical-to-population covariance comparison). *Fix $h \in [H]$, $t \in [T]$. Define the empirical Fisher at the data-realized p_h^t and θ^* as in (G.64). By Assumption 2.1, take $B_\varphi = 1$ and set $R_\Sigma = B_\varphi^2$. The centered covariance satisfies*

$$\|\Sigma_{x,s}^{p,\theta}\|_{\mathrm{op}} \le \mathrm{Tr}(\Sigma_{x,s}^{p,\theta}) \le R_\Sigma.$$

On $\mathcal{E}_{\mathrm{emp}}$, deterministically in p_h^t,

$$\Big\|\bar{\Sigma}_{s_h,\mu_h^t}^{p_h^t,\theta_h^*} - \widehat{\Gamma}_h^t(\theta_h^*)\Big\|_{\mathrm{op}} \;\le\; \big\|\mu_h^t - \widehat{\mu}_h^t\big\|_1 \cdot R_\Sigma \;\le\; \epsilon_N \cdot R_\Sigma.$$

Proof of Lemma G.7. For $b \sim \nu^{p,\theta}(\cdot \mid x; s)$, the covariance identity gives

$$\mathrm{Tr}(\Sigma^{p,\theta}_{x,s}) = \mathbb{E}\|\varphi^p(s,x,b)\|_2^2 - \|\mathbb{E}\varphi^p(s,x,b)\|_2^2 \le B_\varphi^2.$$

Since the covariance is positive semidefinite, its operator norm is at most its trace. At the same realized p_h^t, the population and empirical covariances differ only in their state weights:

$$\bar{\Sigma}^{p_h^t,\theta_h^*}_{s_h,\mu_h^t} - \widehat{\Gamma}_h^t(\theta_h^*) = \sum_{x\in\mathcal{X}} [\mu_h^t(x) - \widehat{\mu}_h^t(x)]\Sigma^{p_h^t,\theta_h^*}_{x,s_h^t}.$$

The operator-norm triangle inequality bounds this difference by $\|\mu_h^t - \widehat{\mu}_h^t\|_1 R_\Sigma$, which is at most $\epsilon_N R_\Sigma$ on $\mathcal{E}_{\mathrm{emp}}$. □

G.6 Regret decomposition, optimism, and Bellman/joint-distribution error bounds

Performance difference under the FSO protocol. For every h, let $(s,\mathbf{x},x) \mapsto \widetilde{\nu}_{h,\mathbf{x}}(\cdot \mid x; s)$ be a Borel stochastic kernel from $\mathcal{S}\times\mathcal{X}^N\times\mathcal{X}$ to $\Delta(\mathcal{B})$. For a bounded Borel action-value function U_h, define the batch-averaged planning operator

$$(\mathcal{T}_h^{g,\widetilde{\nu}}U_h)(s,\mu) := \mathbb{E}_{\mathbf{x}\sim\mu^{\otimes N}}\left[\langle U_h(s,\cdot,\mu\otimes\widetilde{\nu}_{h,\mathbf{x}}(\cdot\mid\cdot;s)), g_h(s,\widehat{\mu}_{\mathbf{x}})\rangle_{\mathcal{A}}\right].$$

Lemma G.8 (Performance difference for batch-adaptive FSO commitment rules)**.** *Let $g=\{g_h\}_{h=1}^H$ be an admissible FSO commitment rule. For every h, let $(s,\mathbf{x},x)\mapsto\widetilde{\nu}_{h,\mathbf{x}}(\cdot\mid x;s)$ be the Borel stochastic kernel above, evaluated at $p_{\mathbf{x}} = g_h(s,\widehat{\mu}_{\mathbf{x}})$. Let $\widetilde{U}=\{\widetilde{U}_h\}_{h=1}^H$ and $\widetilde{W}=\{\widetilde{W}_h\}_{h=1}^{H+1}$ be bounded measurable functions with $\widetilde{W}_{H+1}\equiv 0$. Along the realized trajectory, let $x_h^{1:N}$ be the current follower-state batch and define*

$$p_h := g_h(s_h,\widehat{\mu}_{x_h^{1:N}}),\quad \mathcal{P}_h^* := \mu_h\otimes\nu^{p_h,\theta_h^*}(\cdot\mid\cdot;s_h),\quad \widetilde{\mathcal{P}}_h := \mu_h\otimes\widetilde{\nu}_{h,x_h^{1:N}}(\cdot\mid\cdot;s_h).$$

Then

$$\begin{aligned}\widetilde{W}_1(s_0,\mu_0) - J_{\mathrm{FSO}}(g) &= \sum_{h=1}^H \mathbb{E}^g\left[(\widetilde{U}_h - u_h)(s_h,a_h,\mathcal{P}_h^*) - \widetilde{W}_{h+1}(s_{h+1},\mu_{h+1})\right]\\ &+\sum_{h=1}^H \mathbb{E}^g\left[\widetilde{W}_h(s_h,\mu_h) - (\mathcal{T}_h^{g,\widetilde{\nu}}\widetilde{U}_h)(s_h,\mu_h)\right] + \sum_{h=1}^H \mathbb{E}^g\left[\widetilde{U}_h(s_h,a_h,\widetilde{\mathcal{P}}_h) - \widetilde{U}_h(s_h,a_h,\mathcal{P}_h^*)\right].\end{aligned} \tag{G.25}$$

The candidate and true response distributions use the same realized current batch and the same realized local policy. The expectation is taken under the complete FSO trajectory law, including all within-step random variables generated by the protocol. In particular, the batch-dependent local policy p_h is never replaced by its conditional mean.

Proof of Lemma G.8. Let $\mathcal{G}_h^-$ denote the trajectory σ-field immediately before the current state batch is drawn, including the current system state (s_h,μ_h). Conditionally on $\mathcal{G}_h^-$, the protocol first draws $x_h^{1:N}\sim\mu_h^{\otimes N}$, then sets $p_h = g_h(s_h,\widehat{\mu}_{x_h^{1:N}})$, and finally draws $a_h\sim p_h$. Because the empirical-measure map, the commitment rule, and the candidate stochastic kernel are Borel measurable, the batch-averaged operator is Borel measurable and

$$\mathbb{E}^g\left[\widetilde{U}_h(s_h,a_h,\widetilde{\mathcal{P}}_h)\,\middle|\,\mathcal{G}_h^-\right] = (\mathcal{T}_h^{g,\widetilde{\nu}}\widetilde{U}_h)(s_h,\mu_h)\qquad\text{a.s.} \tag{G.26}$$

Under the actual trajectory law, the next system state is generated from $\mathbb{P}_h(\cdot\mid s_h,a_h,\mathcal{P}_h^*)$. The value recursion defining $J_{\mathrm{FSO}}(g)$ and finite-horizon telescoping therefore give

$$\widetilde{W}_1(s_0,\mu_0) - J_{\mathrm{FSO}}(g) = \sum_{h=1}^H \mathbb{E}^g\left[\widetilde{W}_h(s_h,\mu_h) - u_h(s_h,a_h,\mathcal{P}_h^*) - \widetilde{W}_{h+1}(s_{h+1},\mu_{h+1})\right].$$

For each h, add and subtract $\widetilde{U}_h(s_h,a_h,\mathcal{P}_h^*)$, $\widetilde{U}_h(s_h,a_h,\widetilde{\mathcal{P}}_h)$, and $(\mathcal{T}_h^{g,\widetilde{\nu}}\widetilde{U}_h)(s_h,\mu_h)$. After taking expectations, (G.26) cancels the last two batch-averaged quantities. Collecting the remaining terms proves (G.25). □

Lemma G.9 (FSO regret decomposition for batch-adaptive execution)**.** *For each t, define the predictable historical confidence event $\mathcal{E}^t_{\mathrm{conf}} := \mathcal{E}^t_{\mathrm{or}} \cap \mathcal{E}^{<t}_{\mathrm{emp}} \cap \{\theta^*_h \in \Theta^t_h \text{ for every } h \in [H]\}$. On $\bigcap_{t=1}^T \mathcal{E}^t_{\mathrm{conf}}$,*

$$\mathrm{Reg}_{\mathrm{FSO}}(T) \le \mathrm{LBE}' + L_U^{\mathrm{FSO}} \mathrm{JDE}' + \mathrm{Bias}_{\mathrm{batch}}, \tag{G.27}$$

where

$$\mathrm{JDE}' := \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^t \left[\sum_{x \in \mathcal{X}} \mu^t_h(x) \left\| \nu^{p^t_h, \bar{\theta}^t_h}(\cdot \mid x; s^t_h) - \nu^{p^t_h, \theta^*_h}(\cdot \mid x; s^t_h) \right\|_1 \right], \tag{G.28}$$

$$\mathrm{LBE}' := \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^t \left[(\widehat{U}^t_h - u_h)(s^t_h, a^t_h, \mathcal{P}^t_h) - \widehat{W}^t_{h+1}(s^t_{h+1}, \mu^t_{h+1}) \right],$$

$$\mathrm{Bias}_{\mathrm{batch}} := \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^t \left[\widehat{W}^t_h(s^t_h, \mu^t_h) - (\mathcal{T}_h^{g^t, \widetilde{\nu}^t} \widehat{U}^t_h)(s^t_h, \mu^t_h) \right].$$

Moreover,

$$\mathrm{Bias}_{\mathrm{batch}} \le HT\bar{\epsilon}_N \left[C^{\mathrm{act}}_W(B^{\mathrm{FSO}}_T) + C^{\mathrm{act}}_G(B^{\mathrm{FSO}}_T) \right] \le 2HT\bar{\epsilon}_N C^{\mathrm{env}}_{\mathrm{pair}}(B^{\mathrm{FSO}}_T). \tag{G.29}$$

Proof of Lemma G.9. Fix an episode t and work on $\mathcal{E}^t_{\mathrm{conf}}$, which is $\mathcal{F}^{t-1}$-measurable. Pointwise comparator optimism at true queries and $\theta^*_h \in \Theta^t_h$ imply $J(\pi^*) \le \widehat{W}^t_1(s_0, \mu_0)$. Apply Lemma G.8 with the measurable candidate response from (G.5). The first sum in (G.25) is the episode-t contribution to LBE'. The third sum is bounded by

$$\widehat{U}^t_h(s^t_h, a^t_h, \widetilde{\mathcal{P}}^t_h) - \widehat{U}^t_h(s^t_h, a^t_h, \mathcal{P}^t_h) \le L_U^{\mathrm{FSO}} \sum_{x \in \mathcal{X}} \mu^t_h(x) \left\| \nu^{p^t_h, \bar{\theta}^t_h}(\cdot \mid x; s^t_h) - \nu^{p^t_h, \theta^*_h}(\cdot \mid x; s^t_h) \right\|_1.$$

For a fixed hypothetical batch $\mathbf{x}$, abbreviate $p_{\mathbf{x}} = p^t_{h,\mathbf{x}}(s)$ and $\bar{\theta}_{\mathbf{x}} = \bar{\theta}^t_{h,\mathbf{x}}(s)$, and set $G^t_h(s, m; p, \theta) := \langle \widehat{U}^t_h(s, \cdot, m \otimes \nu^{p,\theta}(\cdot \mid \cdot; s)), p \rangle_{\mathcal{A}}$. The second sum is the conditional expectation of $\widehat{W}^t_h(s, \mu) - G^t_h(s, \mu; p_{\mathbf{x}}, \bar{\theta}_{\mathbf{x}})$. The exact identity

$$\begin{aligned} \widehat{W}^t_h(s, \mu) - G^t_h(s, \mu; p_{\mathbf{x}}, \bar{\theta}_{\mathbf{x}}) = {} & [\widehat{W}^t_h(s, \mu) - \widehat{W}^t_h(s, \widehat{\mu}_{\mathbf{x}})] + [\widehat{W}^t_h(s, \widehat{\mu}_{\mathbf{x}}) - G^t_h(s, \widehat{\mu}_{\mathbf{x}}; p_{\mathbf{x}}, \bar{\theta}_{\mathbf{x}})] \\ & + [G^t_h(s, \widehat{\mu}_{\mathbf{x}}; p_{\mathbf{x}}, \bar{\theta}_{\mathbf{x}}) - G^t_h(s, \mu; p_{\mathbf{x}}, \bar{\theta}_{\mathbf{x}})] \end{aligned}$$

has a zero middle term by the empirical planning rule. The two remaining terms are at most $[C^{\mathrm{act}}_W(B^{\mathrm{FSO}}_T) + C^{\mathrm{act}}_G(B^{\mathrm{FSO}}_T)] \|\mu - \widehat{\mu}_{\mathbf{x}}\|_1$. Taking the current-batch conditional expectation, applying Lemma G.2, and summing over (t, h) proves (G.29). Summing the three performance-difference terms over episodes proves (G.27). □

Predictable empirical-design uncertainty for FSO. For the FSO specialization, define

$$y^{\mathrm{or}}_{h,\tau}(W) := f^*_{h,W}(\widehat{z}^\tau_h) + \xi_{h,\tau}(W).$$

Lemma G.10 (Uniform oracle uncertainty on the predictable FSO design)**.** *Under the conditions of Lemma H.1, the event $\mathcal{E}_{\mathrm{or,FSO}}$ has probability at least $1 - \zeta$. On this event, uniformly over $h \in [H]$, $t \in [T]$, every true evaluation point z, and every continuation value W in the deterministic capped FSO class,*

$$\left| \widehat{f}^{t,\mathrm{or}}_h(W)(z) - (\mathbb{B}_h W)(z) \right| \le B^{\mathrm{base}}_T \|\phi(z)\|_{(\widehat{\Lambda}^t_{L,h})^{-1}}. \tag{G.30}$$

Proof of Lemma G.10. This is the specialization of (H.3) to $\iota = \mathrm{FSO}$, using $B^{\mathrm{or}}_{T,\mathrm{FSO}} = B^{\mathrm{base}}_T$, $\widehat{f}^{t,\mathrm{or}}_h = \widehat{f}^{t,\mathrm{or}}_{h,\mathrm{FSO}}$, and $\widehat{\Lambda}^t_{L,h,\mathrm{FSO}} = \widehat{\Lambda}^t_{L,h}$. □

Lemma G.11 (KRR normal-equation norm cap with a same-design label shift)**.** *Let $\Phi : \mathbb{H} \to \mathbb{R}^m$ be a finite design operator and let $\Lambda := \lambda I_{\mathbb{H}} + \Phi^*\Phi$ with $\lambda > 0$. Let $f^* \in \mathbb{H}$, let $y^{\mathrm{or}} = \Phi f^* + \xi$, and let $y = y^{\mathrm{or}} + \delta$. Define $\widehat{f}^{\mathrm{or}} := \Lambda^{-1}\Phi^* y^{\mathrm{or}}$ and $\widehat{f} := \Lambda^{-1}\Phi^* y$. If $\|f^*\|_{\mathbb{H}} \le R$, $\|\Phi^*\xi\|_{\Lambda^{-1}} \le \beta_{\mathrm{noise}}$, and $\sqrt{\lambda}R + \beta_{\mathrm{noise}} \le \mathsf{r}$, then*

$$\|\widehat{f}^{\mathrm{or}}\|_{\mathbb{H}} \le R + \lambda^{-1/2}\beta_{\mathrm{noise}} \le \mathsf{r}\lambda^{-1/2}, \tag{G.31}$$

$$\|\widehat{f} - \widehat{f}^{\mathrm{or}}\|_{\mathbb{H}} \le \lambda^{-1/2}\|\delta\|_2, \qquad \|\widehat{f}\|_{\mathbb{H}} \le (\mathsf{r} + \|\delta\|_2)\lambda^{-1/2}. \tag{G.32}$$

Proof of Lemma G.11. The KRR normal equation gives

$$\widehat{f}^{\mathrm{or}} = (I_{\mathbb{H}} - \lambda\Lambda^{-1})f^* + \Lambda^{-1}\Phi^*\xi.$$

Because $0 \preceq I_{\mathbb{H}} - \lambda\Lambda^{-1} \preceq I_{\mathbb{H}}$, the first term has RKHS norm at most R. Because $\Lambda \succeq \lambda I_{\mathbb{H}}$,

$$\|\Lambda^{-1}\Phi^*\xi\|_{\mathbb{H}}^2 \le \lambda^{-1}\|\Phi^*\xi\|_{\Lambda^{-1}}^2.$$

This proves (G.31). For the shifted labels, $\widehat{f} - \widehat{f}^{\mathrm{or}} = \Lambda^{-1}\Phi^*\delta$. The operator inequality $\Phi\Lambda^{-1}\Phi^* \preceq I_m$ gives

$$\|\Lambda^{-1}\Phi^*\delta\|_{\mathbb{H}}^2 \le \lambda^{-1}\langle\delta, \Phi\Lambda^{-1}\Phi^*\delta\rangle \le \lambda^{-1}\|\delta\|_2^2.$$

The triangle inequality proves (G.32). □

Non-circular inflation of the empirical-design radius. We next index the FSO value and objective Lipschitz constants by a candidate uncertainty radius $\mathsf{r} > 0$. Define $R_{T,\mathrm{FSO}}^{\mathrm{base}}(\mathsf{r}) = \mathsf{r}\lambda^{-1/2}$ as in (C.19) and set

$$L_U(\mathsf{r}) := L_\phi\left(R_{T,\mathrm{FSO}}^{\mathrm{base}}(\mathsf{r}) + \mathsf{r}\lambda^{-1/2}\right) = 2L_\phi\mathsf{r}\lambda^{-1/2}. \tag{G.33}$$

Use the deterministic envelopes $C_{\mathrm{pair}}^{\mathrm{env}}(\mathsf{r}) = L_U(\mathsf{r})$ and $C_{\mathrm{tar}}^{\mathrm{env}}(\mathsf{r}) = C_{\mathrm{tar}}^{(0)} + c_q\mathsf{r}$, where $C_{\mathrm{tar}}^{(0)} = R_U H L_\phi$ and $c_q = 2L_\phi\lambda^{-1/2}$.

For parameter choices satisfying (G.36), use the following actual FSO uncertainty radius.

Suppose $B_T^{\mathrm{base}} > 0$ is fixed by the minimal dyadic calibration and define

$$q_T := \sqrt{T}\,\bar{\epsilon}_N c_q, \quad B_T^{\mathrm{FSO}} := \frac{B_T^{\mathrm{base}} + \sqrt{T}\,\bar{\epsilon}_N C_{\mathrm{tar}}^{(0)}}{1 - q_T}, \quad \beta_{\mathrm{tar}}^{\max} := B_T^{\mathrm{FSO}} - B_T^{\mathrm{base}}. \tag{G.34}$$

Substituting $C_{\mathrm{tar}}^{(0)} = R_U H L_\phi$ and $c_q = 2L_\phi\lambda^{-1/2}$ gives

$$B_T^{\mathrm{FSO}} = \frac{B_T^{\mathrm{base}} + \sqrt{T}\,\bar{\epsilon}_N R_U H L_\phi}{1 - 2\sqrt{T}\,\bar{\epsilon}_N L_\phi\lambda^{-1/2}}.$$

Define

$$L_U^{\mathrm{FSO}} := 2L_\phi B_T^{\mathrm{FSO}}\lambda^{-1/2}. \tag{G.35}$$

If

$$q_T \le \frac{1}{2}, \quad \sqrt{T}\,\bar{\epsilon}_N C_{\mathrm{tar}}^{(0)} \le B_T^{\mathrm{base}}, \tag{G.36}$$

then Lemma G.12 gives $B_T^{\mathrm{FSO}} \le 4B_T^{\mathrm{base}} = B_T^{\mathrm{cap}}$ and shows that the historical target perturbation is absorbed by $\beta_{\mathrm{tar}}^{\max}$.

For the radius-absorption conditions in (G.36), set $r_0 := L_\phi(2\lambda)^{-1/2}$. A deterministic sufficient condition for (G.36) is

$$N \ge C_\epsilon^2 T\,(|\mathcal{Z}| + \log(HT/\zeta))\max\{4c_q^2, r_0^2\}. \tag{G.37}$$

The actual FSO radius and its affine target envelope are defined in (G.34) and (G.39); the absorption and deterministic cap inclusion are proved in Lemma G.12.

Let $\mathfrak{W}_h(\mathsf{r})$ be the deterministic class of FSO value envelopes generated by $\mathcal{U}_{\mathrm{UCB}}(h, R_{T,\mathrm{FSO}}^{\mathrm{base}}(\mathsf{r}), \mathsf{r}, L_U(\mathsf{r})\bar{\epsilon}_N)$ and $\mathfrak{A}_{\mathrm{alg,FSO}}$, and let $\mathfrak{G}_h(\mathsf{r})$ be the corresponding class of fixed-pair objectives $G_U(s,\mu;p,\theta)$. Within each $W \in \mathfrak{W}_h(\mathsf{r})$, the feasible set Θ' remains fixed when μ and μ' are compared. In particular, the realized set Θ_h^t is determined by the episode prefix and is not recomputed as a function of the current hypothetical mean field.

$$\begin{aligned}
C_W^{\mathrm{act}}(\mathsf{r}) &:= \max_{h\in[H]} \sup_{\substack{W\in\mathfrak{W}_h(\mathsf{r}),\, s \\ \mu\neq\mu'}} \frac{|W(s,\mu) - W(s,\mu')|}{\|\mu - \mu'\|_1}, \\
C_G^{\mathrm{act}}(\mathsf{r}) &:= \max_{h\in[H]} \sup_{\substack{G\in\mathfrak{G}_h(\mathsf{r}),\, s,p,\theta \\ \mu\neq\mu'}} \frac{|G(s,\mu;p,\theta) - G(s,\mu';p,\theta)|}{\|\mu - \mu'\|_1}, \\
C_{\mathrm{pair}}^{\mathrm{act}}(\mathsf{r}) &:= \max\left\{C_W^{\mathrm{act}}(\mathsf{r}), C_G^{\mathrm{act}}(\mathsf{r})\right\}.
\end{aligned} \tag{G.38}$$

The common Lipschitz envelope yields $C^{\mathrm{act}}_{\mathrm{pair}}(\mathsf{r}) \le C^{\mathrm{env}}_{\mathrm{pair}}(\mathsf{r}) := L_U(\mathsf{r})$. We next separate the realized target-perturbation coefficient from a deterministic affine envelope:

$$\begin{aligned} C^{\mathrm{act}}_{\mathrm{tar}}(\mathsf{r}) &:= R_U H L_\phi + C^{\mathrm{act}}_W(\mathsf{r}), & C^{(0)}_{\mathrm{tar}} &:= R_U H L_\phi, \\ c_q &:= 2L_\phi \lambda^{-1/2}, & C^{\mathrm{env}}_{\mathrm{tar}}(\mathsf{r}) &:= C^{(0)}_{\mathrm{tar}} + c_q \mathsf{r}. \end{aligned} \tag{G.39}$$

The affine envelope is deterministic in r and closes the radius-inflation inequality without a circular definition.

Lemma G.12 (Non-circular absorption of the finite-sample target perturbation)**.** *Suppose $B^{\mathrm{base}}_T > 0$ is fixed by the minimal dyadic calibration in (C.25).*

Use the coefficients in (G.39), and let B^{FSO}_T and $\beta^{\max}_{\mathrm{tar}}$ be defined by (G.34).

If (G.36) holds, then

$$\beta^{\max}_{\mathrm{tar}} = \sqrt{T}\,\bar\epsilon_N C^{\mathrm{env}}_{\mathrm{tar}}\left(B^{\mathrm{FSO}}_T\right), \quad \sqrt{T}\,\bar\epsilon_N C^{\mathrm{act}}_{\mathrm{tar}}\left(B^{\mathrm{FSO}}_T\right) \le \beta^{\max}_{\mathrm{tar}}, \tag{G.40}$$

and

$$B^{\mathrm{FSO}}_T \le 4B^{\mathrm{base}}_T = B^{\mathrm{cap}}_T. \tag{G.41}$$

Proof of Lemma G.12. The definition of $C^{\mathrm{env}}_{\mathrm{tar}}$ gives $C^{\mathrm{env}}_{\mathrm{tar}}(\mathsf{r}) = C^{(0)}_{\mathrm{tar}} + c_q\mathsf{r}$. Substituting the definition of B^{FSO}_T into $B^{\mathrm{FSO}}_T - B^{\mathrm{base}}_T$ proves the first identity in (G.40). The bound $C^{\mathrm{act}}_{\mathrm{tar}}(\mathsf{r}) \le C^{\mathrm{env}}_{\mathrm{tar}}(\mathsf{r})$ proves the inequality in (G.40). Finally, (G.36) implies

$$B^{\mathrm{FSO}}_T \le \frac{2B^{\mathrm{base}}_T}{1/2} = 4B^{\mathrm{base}}_T,$$

which proves (G.41). □

Remark G.13 (Computability of the inflated radius). All quantities in (G.39) and (G.34) are deterministic functions of the known problem scales and $(H, T, N, \lambda, \zeta)$. The algorithm may therefore compute B^{FSO}_T before sampling, and no runtime fixed-point iteration is required.

The eventwise containment of the realized FSO iterate in $\mathcal{U}_{\mathrm{UCB}}(h, R^{\mathrm{base}}_{T,\mathrm{FSO}}(B^{\mathrm{FSO}}_T), B^{\mathrm{FSO}}_T, L^{\mathrm{FSO}}_U \bar\epsilon_N)$ is proved in Lemma G.11 and Corollary G.15.

Lemma G.14 (Actual-versus-oracle FSO KRR perturbation)**.** *Fix (h,t) and suppose that $\widehat{W}^t_{h+1}$ belongs both to the actual-radius continuation-value class and to the deterministic capped continuation-value class in (G.47). On $\mathcal{E}^t_{\mathrm{or}} \cap \mathcal{E}^{<t}_{\mathrm{emp}}$,*

$$\left\| \left(\delta^{\mathrm{curr},t}_{h,\tau} + \delta^{\mathrm{next},t}_{h,\tau}\right)_{\tau<t} \right\|_2 \le \sqrt{t-1}\, C^{\mathrm{act}}_{\mathrm{tar}}(B^{\mathrm{FSO}}_T)\bar\epsilon_N \le \beta^{\max}_{\mathrm{tar}}, \tag{G.42}$$

$$\left\| \widehat{f}^t_h - \widehat{f}^{t,\mathrm{or}}_h(\widehat{W}^t_{h+1}) \right\|_{\mathbb{H}} \le \beta^{\max}_{\mathrm{tar}} \lambda^{-1/2}, \tag{G.43}$$

$$\|\widehat{f}^t_h\|_{\mathbb{H}} \le B^{\mathrm{FSO}}_T \lambda^{-1/2} = R^{\mathrm{base}}_{T,\mathrm{FSO}}(B^{\mathrm{FSO}}_T). \tag{G.44}$$

For every evaluation point z,

$$\left| \widehat{f}^t_h(z) - \widehat{f}^{t,\mathrm{or}}_h(\widehat{W}^t_{h+1})(z) \right| \le \beta^{\max}_{\mathrm{tar}} \|\phi(z)\|_{(\widehat{\Lambda}^t_{L,h})^{-1}}. \tag{G.45}$$

Proof of Lemma G.14. For every $\tau < t$, define

$$\begin{aligned} \delta^{\mathrm{curr},t}_{h,\tau} &:= f^*_{h,\widehat{W}^t_{h+1}}(z^\tau_h) - f^*_{h,\widehat{W}^t_{h+1}}(\widehat{z}^\tau_h), \\ \delta^{\mathrm{next},t}_{h,\tau} &:= \widehat{W}^t_{h+1}(s^\tau_{h+1}, \widehat{\mu}^\tau_{h+1}) - \widehat{W}^t_{h+1}(s^\tau_{h+1}, \mu^\tau_{h+1}). \end{aligned}$$

The actual label minus the oracle label is their sum. By Assumptions 2.4 and 2.5, $|\delta^{\mathrm{curr},t}_{h,\tau}| \le R_U H L_\phi \bar\epsilon_N$. By the actual-radius membership and the value-envelope bound in Lemma C.23, $|\delta^{\mathrm{next},t}_{h,\tau}| \le C^{\mathrm{act}}_W(B^{\mathrm{FSO}}_T)\bar\epsilon_N$. These coordinatewise bounds and (G.40) prove (G.42). Apply

Lemma G.11 with $R = R_U H$, $\beta_{\text{noise}} = H\sqrt{\mathcal{R}_{T,\text{FSO}}}$, $\mathsf{r} = B_T^{\text{base}}$, and $\delta = (\delta_{h,\tau}^{\text{curr},t} + \delta_{h,\tau}^{\text{next},t})_{\tau<t}$. The uniform noise event (H.2) and the dyadic calibration give

$$R_U H + \frac{H\sqrt{\mathcal{R}_{T,\text{FSO}}}}{\sqrt{\lambda}} \le \frac{B_T^{\text{base}}}{\sqrt{\lambda}}.$$

Equations (G.31) and (G.32), together with $B_T^{\text{FSO}} = B_T^{\text{base}} + \beta_{\text{tar}}^{\max}$, prove (G.43) and (G.44). Finally, the same-design push-through identity and Cauchy–Schwarz give

$$\left|\widehat{f}_h^t(z) - \widehat{f}_h^{t,\text{or}}(\widehat{W}_{h+1}^t)(z)\right| \le \|\phi(z)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}} \left\| \left(\delta_{h,\tau}^{\text{curr},t} + \delta_{h,\tau}^{\text{next},t}\right)_{\tau<t} \right\|_2,$$

which proves (G.45). □

Corollary G.15 (Episodewise actual-predictor uncertainty and FSO class containment)**.** *For every episode t, on $\mathcal{E}_{\text{or}}^t \cap \mathcal{E}_{\text{emp}}^{<t}$, uniformly over $h \in [H]$ and every true evaluation point z,*

$$\left|\widehat{f}_h^t(z) - (\mathbb{B}_h \widehat{W}_{h+1}^t)(z)\right| \le B_T^{\text{FSO}} \|\phi(z)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}}. \tag{G.46}$$

On the same event, the realized iterates satisfy

$$\begin{aligned} \widehat{U}_h^t &\in \mathcal{U}_{\text{UCB}}\left(h, R_{T,\text{FSO}}^{\text{base}}(B_T^{\text{FSO}}), B_T^{\text{FSO}}, L_U^{\text{FSO}}\bar{\epsilon}_N\right), \\ \widehat{W}_h^t &\in \mathcal{V}_{\text{UCB,FSO}}\left(h, R_{T,\text{FSO}}^{\text{base}}(B_T^{\text{FSO}}), B_T^{\text{FSO}}, L_U^{\text{FSO}}\bar{\epsilon}_N\right), \\ \widehat{W}_h^t &\in \mathcal{V}_{\text{UCB,FSO}}\left(h, R_{T,\text{FSO}}^{\text{cap}}(B_T^{\text{base}}), B_T^{\text{cap}}, C_{\text{off}}^{\text{cap}}\right). \end{aligned} \tag{G.47}$$

For each of the two parameter triples in (G.47), use the terminal convention $\mathcal{V}_{\text{UCB,FSO}}(H+1, R, B, C) := \{0\}$.

Proof of Corollary G.15. Fix an episode t and argue backward over $h \in \{H, H-1, \dots, 1\}$. The base case holds because $\widehat{W}_{H+1}^t \equiv 0$ belongs to both the actual-radius and capped continuation-value classes in (G.47). Assume that $\widehat{W}_{h+1}^t$ belongs to both classes. Then Lemma G.10, equation (G.30), gives

$$\left|\widehat{f}_h^{t,\text{or}}(\widehat{W}_{h+1}^t)(z) - (\mathbb{B}_h \widehat{W}_{h+1}^t)(z)\right| \le B_T^{\text{base}} \|\phi(z)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}}. \tag{G.48}$$

Under the same induction hypothesis, Lemma G.14, equation (G.45), gives

$$\left|\widehat{f}_h^t(z) - \widehat{f}_h^{t,\text{or}}(\widehat{W}_{h+1}^t)(z)\right| \le \beta_{\text{tar}}^{\max} \|\phi(z)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}}. \tag{G.49}$$

Adding (G.48) and (G.49) and using $B_T^{\text{FSO}} = B_T^{\text{base}} + \beta_{\text{tar}}^{\max}$ proves (G.46) at step h.

Equation (G.44) gives $\|\widehat{f}_h^t\|_{\mathbb{H}} \le R_{T,\text{FSO}}^{\text{base}}(B_T^{\text{FSO}})$. The definition of the optimistic action value in Algorithm 3, together with the realized posterior-shape representation (C.11), gives the first containment in (G.47). The realized confidence ellipsoid belongs to $\mathfrak{A}_{\text{alg,FSO}}$ by (C.13). Applying the planning rule (3.7) in Algorithm 3 to the actual-radius action-value class gives the second containment in (G.47).

By (G.41), $B_T^{\text{FSO}} \le B_T^{\text{cap}}$ and $R_{T,\text{FSO}}^{\text{base}}(B_T^{\text{FSO}}) \le R_{T,\text{FSO}}^{\text{cap}}(B_T^{\text{base}})$. By the definition $\bar{\epsilon}_N = \min\{\epsilon_N, 2\}$, the definitions of the action-value Lipschitz constants, and $B_T^{\text{FSO}} \le B_T^{\text{cap}}$,

$$L_U^{\text{FSO}}\bar{\epsilon}_N = 2L_\phi B_T^{\text{FSO}} \lambda^{-1/2}\bar{\epsilon}_N \le 4L_\phi B_T^{\text{cap}} \lambda^{-1/2} = C_{\text{off}}^{\text{cap}}. \tag{G.50}$$

Equation (G.50) and the posterior-shape definition (C.11) place $\widehat{U}_h^t$ in the capped UCB class. Applying the same planning rule to the capped UCB class gives the third containment in (G.47). This closes the backward induction and proves both conclusions simultaneously. □

Proof of Lemma C.27. We establish (C.38) by backward induction in $h \in \{H+1, H, \dots, 1\}$, with the standard base case $\widehat{W}_{H+1}^t(s,\mu) \equiv 0 = W_{H+1}^{\pi^*}(s,\mu)$.

Inductive hypothesis: Suppose $\widehat{W}^t_{h+1}(s,\mu) \geq W^{\pi^*}_{h+1}(s,\mu)$ for all (s,μ).

Step 1: Uniform Bellman dominance on the true evaluation domain. Fix episode t and work on $\mathcal{E}^t_{\mathrm{or}} \cap \mathcal{E}^{<t}_{\mathrm{emp}} \cap \mathcal{E}_{\mathrm{MLE,main}}$. For $z = (s,a,\mathcal{P})$, set $f^t_h(z) := u_h(z) + (P_h \widehat{W}^t_{h+1})(z)$. By Corollary G.15,

$$|\widehat{f}^t_h(z) - f^t_h(z)| \leq B^{\mathrm{FSO}}_T \|\phi(z)\|_{(\widehat{\Lambda}^t_{L,h})^{-1}}, \quad \forall z.$$

The pre-truncation optimistic quantity therefore dominates $f^t_h(z)$ at every true query. Since $f^t_h(z) \in [0, H-h+1]$, truncation preserves this one-sided inequality and gives

$$\widehat{U}^t_h(z) \geq f^t_h(z), \quad \forall z \in \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}).$$

The confidence inequality is uniform in z, so it may be evaluated directly at the true comparator query; no empirical comparator query is defined.

Step 2: Inductive optimism. On $\mathcal{E}_{\mathrm{MLE,main}}$, the true parameter satisfies $\theta^*_h \in \Theta^t_h$, and the benchmark local policy belongs to Π_{loc}. Therefore,

$$\begin{aligned}\widehat{W}^t_h(s,\mu) &\geq \sum_{a\in\mathcal{A}} p^*_h(a \mid s,\mu)\widehat{U}^t_h\Big(s,a,\mu \otimes \nu^{p^*_h,\theta^*_h}(\cdot \mid \cdot; s)\Big) \\ &\geq \sum_{a\in\mathcal{A}} p^*_h(a \mid s,\mu)\left[u_h(s,a,\mathcal{P}^*) + (P_h\widehat{W}^t_{h+1})(s,a,\mathcal{P}^*)\right] \geq W^{\pi^*}_h(s,\mu),\end{aligned}$$

where $\mathcal{P}^* = \mu \otimes \nu^{p^*_h,\theta^*_h}(\cdot \mid \cdot; s)$ and the last inequality uses the induction hypothesis.

Step 3: Aggregate consequence. The pointwise inequality gives $\mathbb{E}[\widehat{W}^t_1(s_1,\mu_1)] \geq J(\pi^*)$ for every episode. Summing over t gives the aggregate optimism claim. The uncertainty coefficient is already the actual radius B^{FSO}_T from (G.34). □

Lemma G.16 (Leader Bellman error under finite-sample observation)**.** *There exists a leader bonus-martingale event $\mathcal{E}_{\text{LBE-mart}}$ with $\mathbb{P}(\mathcal{E}_{\text{LBE-mart}}) \geq 1-\zeta$ such that the following conclusion holds. On $\bigcap_{t=1}^T(\mathcal{E}^t_{\mathrm{or}} \cap \mathcal{E}^{<t}_{\mathrm{emp}}) \cap \mathcal{E}_{\text{LBE-mart}}$, the leader Bellman error satisfies*

$$\begin{aligned}\mathrm{LBE}' \leq\ & 4B^{\mathrm{FSO}}_T H\sqrt{TG_K(T,\lambda)} + 2B^{\mathrm{FSO}}_T B_\phi H\sqrt{T}\log\left(\frac{2eH}{\zeta}\right) \\ &+ 2B^{\mathrm{FSO}}_T L_\phi \lambda^{-1/2} HT\bar{\epsilon}_N + L^{\mathrm{FSO}}_U HT\bar{\epsilon}_N. \end{aligned} \tag{G.51}$$

Proof of Lemma G.16. Work on $\bigcap_{t=1}^T(\mathcal{E}^t_{\mathrm{or}} \cap \mathcal{E}^{<t}_{\mathrm{emp}})$. For each (t,h), let $z^t_h = (s^t_h, a^t_h, \mathcal{P}^t_h)$ and define $f^t_h(z) := u_h(z) + (P_h\widehat{W}^t_{h+1})(z)$ and $\Gamma^t_h(z) := B^{\mathrm{FSO}}_T\|\phi(z)\|_{(\widehat{\Lambda}^t_{L,h})^{-1}}$.

Step 1: Pointwise Bellman comparison at the true query. By Corollary G.15, equation (G.46),

$$f^t_h(z^t_h) - \Gamma^t_h(z^t_h) \leq \widehat{f}^t_h(z^t_h) \leq f^t_h(z^t_h) + \Gamma^t_h(z^t_h).$$

Since $\Gamma_{\mathrm{FS}} = L^{\mathrm{FSO}}_U \bar{\epsilon}_N \geq 0$, the pre-truncation quantity satisfies

$$f^t_h(z^t_h) \leq \widehat{f}^t_h(z^t_h) + \Gamma^t_h(z^t_h) + \Gamma_{\mathrm{FS}} \leq f^t_h(z^t_h) + 2\Gamma^t_h(z^t_h) + \Gamma_{\mathrm{FS}}.$$

The bounded reward and continuation value give $f^t_h(z^t_h) \in [0, H-h+1]$. Because truncation onto this interval is monotone and fixes $f^t_h(z^t_h)$,

$$0 \leq \widehat{U}^t_h(z^t_h) - f^t_h(z^t_h) \leq 2\Gamma^t_h(z^t_h) + \Gamma_{\mathrm{FS}}. \tag{G.52}$$

Step 2: Conditional centering of the system transition. Under Assumption 2.2, conditionally on $\mathcal{F}^B_{t,h}$,

$$\mathbb{E}\left[\widehat{W}^t_{h+1}(s^t_{h+1},\mu^t_{h+1}) \mid \mathcal{F}^B_{t,h}\right] = (P_h\widehat{W}^t_{h+1})(z^t_h).$$

The tower property and the definition of f^t_h therefore give

$$\mathbb{E}^t\left[(\widehat{U}^t_h - u_h)(z^t_h) - \widehat{W}^t_{h+1}(s^t_{h+1},\mu^t_{h+1})\right] = \mathbb{E}^t\left[\widehat{U}^t_h(z^t_h) - f^t_h(z^t_h)\right].$$

Summing (G.52) over (t, h) gives

$$\mathrm{LBE}' \le 2\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}^t[\Gamma_h^t(z_h^t)] + L_U^{\mathrm{FSO}} HT\bar{\epsilon}_N. \tag{G.53}$$

Step 3: Conversion from the true query to the empirical executed query. Let $\widehat{z}_h^t = (s_h^t, a_h^t, \widehat{\mathcal{P}}_h^t)$. The triangle inequality for the $(\widehat{\Lambda}_{L,h}^t)^{-1}$ seminorm gives

$$\|\phi(z_h^t)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}} \le \|\phi(\widehat{z}_h^t)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}} + \|\phi(z_h^t) - \phi(\widehat{z}_h^t)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}}.$$

Since $\widehat{\Lambda}_{L,h}^t \succeq \lambda I_{\mathbb{H}}$, Assumption 2.4 yields

$$\|\phi(z_h^t) - \phi(\widehat{z}_h^t)\|_{(\widehat{\Lambda}_{L,h}^t)^{-1}} \le L_\phi \lambda^{-1/2} \|\mathcal{P}_h^t - \widehat{\mathcal{P}}_h^t\|_1.$$

Thus

$$\Gamma_h^t(z_h^t) \le \Gamma_h^t(\widehat{z}_h^t) + B_T^{\mathrm{FSO}} L_\phi \lambda^{-1/2} \|\mathcal{P}_h^t - \widehat{\mathcal{P}}_h^t\|_1.$$

Taking $\mathbb{E}^t$, applying Lemma G.2, and summing gives

$$\sum_{t,h}\mathbb{E}^t[\Gamma_h^t(z_h^t)] \le \sum_{t,h}\mathbb{E}^t[\Gamma_h^t(\widehat{z}_h^t)] + B_T^{\mathrm{FSO}} L_\phi \lambda^{-1/2} HT\bar{\epsilon}_N. \tag{G.54}$$

Step 4: Martingale concentration bound for the empirical-query bonus. For each fixed h, set $X_{t,h} := \Gamma_h^t(\widehat{z}_h^t)$. The design operator and bonus rule are predictable at the episode start, and $\widehat{z}_h^t$ is observed during episode t, so $(X_{t,h})_{t=1}^T$ is adapted to the episode filtration. Since $\widehat{\Lambda}_{L,h}^t \succeq \lambda I_{\mathbb{H}}$, $\lambda \ge 1$, and $\|\phi(z)\|_{\mathbb{H}} \le B_\phi$, one has $0 \le X_{t,h} \le B_T^{\mathrm{FSO}} B_\phi$. Apply the additive bounded-process inequality (E.10) from Lemma E.3 with failure probability ζ/H and take a union bound over h. On the resulting event,

$$\sum_{t=1}^{T}\mathbb{E}^t[X_{t,h}] \le \sum_{t=1}^{T} X_{t,h} + B_T^{\mathrm{FSO}} B_\phi \sqrt{T}\log\left(\frac{2eH}{\zeta}\right) \tag{G.55}$$

simultaneously for every $h \in [H]$.

Step 5: Elliptical potential. For each fixed h, Cauchy–Schwarz and Lemma E.5, equation (E.12), yield

$$\sum_{t=1}^{T}\Gamma_h^t(\widehat{z}_h^t) \le B_T^{\mathrm{FSO}}\sqrt{T\sum_{t=1}^{T}\|\phi(\widehat{z}_h^t)\|^2_{(\widehat{\Lambda}_{L,h}^t)^{-1}}} \le 2B_T^{\mathrm{FSO}}\sqrt{TG_K(T,\lambda)}.$$

Summing over h gives

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\Gamma_h^t(\widehat{z}_h^t) \le 2B_T^{\mathrm{FSO}} H\sqrt{TG_K(T,\lambda)}. \tag{G.56}$$

Step 6: Assembly. Substituting (G.54), (G.55), and (G.56) into (G.53) gives (G.51). Only the LBE-martingale event is introduced in this proof, and its failure probability is at most ζ. □

Lemma G.17 (Direct JDE bound via total variation and matrix Cauchy–Schwarz)**.** *Under the conditions of Lemma G.5, on* $\mathcal{E}_{\mathrm{MLE,main}} \cap \mathcal{E}_{\mathrm{JDE\text{-}mart}}$ *(probability at least* $1 - 2\zeta$*),*

$$\mathrm{JDE}' \lesssim C_{\mathrm{sc}}\eta H\sqrt{\beta_{\mathrm{F}}^T d_F T L_{T,N}} + C_{\mathrm{sc}}^2 C_{\mathrm{JDE}}\beta_{\mathrm{F}}^T d_F H L_{T,N} + HT\bar{\epsilon}_N + R_Y H\log(2H/\zeta)\,. \tag{G.57}$$

Here C_{JDE} *and* R_Σ *are defined in* (C.18)*,* $L_{T,N}$ *is defined in* (C.41)*, and* R_Y *is defined in* (C.42)*. The continuous convention* $L_{T,N} = 0$ *when* $R_\Sigma = 0$ *applies.*

Remark G.18 (Why the JDE leading term has no polynomial $1/\sqrt{N}$ gain). The trace route in Stages 4–5 contains an apparent factor $1/(N\eta^2)$ because $\widehat{\Gamma}_h^t(\widehat{\theta}_h^t) = (N\eta^2)^{-1}(V_h^{t+1} - V_h^t)$. This does not yield a polynomial $1/\sqrt{N}$ improvement in the leading JDE term: the trace-form elliptical-potential bound is applied to increments $X_t = V_h^{t+1} - V_h^t$ with $\mathrm{Tr}(X_t) \le N\eta^2 R_\Sigma$, and the resulting $N\eta^2$ numerator cancels the preceding normalization. The exact cancellation is shown in (G.70).

Proof of Lemma G.17. For a fixed local policy, the response-estimation error compares the responses in (C.2) at $\bar{\theta}_h^t$ and θ_h^*. The proof has six main stages and an intermediate martingale step (Stage 5.5). *Stage 1* applies the TV perturbation bound $\mu \to \widehat{\mu}$ at the response-error level, giving an additive $2HT\bar{\epsilon}_N$ residual outside any square root; this is the central novel step that resolves the $T^{3/4}$ obstruction (see Remark C.32). *Stages 2–3* apply the softmax-Lipschitz bound and Jensen's inequality to the empirical-weighted error to obtain the empirical Fisher norm. *Stage 4* applies matrix Cauchy–Schwarz with $A = V_h^t$. *Stage 5* applies the self-concordance comparison $\widehat{\Gamma}_h^t(\theta^*) \to \widehat{\Gamma}_h^t(\widehat{\theta}^t)$ and the deterministic elliptical-potential lemma to the additive recurrence of V_h^t. *Stage 5.5* uses martingale concentration to control conditional-expectation sums by realized trajectory sums. *Stage 6* sums over t via Cauchy–Schwarz and over $h \in [H]$.

Notation convention. Throughout this proof, $\bar{\theta}_h^t$ denotes the realized optimistic follower parameter obtained from (G.4); it is distinct from the MLE center $\widehat{\theta}_h^t$. We also use the auxiliary additive Fisher matrix V_h^t from (G.15), with the strict additive recurrence

$$V_h^{t+1} \;=\; V_h^t \;+\; \eta^2 \sum_{n=1}^{N} \Sigma^{p_h^t,\widehat{\theta}_h^t}_{x_h^{t,n},s_h^t}.$$

The increment is evaluated at the contemporaneous MLE $\widehat{\theta}_h^t$ (*frozen* for $\tau \le t$); this admits the deterministic elliptical-potential argument used in Stage 5. The algorithmic confidence ellipsoid stored by Algorithm 3 is in $\widehat{F}_h^t$-norm with radius β_{F}^t (eq. (3.5)); via the PSD-equivalence (C.39), this translates to a V_h^t-norm bound with radius $C_{\mathrm{sc}}\beta_{\mathrm{F}}^t$ (the auxiliary form (G.16) of Lemma G.5), which is the form used inside Stage 4.

Stage 1 (key novel step): TV perturbation bound $\mu \to \widehat{\mu}$ at the response-error level. For each realized episode-step pair, define

$$e_h^t(x) := \left\| \nu^{p_h^t,\bar{\theta}_h^t}(\cdot \mid x; s_h^t) - \nu^{p_h^t,\theta_h^*}(\cdot \mid x; s_h^t) \right\|_1 \in [0,2].$$

The following pathwise inequality holds for every realization of the current state batch:

$$\sum_{x\in\mathcal{X}} \mu_h^t(x) e_h^t(x) \le \sum_{x\in\mathcal{X}} \widehat{\mu}_h^t(x) e_h^t(x) + 2\|\mu_h^t - \widehat{\mu}_h^t\|_1. \tag{G.58}$$

Taking $\mathbb{E}^t[\cdot]$ and using Lemma G.2, rather than conditioning on the non-predictable full event $\mathcal{E}_{\mathrm{emp}}$, gives

$$\begin{aligned}\mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \mu_h^t(x) e_h^t(x)\right] &\le \mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \widehat{\mu}_h^t(x) e_h^t(x)\right] + 2\min\left\{2, \sqrt{|\mathcal{X}|/N}\right\} \\ &\le \mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \widehat{\mu}_h^t(x) e_h^t(x)\right] + 2\bar{\epsilon}_N.\end{aligned}$$

Consequently,

$$\mathrm{JDE}' \le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \widehat{\mu}_h^t(x) e_h^t(x)\right] + 2HT\bar{\epsilon}_N. \tag{G.59}$$

Stage 2: Softmax-Lipschitz to bilinear form. For each (s,x,p) and $\theta_1,\theta_2 \in \Theta$, we derive the required centered softmax-Lipschitz bound by a self-contained interpolation argument. The upper bound in Chen et al. (2023a, Lemma G.1) is a related softmax comparison, but the variance-form constant below follows from the path calculation given here.

(i) Interpolation path and Taylor expansion. Let $R_b^{p,\theta} := \langle \varphi^p(s,x,b), \theta \rangle$, set $g_b := R_b^{p,\theta_1} - R_b^{p,\theta_2}$, and define $\theta_\tau := (1-\tau)\theta_2 + \tau\theta_1$ and $\nu_\tau := \nu^{p,\theta_\tau}$ for $\tau \in [0,1]$. Because $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$ is convex, the entire path θ_τ remains in Θ. For $m_\tau := \langle \nu_\tau, g \rangle_{\mathcal{B}}$ and $S_\tau := \mathbb{V}_{b\sim\nu_\tau}(g_b)$, direct differentiation of the softmax path gives

$$\dot{\nu}_\tau(b) = \eta \nu_\tau(b)(g_b - m_\tau), \qquad \ddot{\nu}_\tau(b) = \eta^2 \nu_\tau(b)\big((g_b - m_\tau)^2 - S_\tau\big).$$

Consequently, $\|\dot{\nu}_0\|_1 \le \eta \langle \nu_0, |g - m_0| \rangle_{\mathcal{B}}$ and $\|\ddot{\nu}_\tau\|_1 \le 2\eta^2 S_\tau$. Taylor's formula with integral remainder therefore yields

$$\|\nu_1 - \nu_0\|_1 \le \eta \langle \nu_0, |g - m_0| \rangle_{\mathcal{B}} + 2\eta^2 \int_0^1 (1-\tau) S_\tau \, \mathrm{d}\tau. \tag{G.60}$$

(ii) Variance control along the path, uniformly over candidate parameters. By Lemma H.7, every path advantage lies in $[-B_A^\Theta, 0]$, and hence

$$\frac{\nu_\tau(b)}{\nu_0(b)} = \exp\big(\eta\big(A^p_{\theta_\tau}(s,x,b) - A^p_{\theta_2}(s,x,b)\big)\big) \le e^{\eta B_A^\Theta}.$$

Since variance minimizes the squared deviation over constant centers, it follows that

$$S_\tau \le \langle \nu_\tau, (g - m_0)^2 \rangle_{\mathcal{B}} \le e^{\eta B_A^\Theta} \langle \nu_0, (g - m_0)^2 \rangle_{\mathcal{B}} = e^{\eta B_A^\Theta} S_0.$$

(iii) Centering and the final variance form. Set $\kappa^* := -\langle \nu^{p,\theta_2}, R^{p,\theta_2} - R^{p,\theta_1} \rangle_{\mathcal{B}} = m_0$. Then $R^{p,\theta_2} - R^{p,\theta_1} + \kappa^* = -(g - m_0)$, and therefore

$$S_0 = \big\langle \nu^{p,\theta_2}, (R^{p,\theta_2} - R^{p,\theta_1} + \kappa^*)^2 \big\rangle_{\mathcal{B}} = \|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}}.$$

Jensen's inequality gives

$$\langle \nu_0, |g - m_0| \rangle_{\mathcal{B}} \le \sqrt{S_0} = \sqrt{\|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}}}. \tag{G.61}$$

Substituting the path variance bound and (G.61) into (G.60), and using $C_{\mathrm{JDE}} = \eta^2 e^{\eta B_A^\Theta}$, yields

$$\begin{aligned} \big\|\nu^{p,\theta_1}(\cdot|x;s) - \nu^{p,\theta_2}(\cdot|x;s)\big\|_1 &\le \eta \sqrt{\|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}}} + C_{\mathrm{JDE}} \|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}} \\ &\le 2\eta \sqrt{\|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}}} + C_{\mathrm{JDE}} \|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}}. \end{aligned} \tag{G.62}$$

Here C_{JDE} is defined in (C.18), and B_A^Θ is defined in (C.17). Setting $\theta_1 = \bar{\theta}_h^t$ and $\theta_2 = \theta_h^*$ in (G.62) gives a per-x bound on $e_h^t(x)$.

Stage 3: Aggregate over $\widehat{\mu}_h^t$ via Jensen. For the empirical-weighted JDE in (G.59), applying (G.62) per state and aggregating with Jensen's inequality (the function $\sqrt{\cdot}$ is concave):

$$\begin{aligned} \sum_x \widehat{\mu}_h^t(x)\, e_h^t(x) &\le 2\eta \sum_x \widehat{\mu}_h^t(x) \sqrt{\big\|\bar{\theta}_h^t - \theta_h^*\big\|^2_{\Sigma^{p_h^t,\theta_h^*}_{x,s_h^t}}} + C_{\mathrm{JDE}} \sum_x \widehat{\mu}_h^t(x) \big\|\bar{\theta}_h^t - \theta_h^*\big\|^2_{\Sigma^{p_h^t,\theta_h^*}_{x,s_h^t}} \\ &\le 2\eta \sqrt{\big\|\bar{\theta}_h^t - \theta_h^*\big\|^2_{\widehat{\Gamma}_h^t(\theta_h^*)}} + C_{\mathrm{JDE}} \big\|\bar{\theta}_h^t - \theta_h^*\big\|^2_{\widehat{\Gamma}_h^t(\theta_h^*)}, \end{aligned} \tag{G.63}$$

where the second inequality uses Jensen on $\sqrt{\cdot}$ and the linearity of expectation in the quadratic, and we have introduced the *empirical Fisher matrix at the true parameter* θ^*

$$\widehat{\Gamma}_h^t(\theta_h^*) := \frac{1}{N} \sum_{n=1}^N \Sigma^{p_h^t,\theta_h^*}_{x_h^{t,n},s_h^t} = \mathbb{E}_{X\sim\widehat{\mu}_h^t}\big[\Sigma^{p_h^t,\theta_h^*}_{X,s_h^t}\big]. \tag{G.64}$$

Crucially, $\widehat{\Gamma}_h^t(\theta^*)$ is the same empirical Fisher object that appears (up to self-concordance) in the recurrence increment of V_h^t. No uniform-in-p concentration over $\Delta(\mathcal{A})$ is invoked anywhere: the data-realized $p_h^t \in \Delta(\mathcal{A})$ enters only through Jensen on the data-realized weights $\widehat{\mu}_h^t$, hence the analysis is valid for arbitrary discrete (possibly countably infinite) $\mathcal{A}$.

Stage 4: Matrix Cauchy–Schwarz with $A = V_h^t$. For any vector $z \in \mathbb{R}^{d_F}$, any positive definite matrix A, and any positive semidefinite matrix B,

$$\|z\|_B^2 \;=\; z^\top B\, z \;=\; \mathrm{Tr}[B\, zz^\top] \;\le\; \mathrm{Tr}[A^{-1}B] \cdot \|z\|_A^2\,, \tag{G.65}$$

which follows from the cyclic property of trace and PSD ordering.[2] Applying (G.65) with $z = \bar\theta_h^t - \theta_h^*$, $A = V_h^t$, $B = \widehat\Gamma_h^t(\theta_h^*)$:

$$\big\|\bar\theta_h^t - \theta_h^*\big\|^2_{\widehat\Gamma_h^t(\theta_h^*)} \;\le\; \mathrm{Tr}\big[(V_h^t)^{-1}\widehat\Gamma_h^t(\theta_h^*)\big] \cdot \big\|\bar\theta_h^t - \theta_h^*\big\|^2_{V_h^t}\,. \tag{G.66}$$

By the matrix-norm triangle inequality $\|a+b\|_M^2 \le 2\,\|a\|_M^2 + 2\,\|b\|_M^2$, and using both $\bar\theta_h^t, \theta_h^* \in \Theta_h^t$ together with the auxiliary V-norm form (G.16) of Lemma G.5:

$$\big\|\bar\theta_h^t - \theta_h^*\big\|^2_{V_h^t} \;\le\; 2\,\Big\|\bar\theta_h^t - \widehat\theta_h^t\Big\|^2_{V_h^t} + 2\,\Big\|\widehat\theta_h^t - \theta_h^*\Big\|^2_{V_h^t} \;\le\; 2C_{\mathrm{sc}}\beta_{\mathrm{F}}^t + 2C_{\mathrm{sc}}\beta_{\mathrm{F}}^t \;=\; 4C_{\mathrm{sc}}\beta_{\mathrm{F}}^t. \tag{G.67}$$

Note the explicit C_{sc} factor: the algorithmic confidence ellipsoid is in $\widehat F$-norm with radius β_{F}^t (Lemma G.5, eq. (G.14)), and the corresponding V-norm bound is $C_{\mathrm{sc}}\beta_{\mathrm{F}}^t$ via the PSD-equivalence (C.39).

Stage 5: Self-concordance comparison $\widehat\Gamma_h^t(\theta_h^*) \to \widehat\Gamma_h^t(\widehat\theta_h^t)$ and elliptical-potential. The increment of V_h^t at episode t is $V_h^{t+1} - V_h^t = N\eta^2\widehat\Gamma_h^t(\widehat\theta_h^t)$, evaluated at the contemporaneous MLE $\widehat\theta_h^t$. We need to compare $\widehat\Gamma_h^t(\theta_h^*)$ with $\widehat\Gamma_h^t(\widehat\theta_h^t)$ in the positive semidefinite order before applying the matrix elliptical-potential lemma to the recursion for V_h^t. By Lemma G.1 (softmax self-concordance) applied per state in the Loewner (PSD) order with $\big\|\widehat\theta_h^t - \theta_h^*\big\|_2 \le 2B_\Theta$ (a deterministic consequence of $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$ that holds regardless of the random $\widehat\theta_h^t$), and by preservation of Loewner order under nonnegative averaging over the empirical batch:

$$\widehat\Gamma_h^t(\theta_h^*) \;\preceq\; e^{2\eta\cdot 2B_\Theta}\,\widehat\Gamma_h^t(\widehat\theta_h^t) \;=\; e^{4\eta B_\Theta}\,\widehat\Gamma_h^t(\widehat\theta_h^t),$$

which is the analogue of the covariance sandwich in Chen et al. (2023a, Proposition F.1).Using $C_{\mathrm{sc}} = e^{4\eta B_\Theta}$ from (C.18), together with PSD ordering preserved under congruence by $(V_h^t)^{-1/2}$ and under taking traces,

$$\mathrm{Tr}\big[(V_h^t)^{-1}\widehat\Gamma_h^t(\theta_h^*)\big] \;\le\; C_{\mathrm{sc}} \cdot \mathrm{Tr}\big[(V_h^t)^{-1}\widehat\Gamma_h^t(\widehat\theta_h^t)\big]. \tag{G.68}$$

Now apply the matrix elliptical-potential lemma in its *trace* form (Lemma E.6; cf. also Carpentier et al., 2020, Proposition 1, Chen et al., 2023a, Lemma G.10) directly to the per-episode increments

$$X_t \;:=\; \eta^2 \sum_{n=1}^{N} \Sigma^{p_h^t,\widehat\theta_h^t}_{x_h^{t,n},s_h^t} \;=\; N\eta^2\widehat\Gamma_h^t(\widehat\theta_h^t),$$

which satisfy $\mathrm{Tr}(X_t) \le N\eta^2 R_\Sigma =: L_\Sigma$ (with $R_\Sigma = B_\varphi^2$). The recurrence $V_h^{t+1} = V_h^t + X_t$ with $V_h^1 = \lambda_\theta I_{d_F}$ and $\mathrm{Tr}(X_t) \le L_\Sigma$ yields, by Lemma E.6,

$$\sum_{t=1}^{T} \mathrm{Tr}\big[(V_h^t)^{-1}X_t\big] \;\le\; \frac{L_\Sigma/\lambda_\theta}{\log(1 + L_\Sigma/\lambda_\theta)} \cdot d_F\, \log\Big(1 + \frac{TL_\Sigma}{\lambda_\theta d_F}\Big). \tag{G.69}$$

Note the standard $(L_\Sigma/\lambda_\theta)/\log(1+L_\Sigma/\lambda_\theta)$ multiplicative factor. This factor is essential whenever the increments X_t may have trace exceeding λ_θ; it is $O(1)$ when $L_\Sigma = O(\lambda_\theta)$ and grows as $(L_\Sigma/\lambda_\theta)/\log(L_\Sigma/\lambda_\theta)$ when $L_\Sigma/\lambda_\theta \to \infty$. Without this factor, the bound $\sum_t \mathrm{Tr}[(V_h^t)^{-1}X_t] \le d_F\log(\cdot)$ would fail in the scalar example $V_1 = \lambda_\theta = 1$ and $X_1 = L_\Sigma \gg 1$, because $\mathrm{Tr}[V_1^{-1}X_1] = L_\Sigma \gg \log(1+L_\Sigma)$.

[2] Specifically, $\mathrm{Tr}[B\,zz^\top] = \mathrm{Tr}[A^{1/2}A^{-1/2}BA^{-1/2}A^{1/2}zz^\top] = \mathrm{Tr}[A^{-1/2}BA^{-1/2}\cdot A^{1/2}zz^\top A^{1/2}] \le \big\|A^{-1/2}BA^{-1/2}\big\|_{\mathrm{op}} \cdot \mathrm{Tr}[A^{1/2}zz^\top A^{1/2}] \le \mathrm{Tr}[A^{-1}B]\cdot\|z\|_A^2$.

Translating from X_t to $\widehat{\Gamma}_h^t(\widehat{\theta}_h^t)$ via $X_t = N\eta^2\widehat{\Gamma}_h^t(\widehat{\theta}_h^t)$, and combining (G.68) with (G.69), we obtain the *deterministic, coverage-free* aggregate trace bound:

$$\sum_{t=1}^{T} \mathrm{Tr}\big[(V_h^t)^{-1}\widehat{\Gamma}_h^t(\theta_h^*)\big] \;\le\; \frac{C_{\mathrm{sc}}}{N\eta^2}\cdot\frac{L_\Sigma/\lambda_\theta}{\log(1+L_\Sigma/\lambda_\theta)}\cdot d_F\,\log\Big(1+\frac{TL_\Sigma}{\lambda_\theta d_F}\Big) \tag{G.70}$$
$$= \frac{C_{\mathrm{sc}}\,R_\Sigma/\lambda_\theta}{\log(1+N\eta^2R_\Sigma/\lambda_\theta)}\cdot d_F\,\log\Big(1+\frac{TN\eta^2R_\Sigma}{\lambda_\theta d_F}\Big).$$

The $1/(N\eta^2)$ factor from (G.68) is exactly cancelled by the $L_\Sigma = N\eta^2R_\Sigma$ in the numerator of the elliptical-potential bound, leaving only a $1/\log(1+N\eta^2R_\Sigma/\lambda_\theta)$ residual saving. Under $N\eta^2R_\Sigma \ge \lambda_\theta$, the bound (G.70) simplifies to

$$\sum_{t=1}^{T} \mathrm{Tr}\big[(V_h^t)^{-1}\widehat{\Gamma}_h^t(\theta_h^*)\big] \;\le\; \frac{C_{\mathrm{sc}}\,R_\Sigma}{\lambda_\theta\log 2}\cdot d_F\,\log\Big(1+\frac{TN\eta^2R_\Sigma}{\lambda_\theta d_F}\Big) \;=\; \widetilde{O}\big(C_{\mathrm{sc}}\,R_\Sigma\,d_F/\lambda_\theta\big),$$

where the $\widetilde{O}$ absorbs the $\log(1+TN\eta^2R_\Sigma/(\lambda_\theta d_F))$ factor. No trace residual remains: the empirical-to-population transition was already absorbed into the additive $2HT\bar{\epsilon}_N$ term in Stage 1.

Stage 5.5 (martingale concentration bound from $\sum_t \mathbb{E}^t[\cdot]$ to $\sum_t \cdot$, Bernstein form). The JDE definition (G.28)/(G.57) involves the conditional expectations $\mathbb{E}^t[\cdot]$ over the trajectory at episode t, while the deterministic bounds in Stages 2–5 above and the Cauchy–Schwarz step in Stage 6 below apply to realized trajectory values. We compare the two using the *Bernstein-style* branch (E.11) of Lemma E.3, whose multiplicative form does not produce a $\sqrt{T}$ residual when applied to nonnegative bounded processes. The proof of Chen et al. (2023a, Theorem 6.1) in Appendix D.2 uses equation (G.15) of Corollary G.8 for the analogous second-order follower-response term, while its first-order term is handled separately by equation (G.14).

Define the per-step JDE integrand at the empirical-weighted level (the deterministic bound established by Stages 1–3):

$$Y_h^t := 2\eta\sqrt{S_h^t} \;+\; C_{\mathrm{JDE}}\,S_h^t, \qquad S_h^t := \big\|\bar{\theta}_h^t - \theta_h^*\big\|^2_{\widehat{\Gamma}_h^t(\theta_h^*)}.$$

By construction $Y_h^t \ge 0$. From $\big\|\bar{\theta}_h^t - \theta_h^*\big\|_2 \le 2B_\Theta$ (since $\bar{\theta}_h^t, \theta_h^* \in \Theta = \overline{\mathbb{B}(0,B_\Theta)}$) and $\big\|\widehat{\Gamma}_h^t(\theta_h^*)\big\|_{\mathrm{op}} \le R_\Sigma$ (with $R_\Sigma = B_\varphi^2$), we have $S_h^t \le 4B_\Theta^2R_\Sigma$ and hence $Y_h^t \le R_Y$. Applying Lemma E.3 eq. (E.11) (the Bernstein form for non-negative bounded processes) per step $h \in [H]$ to $X_t = Y_h^t \in [0, R_Y]$, with $R = R_Y$ and confidence ζ/H, gives with probability at least $1-\zeta/H$

$$\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}^t[Y_h^t] - 2R_Y\log(2H/\zeta) \;\le\; \sum_{t=1}^{T} Y_h^t.$$

Rearranging this left inequality yields

$$\sum_{t=1}^{T}\mathbb{E}^t\big[Y_h^t\big] \;\le\; 2\sum_{t=1}^{T} Y_h^t + 4R_Y\log(2H/\zeta). \tag{G.71}$$

Union-bounding over $h \in [H]$ defines the event $\mathcal{E}_{\text{JDE-mart}}$ with probability at least $1-\zeta$. The residual $4R_Y H\log(2H/\zeta) = \widetilde{O}(H)$ after summing over h is T-independent up to logarithmic factors and is retained explicitly in (G.57).

Stage 6: Sum over t via Cauchy–Schwarz, then over $h \in [H]$. On the joint event $\mathcal{E}_{\text{MLE,main}} \cap \mathcal{E}_{\text{JDE-mart}}$ (probability at least $1-2\zeta$), starting from (G.59), applying Stages 2–3 (eq. (G.63)) and the Bernstein concentration bound (G.71) of Stage 5.5, we obtain for each fixed $h \in [H]$:

$$\sum_{t=1}^{T}\mathbb{E}^t\Big[\sum_x \widehat{\mu}_h^t(x)\,e_h^t(x)\Big] \;\le\; \sum_{t=1}^{T}\mathbb{E}^t[Y_h^t] \;\le\; 2\sum_{t=1}^{T} Y_h^t + 4R_Y\log(2H/\zeta).$$

Now combining (G.66), (G.67), and (G.70) on the realized trajectory, and applying Cauchy–Schwarz $\sum_t y_t \le \sqrt{T \sum_t y_t^2}$ to the linear (square-root) part of $\sum_t Y_h^t$:

$$\begin{aligned}\sum_{t=1}^{T} 2\eta \sqrt{\|\bar\theta_h^t - \theta_h^*\|^2_{\widehat\Gamma_h^t(\theta_h^*)}} &\le 2\eta \sqrt{T \sum_{t=1}^{T} \|\bar\theta_h^t - \theta_h^*\|^2_{\widehat\Gamma_h^t(\theta_h^*)}} \\ &\le 2\eta \sqrt{T \cdot 4C_{\mathrm{sc}}\, \beta_{\mathrm{F}}^T \cdot \frac{C_{\mathrm{sc}}\, R_\Sigma/\lambda_\theta}{\log(1+N\eta^2 R_\Sigma/\lambda_\theta)} \cdot d_F \log\left(1 + \frac{TN\eta^2 R_\Sigma}{\lambda_\theta d_F}\right)} \\ &= 4\eta\, C_{\mathrm{sc}} \sqrt{\frac{\beta_{\mathrm{F}}^T\, d_F\, T\, R_\Sigma/\lambda_\theta}{\log(1+N\eta^2 R_\Sigma/\lambda_\theta)} \cdot \log\left(1 + \frac{TN\eta^2 R_\Sigma}{\lambda_\theta d_F}\right)}.\end{aligned}$$

Similarly the second-order C_{JDE} part of $\sum_t Y_h^t$ gives $\sum_{t=1}^T C_{\mathrm{JDE}} \|\bar\theta_h^t - \theta_h^*\|^2_{\widehat\Gamma_h^t(\theta_h^*)} \le 4C_{\mathrm{JDE}}\, C_{\mathrm{sc}}^2\, \beta_{\mathrm{F}}^T \cdot \frac{R_\Sigma/\lambda_\theta}{\log(1+N\eta^2 R_\Sigma/\lambda_\theta)} \cdot d_F \log\big(1 + \frac{TN\eta^2 R_\Sigma}{\lambda_\theta d_F}\big)$ *deterministically* (no Cauchy–Schwarz over t required). Multiplying both displays by the Bernstein factor 2 from (G.71), summing over $h \in [H]$, and combining with the additive Bernstein residual $4R_Y H \log(2H/\zeta) = \widetilde O(H)$ together with the Stage 1 residual from the TV perturbation bound $2HT\bar\epsilon_N$, we obtain (using $\lesssim$ to absorb the universal Bernstein constant 2):

$$\begin{aligned}\mathrm{JDE}' \lesssim\; & C_{\mathrm{sc}}\eta H \sqrt{\frac{\beta_{\mathrm{F}}^T d_F T R_\Sigma/\lambda_\theta}{\log(1+N\eta^2 R_\Sigma/\lambda_\theta)} \log\big(1 + \frac{TN\eta^2 R_\Sigma}{\lambda_\theta d_F}\big)} \\ & + \frac{C_{\mathrm{JDE}} C_{\mathrm{sc}}^2 \beta_{\mathrm{F}}^T d_F H R_\Sigma/\lambda_\theta}{\log(1+N\eta^2 R_\Sigma/\lambda_\theta)} \log\big(1 + \frac{TN\eta^2 R_\Sigma}{\lambda_\theta d_F}\big) + HT\bar\epsilon_N + R_Y H \log(2H/\zeta)\,,\end{aligned}$$

which is the form claimed in (G.57); in particular, the Bernstein residual remains explicit rather than being absorbed into another term.

Concluding remark on regime simplification. In the entropy-regularized regime where $\eta = O(\log|\mathcal{B}|)$ and $N\eta^2 R_\Sigma \ge \lambda_\theta$ (the data-abundant trace-factor regime used in the rate simplification), the factor $1/\log(1 + N\eta^2 R_\Sigma/\lambda_\theta) = O(1)$, and absorbing problem constants $(C_{\mathrm{sc}}, R_\Sigma, 1/\lambda_\theta, \eta)$ into $\widetilde O$-notation, the leading term simplifies to

$$\mathrm{JDE}' \;\le\; \widetilde O\Big(C_{\mathrm{sc}}\eta H \sqrt{\beta_{\mathrm{F}}^T d_F T}\Big) + \widetilde O\big(C_{\mathrm{sc}}^2 C_{\mathrm{JDE}} \beta_{\mathrm{F}}^T d_F H\big) + \widetilde O(HT\bar\epsilon_N) + \widetilde O(R_Y H),$$

where the first two terms have the same dependence on T as the cumulative follower-response error bound underlying Chen et al. (2023a, Theorem 6.1): a $\widetilde O(\sqrt{T})$ first-order term and a second-order term that grows only polylogarithmically with T. Note that no polynomial $1/N$ improvement appears in the JDE leading term: the $1/(N\eta^2)$ factor inherited from (G.68) is exactly cancelled by the $L_\Sigma = N\eta^2 R_\Sigma$ in the elliptical-potential numerator, leaving only the $1/\log(1 + N\eta^2 R_\Sigma/\lambda_\theta)$ residual. The genuine N-dependent improvement in the FSO regret comes from the additive $2HT\bar\epsilon_N$ residual, which is controlled by $\bar\epsilon_N = \widetilde O(\sqrt{|\mathcal{Z}|/N})$, and the leader-side finite-sample terms collected by $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})HT\bar\epsilon_N$, not from the JDE leading term. □

G.7 Main FSO regret theorem

Proof of Theorem 4.1. Let

$$\mathcal{E} := \mathcal{E}_{\mathrm{or}} \cap \mathcal{E}_{\mathrm{emp}} \cap \mathcal{E}_{\mathrm{MLE,main}} \cap \mathcal{E}_{\mathrm{LBE\text{-}mart}} \cap \mathcal{E}_{\mathrm{JDE\text{-}mart}}.$$

By Lemmas G.3, G.5, G.16 and H.1, the first four constituent events each have failure probability at most ζ. The construction of $\mathcal{E}_{\mathrm{JDE\text{-}mart}}$ in the proof of Lemma G.17, equation (G.71), gives $\mathbb{P}(\mathcal{E}_{\mathrm{JDE\text{-}mart}}) \ge 1 - \zeta$. Thus $\mathbb{P}(\mathcal{E}) \ge 1 - 5\zeta$ by the union bound.

Step 1: Regret decomposition. On $\mathcal{E}$, the conditions $\mathcal{E}_{\mathrm{or}}^t \cap \mathcal{E}_{\mathrm{emp}}^{<t}$ and $\theta_h^* \in \Theta_h^t$ hold for every (t,h). Therefore Lemma G.9, equation (G.27), gives

$$\mathrm{Reg}_{\mathrm{FSO}}(T) \le \mathrm{LBE}' + L_U^{\mathrm{FSO}} \mathrm{JDE}' + \mathrm{Bias}_{\mathrm{batch}}. \tag{G.72}$$

Step 2: Leader Bellman error. By Lemma G.16, equation (G.51),

$$\begin{aligned}\mathrm{LBE}' \le & 4B_T^{\mathrm{FSO}} H\sqrt{TG_K(T,\lambda)} + 2B_T^{\mathrm{FSO}} B_\phi H\sqrt{T}\log(2eH/\zeta) \\ & + \left[2B_T^{\mathrm{FSO}} L_\phi \lambda^{-1/2} + L_U^{\mathrm{FSO}}\right] HT\bar{\epsilon}_N. \end{aligned} \tag{G.73}$$

Step 3: Follower joint-distribution error. By Lemma G.17, equation (G.57), and the continuous convention for $L_{T,N}$,

$$\mathrm{JDE}' \lesssim 4C_{\mathrm{sc}}\eta H\sqrt{\beta_{\mathrm{F}}^T d_F T L_{T,N}} + 4C_{\mathrm{sc}}^2 C_{\mathrm{JDE}}\beta_{\mathrm{F}}^T d_F H L_{T,N} + 2HT\bar{\epsilon}_N + 4R_Y H\log(2H/\zeta)\,. \tag{G.74}$$

Multiplying (G.74) by L_U^{FSO} produces the first two terms and the Bernstein term in (4.1), together with $2L_U^{\mathrm{FSO}} HT\bar{\epsilon}_N$.

Step 4: Current-batch mismatch. By Lemma G.9, equation (G.29),

$$\mathrm{Bias}_{\mathrm{batch}} \le 2C_{\mathrm{pair}}^{\mathrm{env}}(B_T^{\mathrm{FSO}}) HT\bar{\epsilon}_N. \tag{G.75}$$

The current-batch first-moment input used in this bound is Lemma G.2; it is taken inside $\mathbb{E}^t$ and does not condition on the current empirical event.

Step 5: Exact finite-sample coefficient. Using $L_U^{\mathrm{FSO}} = L_U(B_T^{\mathrm{FSO}})$, the total coefficient multiplying $HT\bar{\epsilon}_N$ in (G.73), (G.74), and (G.75) is

$$2B_T^{\mathrm{FSO}} L_\phi \lambda^{-1/2} + L_U(B_T^{\mathrm{FSO}}) + 2L_U(B_T^{\mathrm{FSO}}) + 2C_{\mathrm{pair}}^{\mathrm{env}}(B_T^{\mathrm{FSO}}) = C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})$$

by (C.40). Substituting (G.73), (G.74), and (G.75) into (G.72) proves (4.1).

Step 6: Simplified rate. Under the additional conditions stated before (4.2), one has $L_{T,N} = \widetilde{O}(1)$, $B_T^{\mathrm{FSO}} = \widetilde{O}(H\sqrt{\mathfrak{D}_T})$, and $\beta_{\mathrm{F}}^T = \widetilde{O}(d_F)$. The sharpened base-radius identity gives $L_U^{\mathrm{FSO}} = 2L_\phi B_T^{\mathrm{FSO}} \lambda^{-1/2} = \widetilde{O}(H\sqrt{\mathfrak{D}_T})$. Since $C_{\mathrm{pair}}^{\mathrm{env}}(\mathsf{r}) = L_U(\mathsf{r})$, (C.40) also gives $C_{\mathrm{FS}}(B_T^{\mathrm{FSO}}) = \widetilde{O}(H\sqrt{\mathfrak{D}_T})$. Substituting these relations into (4.1), using $G_K^+ = 1 \vee G_K$, and absorbing the lower-order leader-martingale and Bernstein terms proves (4.2). □

Proof of Corollary 4.3. Recall $D_L = 1 + d_L + d_L^2 + d_F^2$ from Section 4.1. The finite-rank complete-cover bounds in Remark C.7 give $G_K^+(T,\lambda) = \widetilde{O}(d_L)$ and $\mathfrak{D}_T = \widetilde{O}(D_L)$. By Proposition C.4, (G.41), and the radius-absorption conditions,

$$B_T^{\mathrm{FSO}} = \widetilde{O}\left(H\sqrt{D_L}\right).$$

The sharpened base-radius definition (C.19) gives

$$L_U^{\mathrm{FSO}} = 2L_\phi B_T^{\mathrm{FSO}} \lambda^{-1/2} = \widetilde{O}\left(H\sqrt{D_L}\right),$$

where the fixed-primitive-constant condition bounds L_ϕ and λ^{-1}. Because $C_{\mathrm{pair}}^{\mathrm{env}}(\mathsf{r}) = L_U(\mathsf{r})$, equation (C.40) also gives

$$C_{\mathrm{FS}}(B_T^{\mathrm{FSO}}) = \widetilde{O}\left(H\sqrt{D_L}\right).$$

By Lemma G.5, the well-specified MLE radius satisfies $\beta_{\mathrm{F}}^T = \widetilde{O}(d_F)$. Under the corollary's condition on R_Σ and N, the continuous factor satisfies $L_{T,N} = \widetilde{O}(1)$; when $R_\Sigma = 0$, both Fisher-dependent JDE terms vanish.

Substituting these relations into the first JDE term of (4.1) gives

$$L_U^{\mathrm{FSO}} H\sqrt{\beta_{\mathrm{F}}^T d_F T L_{T,N}} = \widetilde{O}\left(d_F H^2\sqrt{TD_L}\right).$$

The second JDE term becomes

$$L_U^{\mathrm{FSO}} \beta_{\mathrm{F}}^T d_F H L_{T,N} = \widetilde{O}\left(d_F^2 H^2\sqrt{D_L}\right).$$

The leading leader Bellman-estimation term becomes

$$B_T^{\mathrm{FSO}} H\sqrt{TG_K(T,\lambda)} = \widetilde{O}\Big(H^2\sqrt{Td_L D_L}\Big).$$

By (G.12), the finite-sample term satisfies

$$C_{\mathrm{FS}}(B_T^{\mathrm{FSO}})HT\bar{\epsilon}_N = \widetilde{O}\Big(H^2 T\sqrt{D_L|\mathcal{Z}|/N}\Big).$$

The leader bonus-martingale term is $\widetilde{O}(H^2\sqrt{TD_L})$ and is dominated by the displayed leader term. The JDE Bernstein residual is $\widetilde{O}(H^2\sqrt{D_L})$ and is dominated by the displayed second-order term because $d_F \geq 1$. Combining these bounds proves (4.4). □

H Further auxiliary results and proofs

H.1 Setting-indexed RKHS uncertainty and function-class containment

Setting-indexed oracle notation. For $\iota \in \{\mathrm{CO}, \mathrm{FSO}\}$, define the oracle radius by $B^{\mathrm{or}}_{T,\mathrm{CO}} := B_T^{\mathrm{CO}}$ and $B^{\mathrm{or}}_{T,\mathrm{FSO}} := B_T^{\mathrm{base}}$. Let $z^\tau_{h,\mathrm{CO}} := z^\tau_h$ and $z^\tau_{h,\mathrm{FSO}} := \widehat{z}^\tau_h$, and set

$$\widehat{\Lambda}^t_{L,h,\iota} := \lambda I_{\mathbb{H}} + \sum_{\tau<t} \phi(z^\tau_{h,\iota}) \otimes \phi(z^\tau_{h,\iota}).$$

For a bounded continuation value W, set $f^*_{h,W} := \mathbb{B}_h W$ and $\xi_{h,\tau}(W) := W(s^\tau_{h+1}, \mu^\tau_{h+1}) - (P_h W)(z^\tau_h)$, where z^τ_h is the true transition query. Let $\widehat{f}^{t,\mathrm{or}}_{h,\iota}(W)$ be the KRR estimator built from this design and the oracle labels $f^*_{h,W}(z^\tau_{h,\iota}) + \xi_{h,\tau}(W)$, $\tau < t$. For $j \in [H]$, let $\mathfrak{V}_{j,\mathrm{CO}} := \mathcal{V}_{\mathrm{UCB,CO}}(j, R^{\mathrm{base}}_{T,\mathrm{CO}}(B_T^{\mathrm{CO}}), B_T^{\mathrm{CO}}, 0)$ and $\mathfrak{V}_{j,\mathrm{FSO}} := \mathcal{V}_{\mathrm{UCB,FSO}}(j, R^{\mathrm{cap}}_{T,\mathrm{FSO}}(B_T^{\mathrm{base}}), B_T^{\mathrm{cap}}, C_{\mathrm{off}}^{\mathrm{cap}})$. Set $\mathfrak{V}_{H+1,\iota} := \{0\}$ for each $\iota \in \{\mathrm{CO}, \mathrm{FSO}\}$ and set $\zeta^{\mathrm{or}}_{t,h} := \zeta/(HT)$. Define the deterministic uniform-noise loads by

$$\begin{aligned}
\mathcal{R}_{T,\mathrm{CO}} &:= 4G_K(T,\lambda) + 4\log(HT/\zeta) + 10 \\
&\quad + 4\max_{h\in[H]} \log N_{\infty,\mathrm{CO}}\left(\varepsilon_{\mathrm{cov},T}; h+1, R^{\mathrm{base}}_{T,\mathrm{CO}}(B_T^{\mathrm{CO}}), B_T^{\mathrm{CO}}, 0\right), \\
\mathcal{R}_{T,\mathrm{FSO}} &:= 4G_K(T,\lambda) + 4\log(HT/\zeta) + 10 \\
&\quad + 4\max_{h\in[H]} \log N_{\infty,\mathrm{FSO}}\left(\varepsilon_{\mathrm{cov},T}; h+1, R^{\mathrm{cap}}_{T,\mathrm{FSO}}(B_T^{\mathrm{base}}), B_T^{\mathrm{cap}}, C_{\mathrm{off}}^{\mathrm{cap}}\right).
\end{aligned} \tag{H.1}$$

Lemma H.1 (Setting-indexed oracle RKHS uncertainty and base-function control)**.** *Assume Assumptions 2.2 and 2.5, $\sup_z K(z,z) \leq 1$, and $\lambda \geq (T+1)/T$. Assume that both setting-specific historical inputs are measurable before the corresponding step transition is sampled. For the CO conclusion, assume that $N_{\infty,\mathrm{CO}}$ is finite at $\varepsilon_{\mathrm{cov},T}$ and that (C.28) has a finite calibration; impose the analogous conditions on $N_{\infty,\mathrm{FSO}}$ and (C.25) for the FSO conclusion. For every $\iota \in \{\mathrm{CO}, \mathrm{FSO}\}$ and $(t,h) \in [T]\times[H]$, there is an event $\mathcal{E}^{t,h}_{\mathrm{or},\iota} \in \mathcal{F}^{t-1}$ on which the uniform self-normalized noise bound*

$$\sup_{W\in\mathfrak{V}_{h+1,\iota}} \left\| \sum_{\tau<t} \phi(z^\tau_{h,\iota})\xi_{h,\tau}(W) \right\|_{(\widehat{\Lambda}^t_{L,h,\iota})^{-1}} \leq H\sqrt{\mathcal{R}_{T,\iota}} \tag{H.2}$$

holds simultaneously with the following conclusions for every z and $W \in \mathfrak{V}_{h+1,\iota}$:

$$\left|\widehat{f}^{t,\mathrm{or}}_{h,\iota}(W)(z) - (\mathbb{B}_h W)(z)\right| \leq B^{\mathrm{or}}_{T,\iota}\|\phi(z)\|_{(\widehat{\Lambda}^t_{L,h,\iota})^{-1}}, \tag{H.3}$$

$$\left\|\widehat{f}^{t,\mathrm{or}}_{h,\iota}(W)\right\|_{\mathbb{H}} \leq B^{\mathrm{or}}_{T,\iota}\lambda^{-1/2}. \tag{H.4}$$

Moreover, $\mathbb{P}((\mathcal{E}^{t,h}_{\mathrm{or},\iota})^c) \leq \zeta^{\mathrm{or}}_{t,h}$. Consequently, with $\mathcal{E}^t_{\mathrm{or},\iota} := \bigcap_{h=1}^H \mathcal{E}^{t,h}_{\mathrm{or},\iota}$ and $\mathcal{E}_{\mathrm{or},\iota} := \bigcap_{t=1}^T \mathcal{E}^t_{\mathrm{or},\iota}$, one has $\mathbb{P}(\mathcal{E}_{\mathrm{or},\iota}) \geq 1-\zeta$ for each fixed setting ι.

The norm conclusion in (H.4) is derived from the uniform noise-norm event (H.2) and the KRR normal equation; it is not inferred from the pointwise display (H.3). For CO, $\widehat{\Lambda}^t_{L,h,\mathrm{CO}} = \Lambda^t_{L,h}$ and $\widehat{f}^{t,\mathrm{or}}_{h,\mathrm{CO}}(\widehat{W}^t_{h+1}) = \widehat{f}^t_h$ because the realized CO input and label are already the oracle input and label. Moreover, $\Gamma^t_{h,\mathrm{CO}}(z) := B_T^{\mathrm{CO}} \|\phi(z)\|_{(\widehat{\Lambda}^t_{L,h,\mathrm{CO}})^{-1}} = \Gamma^t_h(z)$. On $\mathcal{E}_{\mathrm{or,CO}}$, backward induction from $\widehat{W}^t_{H+1} \equiv 0$ uses (H.4), the CO planning rule in Algorithm 2, the version-space definition (3.1), and its membership in the deterministic feasible-set family (C.13) to give

$$\begin{aligned} \widehat{U}^t_h &\in \mathcal{U}_{\mathrm{UCB}}(h, R^{\mathrm{base}}_{T,\mathrm{CO}}(B_T^{\mathrm{CO}}), B_T^{\mathrm{CO}}, 0), \\ \widehat{W}^t_h &\in \mathcal{V}_{\mathrm{UCB,CO}}(h, R^{\mathrm{base}}_{T,\mathrm{CO}}(B_T^{\mathrm{CO}}), B_T^{\mathrm{CO}}, 0) \end{aligned} \tag{H.5}$$

for every (t, h).

Proof of Lemma H.1. Fix a setting $\iota \in \{\mathrm{CO}, \mathrm{FSO}\}$, an episode t, and a step h. Let $\mathfrak{V}^{\varepsilon}_{h+1,\iota}$ be a minimum-cardinality ambient $\varepsilon_{\mathrm{cov},T}$-net of the deterministic class $\mathfrak{V}_{h+1,\iota}$ in the supremum norm. Choose its centers as bounded Borel functions and truncate them pointwise to $[0, H]$, which does not increase the covering radius or cardinality. For each fixed $\widetilde{W} \in \mathfrak{V}^{\varepsilon}_{h+1,\iota}$, define

$$\xi^{\iota}_{h,\tau}(\widetilde{W}) := \widetilde{W}(s^{\tau}_{h+1}, \mu^{\tau}_{h+1}) - (P_h \widetilde{W})(z^{\tau}_h).$$

Let $\mathcal{G}_{h,\tau,\iota}$ be the setting-specific history immediately before the step-h transition outcome in episode τ is revealed. The sigma-field $\mathcal{G}_{h,\tau,\iota}$ contains the true query z^{τ}_h and the predictable design input $z^{\tau}_{h,\iota}$, but it excludes $(s^{\tau}_{h+1}, \mu^{\tau}_{h+1})$. For CO, the two inputs coincide. For FSO, the true query is z^{τ}_h and the predictable regression design is $z^{\tau}_{h,\mathrm{FSO}} = \widehat{z}^{\tau}_h$. By Assumption 2.2,

$$\mathbb{E}\Big[\xi^{\iota}_{h,\tau}(\widetilde{W}) \mid \mathcal{G}_{h,\tau,\iota}\Big] = 0, \qquad |\xi^{\iota}_{h,\tau}(\widetilde{W})| \le H.$$

Hoeffding's lemma therefore makes $\xi^{\iota}_{h,\tau}(\widetilde{W})$ conditionally H-sub-Gaussian given $\mathcal{G}_{h,\tau,\iota}$.

Step 1: Uniform self-normalized noise event. Set $\delta^{\mathrm{net}}_{t,h,\iota} := \zeta^{\mathrm{or}}_{t,h} / |\mathfrak{V}^{\varepsilon}_{h+1,\iota}|$. For each fixed $\widetilde{W} \in \mathfrak{V}^{\varepsilon}_{h+1,\iota}$, apply Lemma E.4 at failure probability $\delta^{\mathrm{net}}_{t,h,\iota}$ using the normalized features $\phi(z^{\tau}_{h,\iota})/\sqrt{\lambda}$ and a deterministic sequence of auxiliary regularization parameters $\eta_n \downarrow 0$. Let $F_n(\widetilde{W})$ denote the failure event for the resulting finite-η_n inequality. Writing $\widetilde{K}^t_{h,\iota} := K^t_{h,\iota}/\lambda$, the spectral calculus gives

$$\Big[(\widetilde{K}^t_{h,\iota} + \eta_n I)^{-1} + I\Big]^{-1} \longrightarrow \widetilde{K}^t_{h,\iota}(I + \widetilde{K}^t_{h,\iota})^{-1} = K^t_{h,\iota}(K^t_{h,\iota} + \lambda I)^{-1}.$$

Let $\boldsymbol{\xi} := (\xi^{\iota}_{h,\tau}(\widetilde{W}))_{\tau<t}$. Consequently, the limiting quadratic form is

$$\boldsymbol{\xi}^{\top} K^t_{h,\iota}(K^t_{h,\iota} + \lambda I)^{-1} \boldsymbol{\xi} = \Big\| \sum_{\tau<t} \phi(z^{\tau}_{h,\iota}) \xi^{\iota}_{h,\tau}(\widetilde{W}) \Big\|^2_{(\widehat{\Lambda}^t_{L,h,\iota})^{-1}},$$

and the determinant term converges to $\log\det(I + K^t_{h,\iota}/\lambda)$. Let $F_0(\widetilde{W})$ denote the failure event for the limiting inequality. The convergence of the quadratic-form and determinant terms gives $F_0(\widetilde{W}) \subseteq \liminf_{n\to\infty} F_n(\widetilde{W})$. Fatou's lemma for the failure-event indicators therefore gives

$$\mathbb{P}(F_0(\widetilde{W})) \le \mathbb{P}\Big(\liminf_{n\to\infty} F_n(\widetilde{W})\Big) \le \liminf_{n\to\infty} \mathbb{P}(F_n(\widetilde{W})) \le \delta^{\mathrm{net}}_{t,h,\iota}.$$

Take a union bound over $\widetilde{W} \in \mathfrak{V}^{\varepsilon}_{h+1,\iota}$. Using $\log\det(I + K^t_{h,\iota}/\lambda) \le 2G_K(T, \lambda)$ gives, simultaneously for every net element,

$$\Big\| \sum_{\tau<t} \phi(z^{\tau}_{h,\iota}) \xi^{\iota}_{h,\tau}(\widetilde{W}) \Big\|^2_{(\widehat{\Lambda}^t_{L,h,\iota})^{-1}} \le 2H^2 G_K(T, \lambda) + 2H^2 \log\left(\frac{|\mathfrak{V}^{\varepsilon}_{h+1,\iota}|}{\zeta^{\mathrm{or}}_{t,h}}\right). \tag{H.6}$$

For any $W \in \mathfrak{V}_{h+1,\iota}$, choose $\widetilde{W} \in \mathfrak{V}^{\varepsilon}_{h+1,\iota}$ with $\|W - \widetilde{W}\|_\infty \le \varepsilon_{\mathrm{cov},T}$. Set $d_\tau := \xi^{\iota}_{h,\tau}(W) - \xi^{\iota}_{h,\tau}(\widetilde{W})$, so $|d_\tau| \le 2\varepsilon_{\mathrm{cov},T}$. If $\Phi^t_{h,\iota}$ denotes the design operator, then

$$\left\|(\Phi^t_{h,\iota})^* d\right\|^2_{(\widehat{\Lambda}^t_{L,h,\iota})^{-1}} = \left\langle d, \Phi^t_{h,\iota}(\widehat{\Lambda}^t_{L,h,\iota})^{-1}(\Phi^t_{h,\iota})^* d \right\rangle \le \|d\|_2^2 \le 4(t-1)\varepsilon^2_{\mathrm{cov},T},$$

where the first inequality uses $\Phi^t_{h,\iota}(\widehat{\Lambda}^t_{L,h,\iota})^{-1}(\Phi^t_{h,\iota})^* \preceq I$. Because $\varepsilon_{\mathrm{cov},T} = \min\{H/T, 1/2\}$, one has $T\varepsilon^2_{\mathrm{cov},T} \le H^2$ and hence $\|(\Phi^t_{h,\iota})^* d\|_{(\widehat{\Lambda}^t_{L,h,\iota})^{-1}} \le 2H$. Set $A_{T,\iota} := G_K(T,\lambda) + \log|\mathfrak{V}^{\varepsilon}_{h+1,\iota}| + \log(HT/\zeta)$. Equation (H.6) and the triangle inequality give the full-class bound $H(\sqrt{2A_{T,\iota}} + 2)$. Since $(\sqrt{2A}+2)^2 \le 4A + 10$ for $A \ge 0$, the covering-number definitions in (H.1) give (H.2) with failure probability at most $\zeta^{\mathrm{or}}_{t,h}$.

Step 2: Pointwise oracle uncertainty. Fix $W \in \mathfrak{V}_{h+1,\iota}$ and abbreviate $f^* := \mathbb{B}_h W$, $\Lambda := \widehat{\Lambda}^t_{L,h,\iota}$, and $S := \sum_{\tau<t} \phi(z^\tau_{h,\iota})\xi^{\iota}_{h,\tau}(W)$. The oracle KRR normal equation gives

$$\widehat{f}^{t,\mathrm{or}}_{h,\iota}(W) = (I - \lambda\Lambda^{-1})f^* + \Lambda^{-1}S. \tag{H.7}$$

Hence, for every evaluation point z,

$$\left|\widehat{f}^{t,\mathrm{or}}_{h,\iota}(W)(z) - f^*(z)\right| \le \left(\sqrt{\lambda}\|f^*\|_{\mathbb{H}} + \|S\|_{\Lambda^{-1}}\right)\|\phi(z)\|_{\Lambda^{-1}} \le H\left(\sqrt{\lambda}R_U + \sqrt{\mathcal{R}_{T,\iota}}\right)\|\phi(z)\|_{\Lambda^{-1}},$$

where the second inequality uses Assumption 2.5 and (H.2). By (H.1),

$$2\lambda R_U^2 + 2\mathcal{R}_{T,\mathrm{CO}} = \mathcal{Q}_{T,\mathrm{CO}}(B^{\mathrm{CO}}_T), \qquad 2\lambda R_U^2 + 2\mathcal{R}_{T,\mathrm{FSO}} = \mathcal{Q}_T(B^{\mathrm{base}}_T).$$

The dyadic defining inequalities (C.28) and (C.25), together with $(a+b)^2 \le 2a^2 + 2b^2$, therefore give

$$H\left(\sqrt{\lambda}R_U + \sqrt{\mathcal{R}_{T,\iota}}\right) \le B^{\mathrm{or}}_{T,\iota}. \tag{H.8}$$

This proves (H.3).

Step 3: Oracle KRR norm control on the same event. Since the spectrum of $I - \lambda\Lambda^{-1}$ lies in $[0,1]$ and $\Lambda \succeq \lambda I_{\mathbb{H}}$, equation (H.7) gives

$$\left\|\widehat{f}^{t,\mathrm{or}}_{h,\iota}(W)\right\|_{\mathbb{H}} \le \|f^*\|_{\mathbb{H}} + \|\Lambda^{-1}S\|_{\mathbb{H}} \le R_U H + \lambda^{-1/2}\|S\|_{\Lambda^{-1}} \le R_U H + \frac{H\sqrt{\mathcal{R}_{T,\iota}}}{\sqrt{\lambda}} \le \frac{B^{\mathrm{or}}_{T,\iota}}{\sqrt{\lambda}},$$

where the last inequality is (H.8) divided by $\sqrt{\lambda}$. This proves (H.4) directly from the uniform noise event and the normal equation.

Step 4: Global oracle events. Because $\zeta^{\mathrm{or}}_{t,h} = \zeta/(HT)$, a union bound over (t,h) gives $\mathbb{P}(\mathcal{E}_{\mathrm{or},\iota}) \ge 1 - \zeta$ for each fixed setting ι.

Step 5: CO backward containment. Fix an episode t and work on $\mathcal{E}^t_{\mathrm{or,CO}}$. The terminal continuation value $\widehat{W}^t_{H+1} \equiv 0$ belongs to $\mathfrak{V}_{H+1,\mathrm{CO}}$. Assume inductively that $\widehat{W}^t_{h+1} \in \mathfrak{V}_{h+1,\mathrm{CO}}$. In CO, the realized design and labels equal their oracle counterparts, so $\widehat{f}^t_h = \widehat{f}^{t,\mathrm{or}}_{h,\mathrm{CO}}(\widehat{W}^t_{h+1})$. Equation (H.4) gives $\|\widehat{f}^t_h\|_{\mathbb{H}} \le R^{\mathrm{base}}_{T,\mathrm{CO}}(B^{\mathrm{CO}}_T)$. The realized posterior-standard-deviation shape belongs to $\mathfrak{F}^{\mathrm{post}}_{\lambda,T}$ by (C.11). The optimistic update (C.10) therefore places $\widehat{U}^t_h$ in the first class of (H.5) by the definition (C.12). The CO version space (3.1) belongs to the deterministic family $\mathfrak{A}_{\mathrm{alg,CO}}$ in (C.13). Applying the joint planning rule (3.2) in Algorithm 2 places $\widehat{W}^t_h$ in the second class of (H.5) by the definition (C.14). This closes the backward induction and proves the CO containment for every (t,h) on $\mathcal{E}_{\mathrm{or,CO}}$. □

H.2 Proof of Proposition H.2

A sufficient condition for Assumption 2.5 is the pair of RKHS norm bounds

$$\|u_h\|_{\mathbb{H}} \le 1, \qquad \|\omega^*_h\|_{\mathbb{H},\mathrm{TV}} \le \sqrt{d_L}, \tag{H.9}$$

which yields $R_U = 1 + \sqrt{d_L}$ (independent of $|\mathcal{S}|$ and of the inner-sup domain Θ').

Proposition H.2 (Linear MFG as RKHS special case)**.** *Any linear MFG (Definition 2.6) satisfying (H.9) fulfils Assumption 2.5 with $R_U = 1 + \sqrt{d_L}$, uniformly over all nonempty compact feasible sets $\Theta' \subseteq \Theta$ for which $W_{\max}^{U,\Theta'}$ is Borel measurable.*

Proof of Proposition H.2. Throughout this proof, d_L denotes the leader-side feature dimension appearing in the linear transition factorization (Definition 2.6, where ω_h^* takes values in $\mathbb{R}^{d_L}$); it is distinct from the follower-side parameter dimension d_F used in Assumption 2.1. The $\sqrt{d_L}$ factor in R_U arises from $\|\omega_h^*\|_{\mathbb{H},\mathrm{TV}} \le \sqrt{d_L}$ via the integrated TV-norm closure condition (H.9).

Fix any nonempty compact feasible set $\Theta' \subseteq \Theta$ and any Borel measurable $U : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{X} \times \mathcal{B}) \to [0, H]$ for which $W := W_{\max}^{U,\Theta'}$ (as in (2.10)) is Borel measurable, as required in Assumption 2.5. We verify that $\mathbb{B}_h W \in \mathcal{U}^*$ with RKHS-norm bound $R_U H$ where $R_U = 1 + \sqrt{d_L}$, uniformly over the choice of Θ'. The proof uses only two properties of W: (i) boundedness $\|W\|_\infty \le H$, which is checked directly below, and (ii) Borel measurability of $(s, \mu) \mapsto W(s, \mu)$, which is part of the quantified domain in Assumption 2.5 and is available for the algorithmic instances by the planning requirement in Assumption C.14, with sufficient topological conditions given in Remark C.17.

Boundedness of W and the measurability input. For any such compact $\Theta' \subseteq \Theta$, $\|W_{\max}^{U,\Theta'}\|_\infty \le H$ pointwise, since $W_{\max}^{U,\Theta'}(s, \mu) = \sup_{p \in \Pi_{\mathrm{loc}}, \theta \in \Theta'} \langle U(s, \cdot, \mu \otimes \nu^{p,\theta}), p \rangle_{\mathcal{A}} \le \sup_p \sup_a |U| \le \|U\|_\infty \le H$, independently of $\Theta' \subseteq \Theta$. As stated above, Borel measurability is part of the structural domain in Assumption 2.5; in the algorithmic applications, it is supplied by Assumption C.14. The Carathéodory and compactness conditions in Remark C.17 provide sufficient conditions for this conclusion.

Bellman operator decomposition. Writing $z = (s, a, \mathcal{P})$, by the definition of the Bellman operator $\mathbb{B}_h$ and the linear factorization in Definition 2.6,

$$\begin{aligned} \mathbb{B}_h W(s, a, \mathcal{P}) &= u_h(s, a, \mathcal{P}) + (P_h W)(s, a, \mathcal{P}) \\ &= \langle u_h(\cdot), \phi(z) \rangle_{\mathbb{H}} + \sum_{s' \in \mathcal{S}} \int_{\mu' \in \Delta(\mathcal{X})} \langle \phi(z), \omega_h^*(s', \mathrm{d}\mu') \rangle_{\mathbb{H}} \, W(s', \mu') \\ &= \Big\langle u_h(\cdot) + \underbrace{\textstyle\sum_{s' \in \mathcal{S}} \int W(s', \mu') \, \omega_h^*(s', \mathrm{d}\mu')}_{=: \, g_W^h \, \in \mathbb{H}}, \, \phi(z) \Big\rangle_{\mathbb{H}}, \end{aligned}$$

where the bracket-extraction uses bilinearity of the RKHS inner product in the first argument and the integrability of ω_h^* as an $\mathbb{H}$-valued signed measure.

RKHS-norm bound for g_W^h via the integrated TV-norm. If $\|W\|_\infty = 0$, then $g_W^h = 0$, and the required bound is immediate. Otherwise, apply the definition of $\|\omega_h^*\|_{\mathbb{H},\mathrm{TV}}$ (C.7) to the rescaled test function $f = W / \|W\|_\infty$, which satisfies $\|f\|_\infty \le 1$, to obtain

$$\begin{aligned} \|g_W^h\|_{\mathbb{H}} &= \left\| \sum_{s' \in \mathcal{S}} \int W(s', \mu') \, \omega_h^*(s', \mathrm{d}\mu') \right\|_{\mathbb{H}} = \|W\|_\infty \cdot \left\| \sum_{s' \in \mathcal{S}} \int f(s', \mu') \, \omega_h^*(s', \mathrm{d}\mu') \right\|_{\mathbb{H}} \\ &\le \|W\|_\infty \cdot \|\omega_h^*\|_{\mathbb{H},\mathrm{TV}} \le H \sqrt{d_L}. \end{aligned}$$

Conclusion. Combining the two parts,

$$\|\mathbb{B}_h W\|_{\mathbb{H}} \le \|u_h(\cdot)\|_{\mathbb{H}} + \|g_W^h\|_{\mathbb{H}} \le 1 + H\sqrt{d_L} \le (1 + \sqrt{d_L}) \, H =: R_U \, H,$$

where $R_U := 1 + \sqrt{d_L}$ is independent of $|\mathcal{S}|$. Therefore, $\mathbb{B}_h W_{\max}^{U,\Theta'} \in \mathcal{U}^*$, verifying Assumption 2.5 for this compact feasible set Θ'.

Comparison with the per-s' formulation. For finite $\mathcal{S}$, a precise per-s' condition is

$$\sup_{\substack{f \text{ Borel} \\ \|f\|_\infty \le 1}} \left\| \int_{\Delta(\mathcal{X})} f(\mu') \, \omega_h^*(s', \mathrm{d}\mu') \right\|_{\mathbb{H}} \le 1 \qquad \text{for each } s' \in \mathcal{S}.$$

For $\|W\|_\infty \le H$, the triangle inequality then gives

$$\|g_W^h\|_{\mathbb{H}} \le \sum_{s' \in \mathcal{S}} \left\| \int W(s', \mu') \, \omega_h^*(s', \mathrm{d}\mu') \right\|_{\mathbb{H}} \le H|\mathcal{S}|,$$

and hence the sufficient closure radius $R_U = 1 + |\mathcal{S}|$. The integrated condition (H.9) controls the combined operator directly and yields $R_U = 1 + \sqrt{d_L}$, thereby avoiding the explicit $|\mathcal{S}|$-dependence introduced by the per-state triangle bound, and matching the standard linear-MDP dimension scaling (Jin et al., 2020).

Measurability and well-definedness of the vector-measure integral. In the finite-rank coefficient representation used in this proposition, ω_h^* is a countably additive $\mathbb{R}^{d_L}$-valued vector measure, identified with the corresponding finite-dimensional subspace of $\mathbb{H}$. The map $(s', \mu') \mapsto W(s', \mu')$ is bounded and Borel measurable by the preceding paragraphs. Each coordinate of ω_h^* is a finite signed measure, so $\sum_{s'} \int W(s', \mu')\, \omega_h^*(s', \mathrm{d}\mu')$ is well defined coordinatewise as the standard integral of a bounded scalar function with respect to a vector measure (Diestel & Uhl, 1977). The operator-norm bound $\left\| \int W \, \mathrm{d}\omega_h^* \right\|_{\mathbb{H}} \le \|W\|_\infty \, \|\omega_h^*\|_{\mathbb{H},\mathrm{TV}}$ follows directly from (C.7). □

H.3 An infinite-rank certificate for Bellman completeness

H.3.1 A nonlinear geometric-transition model

Example H.3 (A nonlinear model with an infinite-rank Bellman-complete kernel)**.** *Let $\mathcal{S} = \mathbb{N}_0$, $\mathcal{A} = \mathcal{B} = \{0, 1\}$, $\mathcal{X} = \{x_0\}$, and $\mu_0 = \delta_{x_0}$, with initial system state $(0, \mu_0)$ and $\Pi_{\mathrm{loc}} = \Delta(\mathcal{A})$. Set $\varphi(s, a, x_0, b) = ab$ and $\theta_h^* = \theta^* \in (0, \min\{1, B_\Theta\}]$ for every h, where $B_\Theta > 0$. For $z = (s, a, \mathcal{P})$, write $\mathcal{P}_B(1) = \mathcal{P}(\{(x_0, 1)\})$ and $\rho(z) = (a + \mathcal{P}_B(1))/4$. Define, on the entire input domain,*

$$u_h(z) = \rho(z)(1 - \rho(z)), \qquad \mathbb{P}_h(\{(n, \mu_0)\} \mid z) = (1 - \rho(z))\rho(z)^n, \qquad n \in \mathbb{N}_0, \tag{H.10}$$

$$K(z, z') = \frac{(1 - \rho(z))(1 - \rho(z'))}{1 - 2\rho(z)\rho(z')}. \tag{H.11}$$

Here, the probability at $n = 0$ is $1 - \rho(z)$, including when $\rho(z) = 0$. The follower share under commitment p is $\nu^{p,\theta^*}(1 \mid x_0; s) = (1 + e^{-\eta\theta^* p(1)})^{-1}$, so the leader's reward function and the system transition kernel depend nonlinearly on the endogenous follower response.

The kernel is normalized and has infinite rank, and the transition kernel admits no exact finite-dimensional linear factorization even on attainable inputs. Nevertheless, its RKHS satisfies

$$\|\mathbb{B}_h W\|_{\mathbb{H}} \le \frac{1}{\sqrt{2}} + \sqrt{2}H \le \frac{3}{\sqrt{2}}H, \quad \text{for every bounded Borel } W : \mathcal{S} \times \Delta(\mathcal{X}) \to [0, H]. \tag{H.12}$$

Consequently, Assumption 2.5 holds with $R_U = 3/\sqrt{2}$ uniformly over its entire domain of compact sets Θ'. The certificate and verification below establish the global kernel, regularity, planning, and complexity properties.

The following sufficient condition controls every bounded Borel continuation value, which makes the uniformity over follower-parameter feasible sets immediate. It concerns the reward–transition pair; compatibility with the deterministic population factorization is verified separately for Example H.3.

Lemma H.4 (Countable-mixture certificate)**.** *Let $w_j : \mathcal{Z}_{lf} \to [0, 1]$, $j \in \mathbb{N}_0$, be Borel functions with $\sum_{j \ge 0} w_j(z) = 1$. Suppose there are numbers $c_j > 0$ such that*

$$C_c^2 := \sum_{j \ge 0} c_j^2 < \infty, \qquad \sup_z \sum_{j \ge 0} \frac{w_j(z)^2}{c_j^2} \le 1.$$

Set $\psi(z) = (w_j(z)/c_j)_{j \ge 0} \in \ell_2$ and $K(z, z') = \langle \psi(z), \psi(z') \rangle_{\ell_2}$, and let $\mathbb{H}$ be the RKHS of K. For each h and j, let $Q_{h,j}$ be a Borel probability measure on $\mathcal{S} \times \Delta(\mathcal{X})$ and suppose

$$\mathbb{P}_h(E \mid z) = \sum_{j \ge 0} w_j(z) Q_{h,j}(E), \qquad u_h \in \mathbb{H}, \quad \|u_h\|_{\mathbb{H}} \le R_r, \tag{H.13}$$

for every Borel set E, with R_r independent of h. Then $K(z, z) \le 1$ and every bounded Borel $W : \mathcal{S} \times \Delta(\mathcal{X}) \to [0, H]$ satisfies

$$\|P_h W\|_{\mathbb{H}} \le C_c H, \qquad \|\mathbb{B}_h W\|_{\mathbb{H}} \le R_r + C_c H \le (R_r + C_c)H. \tag{H.14}$$

In particular, Assumption 2.5 holds with $R_U = R_r + C_c$ whenever these primitives form an admissible MFG.

Proof of Lemma H.4. The feature map ψ is Borel as an ℓ_2-valued map, because its finite-coordinate truncations are Borel and converge pointwise in ℓ_2. Its norm is at most one, so K is a Borel positive semidefinite kernel with $K(z,z) \le 1$. The nonnegative mixture in (H.13) defines a Borel probability kernel: countable additivity follows from Tonelli's theorem, and its mass on the whole next-state space is $\sum_j w_j(z) = 1$.

For a bounded Borel W, set

$$a_{h,j}(W) = c_j \int W \, \mathrm{d}Q_{h,j}.$$

Then $a_h(W) \in \ell_2$ and $\|a_h(W)\|_{\ell_2}^2 \le H^2 \sum_j c_j^2 = C_c^2 H^2$. Tonelli's theorem gives

$$(P_h W)(z) = \sum_{j \ge 0} w_j(z) \int W \, \mathrm{d}Q_{h,j} = \langle \psi(z), a_h(W) \rangle_{\ell_2}.$$

For any coefficient vector $a \in \ell_2$, the function $z \mapsto \langle \psi(z), a \rangle$ belongs to $\mathbb{H}$ and has RKHS norm at most $\|a\|_{\ell_2}$. Indeed, projecting a onto the closed span of $\{\psi(z) : z \in \mathcal{Z}_{lf}\}$ leaves this function unchanged and identifies that span isometrically with the canonical RKHS. This proves the first inequality in (H.14); the remaining inequalities follow from the triangle inequality and $H \ge 1$.

For every pair (U, Θ') in the domain of Assumption 2.5, the envelope $W_{\max}^{U,\Theta'}$ is Borel by the stated domain restriction and lies in $[0, H]$ because p is a probability distribution. Applying the same bound to this envelope gives a constant independent of U and Θ'. No attainment assertion for arbitrary bounded Borel U is needed. □

Proof of Example H.3. We verify the model on all inputs and distinguish the coordinate feature map ψ from the canonical RKHS feature map $\phi(z) = K(z, \cdot)$ used in the main text.

Model and normalized kernel. For every $\mathcal{P} \in \Delta(\mathcal{X} \times \mathcal{B})$ and $a \in \{0,1\}$, one has $0 \le \rho(z) \le 1/2$. Thus $w_n(z) = (1-\rho(z))\rho(z)^n$ are nonnegative Borel functions summing to one, and (H.10) defines a probability kernel on the entire input domain. The population evolution map is $\Phi_h \equiv \mu_0$, and the leader-state kernel is the displayed geometric distribution, verifying Assumption 2.2. The true follower rewards are $r_h(x_0, b; s, a) = \theta^* ab \in [0,1]$, while $0 \le u_h \le 1/4$. The scalar follower feature is bounded by one, verifying Assumption 2.1. The softmax formula (C.2) yields the follower share stated in the example. Taking

$$Q_{h,n} = \delta_{(n,\mu_0)}, \qquad c_n = 2^{-n/2}, \qquad \psi_n(z) = 2^{n/2}(1-\rho(z))\rho(z)^n$$

gives $\sum_n c_n^2 = 2$ and

$$\|\psi(z)\|_{\ell_2}^2 = \frac{(1-\rho(z))^2}{1-2\rho(z)^2} \le 1.$$

The last inequality follows from $2\rho - 3\rho^2 \ge 0$ on $[0, 1/2]$. Summing the geometric series for $\langle \psi(z), \psi(z') \rangle$ gives (H.11). Also $u_h = 2^{-1/2}\psi_1$, so $\|u_h\|_{\mathbb{H}} \le 2^{-1/2}$. The mixture certificate with $C_c = \sqrt{2}$ proves (H.12) for every bounded Borel W, including all the envelopes required by the current closure assumption.

Infinite rank of both the kernel and the transition. Fix $s = 0$ and $a = 0$ and vary $p(1)$ in any nondegenerate subinterval of $(0,1)$. Since $\eta\theta^* > 0$, the corresponding follower shares, and hence the values of ρ, vary strictly monotonically. These are attainable inputs: the initial state is 0 and action 0 has positive probability under each such commitment. For any integer $M \ge 1$, choose M distinct resulting values $\rho_1, \dots, \rho_M$, and denote the corresponding inputs by $z_1, \dots, z_M$. The matrix of transition probabilities to states $0, \dots, M-1$ is

$$A_{i,n+1} = (1-\rho_i)\rho_i^n, \qquad \det A = \prod_{i=1}^{M}(1-\rho_i) \prod_{1 \le i < j \le M} (\rho_j - \rho_i) \ne 0.$$

Any exact D-dimensional transition factorization would force this matrix to have rank at most D for every M, contradicting the choice $M > D$. Moreover, the first M coordinate features

form the invertible matrix $A\,\mathrm{diag}(2^{n/2})_{n=0}^{M-1}$. Consequently, the M full feature vectors are linearly independent and their kernel Gram matrix is positive definite. As M is arbitrary, the kernel has infinite rank. This verifies more than adding redundant coordinates to an otherwise finite-dimensional linear transition model.

Regularity, sampling, and planning. Regarding ψ as a function of $\rho \in [0, 1/2]$, its coordinate derivatives are

$$\psi_0'(\rho) = -1, \qquad \psi_n'(\rho) = 2^{n/2}\big[n(1-\rho)\rho^{n-1} - \rho^n\big], \quad n \geq 1.$$

The derivative series is uniformly square summable on this interval, since its terms admit a bound proportional to $(n+1)^2 2^{-n}$. Summing it gives

$$\|\psi'(\rho)\|_{\ell_2}^2 = \frac{3 - 8\rho + 6\rho^2}{(1-2\rho^2)^3} \leq 24.$$

The formula at $\rho = 0$ follows by continuity. Since $|\rho(s,a,\mathcal{P}) - \rho(s,a,\mathcal{P}')| \leq \|\mathcal{P} - \mathcal{P}'\|_1/4$, the coordinate feature map is Lipschitz in the joint distribution with constant $\sqrt{6}/2$. The kernel identity gives

$$\|\phi(z) - \phi(z')\|_{\mathbb{H}}^2 = K(z,z) + K(z',z') - 2K(z,z') = \|\psi(z) - \psi(z')\|_{\ell_2}^2,$$

so Assumption 2.4 holds with the same constant for the canonical feature map. For $n \geq 1$, $w_n(\rho)$ is nondecreasing on $[0, 1/2]$, whereas $w_0(\rho) = 1 - \rho$ is decreasing. It follows that

$$\|w(\rho) - w(\rho')\|_1 = 2|\rho - \rho'|.$$

In particular, the leader transition has joint-distribution Lipschitz constant at most $1/2$ and the population evolution map has Lipschitz constant zero. These also give admissible constants $L_{\mathcal{T}} = 1/2$ and $L_\Phi = 0$ in Assumption C.43. Under FSO, the state batch consists of N copies of x_0 and the follower actions may be sampled conditionally independently from the stated response law, consistently with Assumption 2.3.

For algorithmic planning, every RKHS function is continuous in the joint-distribution input, and so is each posterior-standard-deviation bonus, because its defining operator is bounded and ϕ is norm-continuous. Truncation and adding a scalar offset preserve this continuity. Thus the planning objective for any of the algorithmic UCB functions is Borel in the state and continuous in (p, θ). The finite action simplex and any nonempty compact $\Theta' \subseteq \Theta$ give a nonempty compact argmax and a measurable selector as in Assumption C.14. The same argument with the data as a measurable parameter applies to the algorithmic compact feasible-set correspondences. The true finite-horizon CO benchmark is attained as well: backward induction uses bounded continuation values, the ℓ_1 continuity of w, and compactness of the action simplex to construct a Borel maximizing policy. These attainment statements concern the continuous planning objectives just described, and do not assert attainment for every bounded Borel U in the structural assumption.

Information gain and the complete value-envelope class. Truncating the coordinate features after $n = m$ gives, uniformly in z,

$$\sum_{n>m} \psi_n(z)^2 = \frac{(1-\rho(z))^2 (2\rho(z)^2)^{m+1}}{1 - 2\rho(z)^2} \leq 2^{-m}.$$

For any $q \leq T$ query points, write their Gram matrix as $K_{1:q} = K_{1:q}^{(m)} + E_m$. Both terms are positive semidefinite, $\mathrm{rank}(K_{1:q}^{(m)}) \leq m+1$, $\mathrm{tr}(K_{1:q}^{(m)}) \leq T$, and $\mathrm{tr}(E_m) \leq T 2^{-m}$. The log-determinant identity followed by $\log\det(I + C) \leq \mathrm{tr}(C)$ for positive semidefinite C yields

$$\log\det(I + (K_{1:q}^{(m)} + E_m)/\lambda) \leq \log\det(I + K_{1:q}^{(m)}/\lambda) + \mathrm{tr}(E_m)/\lambda.$$

Concavity of the logarithm bounds the first term using its rank and trace. Taking the supremum in (C.6) therefore gives

$$G_K(T,\lambda) \leq \frac{m+1}{2} \log\left(1 + \frac{T}{(m+1)\lambda}\right) + \frac{T 2^{-m}}{2\lambda}. \tag{H.15}$$

Taking $m = \lceil \log_2(1 + T/\lambda) \rceil$ gives $G_K(T, \lambda) = O(\log^2(2 + T))$ for $\lambda = \Theta(1)$.

Information gain alone is not the full complexity requirement in the current theorem, so we also verify its value-envelope covering condition. In this example the kernel and follower feature are independent of s, and $\Delta(\mathcal{X}) = \{\mu_0\}$. Every RKHS function and every posterior-standard-deviation shape is consequently independent of s. For each fixed UCB function and each fixed feasible set, the value envelope in (C.14) is therefore a constant function of (s, μ_0) with value in $[0, H]$. Thus, for both settings and every choice of deterministic radii,

$$N_{\infty,\iota}(\varepsilon; h, R, B, C_{\text{off}}) \le 1 + \lceil H/\varepsilon \rceil, \qquad \iota \in \{\text{CO}, \text{FSO}\}, \qquad \varepsilon > 0.$$

This bound is uniform over the entire algorithmic feasible-set family, not only a realized feasible set. Together with (H.15), it makes the setting-specific radius loads bounded as functions of the candidate radius for fixed T and the other inputs, and hence the deterministic dyadic calibrations are finite. The example provides an explicit infinite-rank structural witness with controlled statistical complexity. □

H.4 Proof of Lemma H.5

We first state the lemma for the reader's convenience.

Lemma H.5 (Regret decomposition)**.** *Under the assumptions of Theorem C.10, with probability* $\ge 1 - \zeta$,

$$\begin{aligned}\mathrm{Reg}_{\mathrm{CO}}(T) \le & \underbrace{\sum_{t,h} \mathbb{E}^t\big[(\widehat{U}_h^t - u_h)(s_h, a_h, \mathcal{P}_h) - \widehat{W}_{h+1}^t(s_{h+1}, \mu_{h+1})\big]}_{\textit{LBE}} \\ & + L_U \underbrace{\sum_{t,h} \mathbb{E}^t\Big[\textstyle\sum_x \mu_h(x) \Big\| \widehat{\nu}_h^t(\cdot \mid x; s_h) - \nu_h^{\pi^t}(\cdot \mid x; s_h, \mu_h) \Big\|_1\Big]}_{\textit{JDE}},\end{aligned} \tag{H.16}$$

where, writing $p_h^t = \pi_h^t(\cdot|s_h, \mu_h)$ *for the realized local policy executed at episode* t, $\widehat{\nu}_h^t(\cdot|x; s_h) := \nu^{p_h^t, \theta_h^t}(\cdot|x; s_h)$ *is the algorithm's predicted follower response under the planner's joint argmax pair* (p_h^t, θ_h^t) *from (3.2), and* $\nu_h^{\pi^t}(\cdot \mid x; s_h, \mu_h) = \nu^{p_h^t, \theta_h^*}(\cdot \mid x; s_h)$ *is the* true *follower response generated under the* same executed local policy p_h^t *but the true parameter* θ_h^**. Both responses are evaluated along the trajectory induced by* π^t*: this is precisely the quantal-response-error object produced by applying the performance-difference Lemma E.2 to the executed policy* π^t *in Step 3 below (cf. the* $\widehat{\nu}_h - \nu_h^\pi$ *term of (E.3) with* $\pi = \pi^t$*).*

Proof of Lemma H.5. **Outline.** The proof proceeds in three steps. *Step 1* introduces two operators that compactly represent the leader's value-function update under different choices of follower-response model. *Step 2* decomposes the per-episode suboptimality into an optimism term (i) and a performance-difference term (ii); we show that, on the high-probability uncertainty-quantification event of Lemma H.1, term (i) is nonpositive. *Step 3* applies the performance-difference Lemma E.2 to term (ii), yielding the leader's-Bellman-error / follower-response-error decomposition stated in Lemma H.5.

Step 1: Two value-iteration operators. For any leader-policy π, parameter θ, and follower-response policy ν, define

$$\begin{aligned}\big(T_h^{\pi,\theta} U_h\big)(s, \mu) &:= \langle U_h(s, \cdot, \mu \otimes \nu_h^{\pi,\theta}(s, \mu)), \pi_h(\cdot \mid s, \mu)\rangle_{\mathcal{A}}, \\ \big(T_h^{\pi,\nu} U_h\big)(s, \mu) &:= \langle U_h(s, \cdot, \mu \otimes \nu_h(s, \mu)), \pi_h(\cdot \mid s, \mu)\rangle_{\mathcal{A}}.\end{aligned}$$

The two operators differ only in how the follower-response policy is supplied: $T_h^{\pi,\theta}$ generates it from a follower-reward parameter θ via the quantal-response map $\theta \mapsto \nu^{\pi,\theta}$, whereas $T_h^{\pi,\nu}$ takes the response policy ν as a primitive. Under Scheme 1 (3.2), the leader's optimistic value function admits the compact representation

$$\widehat{W}_h^t(s, \mu) \;=\; \max_{p \in \Pi_{\mathrm{loc}},\, \theta \in \Theta_h^{t-1}} \big(T_h^{p,\theta} \widehat{U}_h^t\big)(s, \mu). \tag{H.17}$$

Step 2: Suboptimality decomposition and nonpositivity of term (i). By the definition of the CO regret in (C.4),

$$\mathrm{SubOpt}(\pi^t) \;=\; J(\pi^*) - J(\pi^t) \;=\; \underbrace{J(\pi^*) - \mathbb{E}\left[\widehat{W}_1^t(s_1,\mu_1)\right]}_{\text{(i)}} \;+\; \underbrace{\mathbb{E}\left[\widehat{W}_1^t(s_1,\mu_1)\right] - J(\pi^t)}_{\text{(ii)}},$$

where the expectation is taken over the initial system state $(s_1,\mu_1)\sim\rho_0$ with the standing degenerate initial distribution $\rho_0=\delta_{(s_0,\mu_0)}$.

We first show that term (i) is nonpositive on the leader-side uncertainty-quantification event of Lemma H.1 (probability at least $1-\zeta$). By the definition (C.3) of J and by adding and subtracting the operator-induced terms,

$$\begin{aligned}
J(\pi^*) &= \mathbb{E}^{\pi^*}\Big[\sum_{h\in[H]} u_h(s_h,a_h,\mu_h\otimes\nu_h^{\pi^*}(s_h,\mu_h))\Big] \\
&= \mathbb{E}^{\pi^*}\sum_{h\in[H]}\Big[u_h(s_h,a_h,\mu_h\otimes\nu_h^{\pi^*}(s_h,\mu_h)) + \big(T_h^{\pi^*,\nu^{\pi^*}}\widehat{U}_h^t\big)(s_h,\mu_h) - \widehat{U}_h^t(s_h,a_h,\mu_h\otimes\nu_h^{\pi^*}(s_h,\mu_h))\Big] \\
&= \sum_{h\in[H]}\Big\{\mathbb{E}^{\pi^*}\Big[\big(T_h^{\pi^*,\nu^{\pi^*}}\widehat{U}_h^t\big)(s_h,\mu_h)\Big] - \mathbb{E}^{\pi^*}\Big[\big(\widehat{U}_h^t - u_h\big)(s_h,a_h,\mu_h\otimes\nu_h^{\pi^*}(s_h,\mu_h))\Big]\Big\},
\end{aligned} \tag{H.18}$$

where the second equality merely adds and subtracts $\widehat{U}_h^t$, and the last equality uses the fact that the operator-induced terms cancel:

$$\begin{aligned}
\mathbb{E}^{\pi^*}\sum_{h\in[H]}\big(T_h^{\pi^*,\nu^{\pi^*}}\widehat{U}_h^t\big)(s_h,\mu_h) &= \mathbb{E}^{\pi^*}\sum_{h\in[H]}\big\langle \widehat{U}_h^t(s_h,\cdot,\mu_h\otimes\nu_h^{\pi^*}(s_h,\mu_h)),\, \pi_h^*(\cdot\,|\,s_h,\mu_h)\big\rangle_{\mathcal{A}} \\
&= \mathbb{E}^{\pi^*}\sum_{h\in[H]}\widehat{U}_h^t(s_h,a_h,\mu_h\otimes\nu_h^{\pi^*}(s_h,\mu_h)),
\end{aligned}$$

where the second equality uses $a_h\sim\pi_h^*(\cdot\,|\,s_h,\mu_h)$ to identify the inner product over $\mathcal{A}$ with the corresponding expectation over a_h.

Next, by a standard telescoping argument applied to the optimistic value sequence,

$$\begin{aligned}
\mathbb{E}\left[\widehat{W}_1^t(s_1,\mu_1)\right] &= \sum_{h=1}^{H}\mathbb{E}\left[\widehat{W}_h^t(s_h,\mu_h) - \widehat{W}_{h+1}^t(s_{h+1},\mu_{h+1})\right] + \mathbb{E}\left[\widehat{W}_{H+1}^t(s_{H+1},\mu_{H+1})\right] \\
&= \sum_{h=1}^{H}\mathbb{E}\left[\widehat{W}_h^t(s_h,\mu_h) - \widehat{W}_{h+1}^t(s_{h+1},\mu_{h+1})\right],
\end{aligned} \tag{H.19}$$

where the boundary term vanishes by the convention $\widehat{W}_{H+1}^t\equiv 0$.

Combining (H.18) and (H.19), and inserting and subtracting the operator $T_h^{\pi^*,\check{\nu}^t}\widehat{U}_h^t$, where for each realized (s_h,μ_h) we fix $p_h^*=\pi_h^*(\cdot|s_h,\mu_h)$ and choose an optimistic follower parameter

$$\check{\theta}_h^t \in \operatorname*{argmax}_{\theta\in\Theta_h^{t-1}}\big\langle \widehat{U}_h^t(s_h,\cdot,\mu_h\otimes\nu^{p_h^*,\theta}(\cdot|\cdot;s_h)),\, p_h^*\big\rangle_{\mathcal{A}},$$

with $\check{\nu}_h^t := \nu^{p_h^*, \check{\theta}_h^t}(\cdot|\cdot; s_h)$, we obtain

$$
\begin{aligned}
\text{(i)} &= J(\pi^*) - \mathbb{E}\left[\widehat{W}_1^t(s_1, \mu_1)\right] \\
&= \sum_{h\in[H]} \left\{ \mathbb{E}^{\pi^*}\left[\left(T_h^{\pi^*, \nu^{\pi^*}} \widehat{U}_h^t\right)(s_h, \mu_h)\right] - \mathbb{E}^{\pi^*}\left[\left(\widehat{U}_h^t - u_h\right)(s_h, a_h, \mu_h \otimes \nu_h^{\pi^*}(s_h, \mu_h))\right]\right\} \\
&\quad - \sum_{h=1}^{H} \mathbb{E}\left[\widehat{W}_h^t(s_h, \mu_h) - \widehat{W}_{h+1}^t(s_{h+1}, \mu_{h+1})\right] \\
&= \underbrace{\sum_{h=1}^{H} \mathbb{E}\left[-\left(\widehat{U}_h^t - u_h\right)(s_h, a_h, \mu_h \otimes \nu_h^{\pi^*}(s_h, \mu_h)) + \widehat{W}_{h+1}^t(s_{h+1}, \mu_{h+1})\right]}_{\text{(i.a)}} \\
&\quad + \underbrace{\sum_{h=1}^{H} \mathbb{E}\left[-\widehat{W}_h^t(s_h, \mu_h) + \left(T_h^{\pi^*, \check{\nu}^t} \widehat{U}_h^t\right)(s_h, \mu_h)\right]}_{\text{(i.b)}} \\
&\quad + \underbrace{\sum_{h=1}^{H} \mathbb{E}\left[\left(T_h^{\pi^*, \nu^{\pi^*}} - T_h^{\pi^*, \check{\nu}^t}\right) \widehat{U}_h^t(s_h, \mu_h)\right]}_{\text{(i.c)}},
\end{aligned}
$$

where the expectation is taken over the trajectory induced by π^* under the true model.

We now show that each of the three terms (i.a), (i.b), (i.c) is nonpositive. *For term (i.a):* By the uncertainty-quantification Lemma H.1, with probability at least $1 - \zeta$,

$$
\left| \langle \phi(z), \widehat{f}_h^t \rangle_{\mathbb{H}} - u_h(z) - (P_h \widehat{W}_{h+1}^t)(z) \right| \;\le\; \Gamma_h^t(z), \quad \forall\, z \in \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}),\; h \in [H],\; t \in [T]. \tag{H.20}
$$

The lower side of (H.20) gives

$$
\langle \phi(z), \widehat{f}_h^t \rangle_{\mathbb{H}} + \Gamma_h^t(z) \ge u_h(z) + (P_h \widehat{W}_{h+1}^t)(z).
$$

The reward bound and $\widehat{W}_{h+1}^t \in [0, H-h]$ imply $u_h(z) + (P_h \widehat{W}_{h+1}^t)(z) \in [0, H-h+1]$. Since $\mathrm{Trun}_{[0,H-h+1]}$ is nondecreasing and fixes every point of this interval, (C.10) yields

$$
\widehat{U}_h^t(z) \ge u_h(z) + (P_h \widehat{W}_{h+1}^t)(z).
$$

At the query $z = (s_h, a_h, \mathcal{P}_h)$ in term (i.a), the transition identity gives $\mathbb{E}[\widehat{W}_{h+1}^t(s_{h+1}, \mu_{h+1}) \mid s_h, a_h, \mathcal{P}_h] = (P_h \widehat{W}_{h+1}^t)(s_h, a_h, \mathcal{P}_h)$. Consequently, the integrand of term (i.a) has nonpositive conditional expectation, so term (i.a) is nonpositive on the same event.

For term (i.b): Since $\theta_h^* \in \Theta_h^{t-1}$ by the definition of the step-h version space (3.1), the joint maximization (H.17) dominates any single-policy joint, so

$$
\begin{aligned}
\widehat{W}_h^t(s_h, \mu_h) &= \max_{p \in \Pi_{\mathrm{loc}},\, \theta \in \Theta_h^{t-1}} \left(T_h^{p,\theta} \widehat{U}_h^t\right)(s_h, \mu_h) \\
&\ge \max_{\theta \in \Theta_h^{t-1}} \left(T_h^{\pi^*, \theta} \widehat{U}_h^t\right)(s_h, \mu_h) = \left(T_h^{\pi^*, \check{\nu}^t} \widehat{U}_h^t\right)(s_h, \mu_h),
\end{aligned}
$$

where the equality uses the definition of $\check{\nu}^t$ as the maximizer; hence (i.b) is nonpositive.

For term (i.c): Since $\theta_h^* \in \Theta_h^{t-1}$, the parameter associated with $\nu_h^{\pi^*}$ (namely θ_h^*) is feasible in the maximization defining $\check{\nu}^t$, while $\check{\nu}^t$ achieves the maximum; therefore

$$
\left(T_h^{\pi^*, \nu^{\pi^*}} \widehat{U}_h^t\right)(s_h, \mu_h) \;\le\; \left(T_h^{\pi^*, \check{\nu}^t} \widehat{U}_h^t\right)(s_h, \mu_h),
$$

so term (i.c) is also nonpositive.

Combining (i.a)–(i.c), we conclude that on the UQ event of Lemma H.1,

$$
\mathrm{SubOpt}(\pi^t) \;\le\; \mathbb{E}\left[\widehat{W}_1^t(s_1, \mu_1)\right] - J(\pi^t), \tag{H.21}
$$

which holds with probability at least $1 - \zeta$.

Step 3: Performance-difference decomposition. Summing (H.21) over $t \in [T]$ and applying Lemma E.2 (performance-difference lemma), equation (E.3), to the tuple $(\pi^t, \widehat{W}^t, \widehat{U}^t, \widehat{\nu}^t)$ gives the following bound. Here, $\widehat{\nu}^t$ denotes the optimizer of Scheme 1 under π^t. The Lipschitz precondition (E.1) holds with $L_U = L_U^{\mathrm{CO}}$ by (H.5) and Corollary F.1. Thus, on the oracle-UQ event,

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{CO}}(T) = \sum_{t=1}^{T} \mathrm{SubOpt}(\pi^t) \;&\le\; \sum_{t=1}^{T} \Big(\mathbb{E}\left[\widehat{W}_1^t(s_1,\mu_1)\right] - J(\pi^t) \Big) \\
&\le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[(\widehat{U}_h^t - u_h)(s_h,a_h,\mathcal{P}_h) - \widehat{W}_{h+1}^t(s_{h+1},\mu_{h+1}) \right] \\
&\quad + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\widehat{W}_h^t(s_h,\mu_h) - \big(T_h^{\pi^t,\widehat{\nu}^t}\widehat{U}_h^t\big)(s_h,\mu_h) \right] \\
&\quad + L_U \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \mu_h(x) \left\| \widehat{\nu}_h^t(\cdot \mid x; s_h) - \nu_h^{\pi^t}(\cdot \mid x; s_h,\mu_h) \right\|_1 \right] \\
&= \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[(\widehat{U}_h^t - u_h)(s_h,a_h,\mathcal{P}_h) - \widehat{W}_{h+1}^t(s_{h+1},\mu_{h+1}) \right]}_{\text{leader's Bellman error}} \\
&\quad + \underbrace{L_U \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \mu_h(x) \left\| \widehat{\nu}_h^t(\cdot \mid x; s_h) - \nu_h^{\pi^t}(\cdot \mid x; s_h,\mu_h) \right\|_1 \right]}_{\text{follower-response error}}.
\end{aligned}$$

The penultimate equality uses the identity $\widehat{W}_h^t(s_h,\mu_h) = \big(T_h^{\pi^t,\widehat{\nu}^t}\widehat{U}_h^t\big)(s_h,\mu_h)$, which follows from Scheme 1 (3.2) together with the representation (H.17). Throughout this expression, $\widehat{\nu}_h^t$ denotes the response induced by the optimistic follower parameter under the executed local policy, and $\nu_h^{\pi^t}$ denotes the corresponding true response evaluated along the executed trajectory:

$$\widehat{\nu}_h^t(\cdot|x;s_h) \;:=\; \nu^{p_h^t,\theta_h^t}(\cdot|x;s_h), \qquad \nu_h^{\pi^t}(\cdot \mid x; s_h,\mu_h) \;=\; \nu^{p_h^t,\theta_h^*}(\cdot \mid x; s_h),$$

where $p_h^t = \pi_h^t(\cdot|s_h,\mu_h)$ is the executed local policy. Thus, the JDE term compares the predicted and true follower responses under the same executed local policy p_h^t, differing only in the follower parameter.

The JDE term compares the predicted response with $\nu_h^{\pi^t}$ under the same executed policy; ν^{π^*} denotes the QSE comparator response.

The conditional expectation $\mathbb{E}^t$ is taken over the trajectory $\{s_h,\mu_h\}_{h\in[H]}$ generated by the leader's policy π^t under the true model. This completes the proof of Lemma H.5. □

H.5 Proof of Lemma H.6

Lemma H.6 (CO leader Bellman error)**.** *On the intersection of $\mathcal{E}_{\mathrm{or,CO}}$ and the CO leader LBE-martingale event,*

$$\mathrm{LBE} \le 4B_T^{\mathrm{CO}} H\sqrt{T G_K(T,\lambda)} + 2B_T^{\mathrm{CO}} B_\phi H\sqrt{T}\log\left(\frac{2eH}{\zeta}\right). \tag{H.22}$$

The intersection has probability at least $1-2\zeta$.

Proof of Lemma H.6. For every episode t, the CO predictor equals its oracle counterpart because the CO design input is the true observed input. On the predictable historical event $\mathcal{E}_{\mathrm{or,CO}}^t$, Lemma H.1 gives

$$\left| \widehat{f}_h^t(z) - (\mathbb{B}_h\widehat{W}_{h+1}^t)(z) \right| \le \Gamma_{h,\mathrm{CO}}^t(z), \qquad \Gamma_{h,\mathrm{CO}}^t(z) := B_T^{\mathrm{CO}} \|\phi(z)\|_{(\widehat{\Lambda}_{L,h,\mathrm{CO}}^t)^{-1}}.$$

The truncation defining the optimistic action value and the Bellman identity therefore imply

$$\mathrm{LBE} \le 2 \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}^t \left[\Gamma^t_{h,\mathrm{CO}}(z^t_h) \right].$$

Since $\lambda \ge 1$, one has $0 \le \Gamma^t_{h,\mathrm{CO}}(z) \le B_T^{\mathrm{CO}} B_\phi$. Apply the nonnegative-process form of Lemma E.3 at failure probability ζ/H for every step and take a union bound. On the resulting LBE-martingale event,

$$\sum_{t=1}^{T} \mathbb{E}^t \left[\Gamma^t_{h,\mathrm{CO}}(z^t_h) \right] \le \sum_{t=1}^{T} \Gamma^t_{h,\mathrm{CO}}(z^t_h) + B_T^{\mathrm{CO}} B_\phi \sqrt{T} \log\left(\frac{2eH}{\zeta} \right)$$

simultaneously for all $h \in [H]$. By the Cauchy–Schwarz inequality, Lemma E.5, and the definition of G_K,

$$\sum_{t=1}^{T} \sum_{h=1}^{H} \Gamma^t_{h,\mathrm{CO}}(z^t_h) \le 2 B_T^{\mathrm{CO}} H \sqrt{T G_K(T, \lambda)}.$$

Combining the last three displays proves (H.22). The oracle event and the LBE-martingale event each fail with probability at most ζ. □

H.6 Softmax comparison tools for follower-response analysis

Setting up the comparison framework. For the follower reward model in Definition 2.6, the softmax quantal response admits the closed form

$$\nu^\theta(b \,|\, x; s, \mu, \pi) \;\propto\; \exp\big(\eta \cdot r_\theta(x, b, s, \mu, \pi)\big),$$

i.e., $\nu^\theta(b \,|\, x; s, \mu, \pi) = \exp\{\eta r_\theta(x, b, s, \mu, \pi)\} / \sum_{b' \in \mathcal{B}} \exp\{\eta r_\theta(x, b', s, \mu, \pi)\}$. For any two parameter values $\theta, \widetilde{\theta} \in \Theta$, we write $r, \widetilde{r}$ for the corresponding rewards and $\nu, \widetilde{\nu}$ for the corresponding responses. We establish the advantage bounds and Hellinger comparisons used in the follower-response analysis.

To this end, we re-express the followers' value functions in the softmax form:

$$\begin{aligned} Q_\theta(x, b, s, \mu, \pi) &= r_\theta(x, b, s, \mu, \pi), \\ V_\theta(x, s, \mu, \pi) &= \frac{1}{\eta} \log\Big(\sum_{b \in \mathcal{B}} \exp\big(\eta \cdot Q_\theta(x, b, s, \mu, \pi)\big) \Big), \\ A_\theta(x, b, s, \mu, \pi) &= Q_\theta(x, b, s, \mu, \pi) - V_\theta(x, s, \mu, \pi) \quad \text{(advantage)}. \end{aligned}$$

Notational simplification. For the softmax comparisons in this subsection, we fix a follower state x, a leader state s, a mean field μ, a commitment policy π, and a step $h \in [H]$, suppressing them whenever no ambiguity arises. We thus write $Q_\theta \equiv Q_\theta(x, \cdot, s, \mu, \pi) \in \mathbb{R}^{|\mathcal{B}|}$ (a vector indexed by b), $V_\theta \equiv V_\theta(x, s, \mu, \pi) \in \mathbb{R}$, and $A_\theta \equiv A_\theta(x, \cdot, s, \mu, \pi) \in \mathbb{R}^{|\mathcal{B}|}$, and analogously $\nu^\theta(b) \equiv \nu^\theta(b \,|\, x; s, \mu, \pi)$. With these abbreviations, the softmax response satisfies the identity

$$\nu^\theta(b) \;=\; \exp\big(\eta \cdot A_\theta(b)\big), \qquad \forall\, b \in \mathcal{B},$$

which follows directly from the definition of V_θ as a log-sum-exp and the relation $A_\theta = Q_\theta - V_\theta$. For a second parameter $\widetilde{\theta} \in \Theta$, we use tilded symbols $(\widetilde{Q}, \widetilde{V}, \widetilde{A}, \widetilde{\nu})$ for the analogous quantities at $\widetilde{\theta}$.

For later softmax comparisons, we distinguish the advantage bound at the true parameter from the bound uniform over candidate parameters. The bound B_A^Θ is defined in (C.17); define

$$B_A^* := 2 + \eta^{-1} \log |\mathcal{B}|. \tag{H.23}$$

The following Lemma H.7 shows that the true normalized advantage is bounded by B_A^*, whereas every candidate advantage indexed by $\theta \in \Theta$ is bounded by B_A^Θ.

Lemma H.7 (Advantage bounds at the true parameter and uniformly over candidate parameters)**.** *For $p \in \Delta(\mathcal{A})$ and $\theta \in \Theta$, define*

$$Q_\theta^p(s,x,b) := \langle \varphi^p(s,x,b), \theta \rangle, \qquad V_\theta^p(s,x) := \eta^{-1} \log \sum_{b' \in \mathcal{B}} \exp\big(\eta Q_\theta^p(s,x,b')\big),$$

and

$$A_\theta^p(s,x,b) := Q_\theta^p(s,x,b) - V_\theta^p(s,x).$$

Then, uniformly over (s,x,b,p),

$$A_{\theta_h^*}^p(s,x,b) \in [-B_A^*, 0],$$

and, uniformly over every $\theta \in \Theta$,

$$A_\theta^p(s,x,b) \in [-B_A^\Theta, 0].$$

Proof of Lemma H.7. The log-sum-exp inequality gives

$$\max_{b'} Q_\theta^p(s,x,b') \le V_\theta^p(s,x) \le \max_{b'} Q_\theta^p(s,x,b') + \eta^{-1} \log |\mathcal{B}|.$$

Hence $A_\theta^p(s,x,b) \le 0$. Moreover, $\|\varphi^p(s,x,b)\|_2 \le B_\varphi$ and $\|\theta\|_2 \le B_\Theta$ imply

$$-B_\varphi B_\Theta \le Q_\theta^p(s,x,b) \le B_\varphi B_\Theta.$$

Therefore

$$A_\theta^p(s,x,b) \ge -2B_\varphi B_\Theta - \eta^{-1} \log |\mathcal{B}| \ge -B_A^\Theta.$$

For the true parameter, the normalized reward convention $r_h \in [0,1]$ gives the looser but convenient envelope $A_{\theta_h^*}^p(s,x,b) \ge -B_A^*$. □

For the covariance comparisons, fix (s,μ,π), $x \in \mathcal{X}$, and $\theta \in \Theta$. Define the policy-averaged feature

$$\varphi^\pi(x,b,s,\mu) := \sum_{a \in \mathcal{A}} \pi(a \mid s,\mu)\, \varphi(s,a,x,b),$$

its centered version $\Delta\varphi^\pi(x,b,s,\mu) := \varphi^\pi(x,b,s,\mu) - \mathbb{E}_{x,s,\mu}^{\pi,\theta}[\varphi^\pi(x,b,s,\mu)]$, and the per-state Fisher (covariance) matrix

$$\Sigma_{x,s,\mu}^{\pi,\theta} := \mathbb{C}_{x,s,\mu}^{\pi,\theta}\big[\varphi^\pi(x,b,s,\mu)\big] = \mathbb{E}_{x,s,\mu}^{\pi,\theta}\big[\Delta\varphi^\pi(x,b,s,\mu)\, \Delta\varphi^\pi(x,b,s,\mu)^\intercal\big], \tag{H.24}$$

where the expectation is taken with respect to $b \sim \nu^\theta(\cdot \mid x; s, \mu, \pi)$.

The next lemma establishes a lower bound on the Hellinger distance between two probability distributions over $\mathcal{B}$.

Lemma H.8 (Lemma G.2 in Chen et al. (2023a))**.** *Let $\nu, \widetilde{\nu} \in \Delta(\mathcal{B})$ be quantal responses with parameters θ and $\widetilde{\theta}$. We let $D_{\mathrm{H}}(\cdot,\cdot)$ denote the Hellinger distance between probability distributions. Then we have*

$$D_{\mathrm{H}}^2(\nu, \widetilde{\nu}) \ge \frac{\eta^2}{8(1+\eta B_A^\Theta)^2} \left\langle \nu\,,\, (\widetilde{A} - A)^2 \right\rangle_{\mathcal{B}}.$$

For myopic followers under the linear reward model, the softmax advantage is $A_\theta = Q_\theta - V_\theta$, where

$$V_\theta = \eta^{-1} \log \sum_{b \in \mathcal{B}} \exp(\eta Q_\theta(b))$$

is the log-partition normalizer and is constant in b. Hence

$$A_\theta - A_{\widetilde{\theta}} = (Q_\theta - Q_{\widetilde{\theta}}) - (V_\theta - V_{\widetilde{\theta}}),$$

and taking variance under any response distribution ν removes the constant $V_\theta - V_{\widetilde{\theta}}$. Therefore

$$\mathbb{V}_\nu[A_\theta - A_{\widetilde{\theta}}] = \mathbb{V}_\nu[Q_\theta - Q_{\widetilde{\theta}}] = \mathbb{V}_\nu[r_\theta - r_{\widetilde{\theta}}] = \left\|\theta - \widetilde{\theta}\right\|_\Sigma^2. \tag{H.25}$$

Moreover, by $\mathbb{E}_\nu[X^2] \geq \mathbb{V}_\nu[X]$,

$$\left\langle \nu\,,\, \left(A_\theta - A_{\widetilde{\theta}}\right)^2 \right\rangle \;\geq\; \mathbb{V}_\nu\big[A_\theta - A_{\widetilde{\theta}}\big] \;=\; \left\|\theta - \widetilde{\theta}\right\|_\Sigma^2 . \tag{H.26}$$

This is the variance identity used in the sequel.

By Lemma H.8 and (H.26), we have

$$\begin{aligned}
&8(\eta^{-1} + B_A^\Theta)^2\, D_{\mathrm{H}}^2\big(\nu^{\theta^*}(\cdot \,|\, x; s, \mu, \pi), \nu^{\widetilde{\theta}}(\cdot \,|\, x; s, \mu, \pi)\big) \\
&\quad\geq \left\langle \nu^{\theta^*}(\cdot \,|\, x; s, \mu, \pi)\,,\, \left(A_{\widetilde{\theta}}(x, \cdot, s, \mu, \pi) - A_{\theta^*}(x, \cdot, s, \mu, \pi)\right)^2 \right\rangle \\
&\quad\geq \mathbb{V}^{\pi,\theta^*}_{x,s,\mu}\big[A_{\widetilde{\theta}}(x, b, s, \mu, \pi) - A_{\theta^*}(x, b, s, \mu, \pi)\big] \\
&\quad= \mathbb{V}^{\pi,\theta^*}_{x,s,\mu}\big[(Q_{\widetilde{\theta}} - Q_{\theta^*})(x, b, s, \mu, \pi)\big],
\end{aligned} \tag{H.27}$$

where the first inequality is Lemma H.8; the second uses the standard identity $\mathbb{E}[X^2] = \mathbb{V}[X] + (\mathbb{E}X)^2 \geq \mathbb{V}[X]$ applied to $X := A_{\widetilde{\theta}} - A_{\theta^*}$ to drop the squared mean, i.e.,

$$\mathbb{E}[(X - \mathbb{E}X)^2] \;=\; \mathbb{E}[X^2] - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 \;=\; \mathbb{E}[X^2] - (\mathbb{E}X)^2 \;\leq\; \mathbb{E}[X^2];$$

and the final equality uses the linear-myopic identity (H.25) (equivalently, equation (G.10) of Chen et al. (2023a)), which states that $\mathbb{V}_\nu[A_\theta - A_{\widetilde{\theta}}] = \mathbb{V}_\nu[Q_\theta - Q_{\widetilde{\theta}}]$ since $V_\theta - V_{\widetilde{\theta}}$ is constant in b.

In what follows, we include the subscript $h \in [H]$. Summing (H.27) over $i \in [t-1]$ leads to

$$\begin{aligned}
&8(\eta^{-1} + B_A^\Theta)^2 \cdot \sum_{i=1}^{t-1} D_{\mathrm{H}}^2\left(\nu^{\theta^*_h}(\cdot \,|\, x; s_h^i, \mu_h^i, \pi_h^i), \nu^{\widetilde{\theta}}(\cdot \,|\, x; s_h^i, \mu_h^i, \pi_h^i)\right) \\
&\qquad\geq \sum_{i=1}^{t-1} \mathbb{V}^{\pi_h^i, \theta_h^*}_{x, s_h^i, \mu_h^i}[(Q_{\widetilde{\theta}_h} - Q_{\theta_h^*})(x, b, s_h^i, \mu_h^i, \pi_h^i)] = \sum_{i=1}^{t-1} \left\|\theta_h^* - \widetilde{\theta}_h\right\|^2_{\Sigma^{\pi_h^i, \theta_h^*}_{x, s_h^i, \mu_h^i}} .
\end{aligned}$$

H.7 Proof of Lemma H.9

We first state the lemma for the reader's convenience.

Lemma H.9 (Joint distribution error)**.** *Under the assumptions of Theorem C.10, with probability* $\geq 1 - \zeta$,

$$\mathrm{JDE} \;\lesssim\; \eta B_\Theta H \sqrt{d_F T \log(HT/\zeta)} + C_{\mathrm{JDE}} B_\Theta^2 d_F H \log(HT/\zeta). \tag{H.28}$$

Proof of Lemma H.9. **Outline.** The proof of Lemma H.9 bounds the joint-distribution error $\sum_x \mu(x) \left\|(\widehat{\nu}_h^t - \nu^{\theta_h^*})(\cdot \,|\, x)\right\|_1$ by reducing it, via a sequence of progressively more specific bounds, to a single elliptical-potential argument applied to a μ-weighted aggregated Fisher matrix, thereby avoiding any spurious $|\mathcal{X}|$ factor in the regret. The reduction proceeds as follows:

(a) First, we apply the centered L_1-form softmax-Lipschitz inequality (G.62) of Lemma G.17 (cf. Chen et al., 2023a, Lemma G.1), which gives, for any $\theta_1, \theta_2 \in \Theta$ and any (p, x, s),

$$\left\|\nu^{p,\theta_1}(\cdot|x; s) - \nu^{p,\theta_2}(\cdot|x; s)\right\|_1 \;\leq\; 2\eta \sqrt{\|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}}} \;+\; C_{\mathrm{JDE}}\, \|\theta_1 - \theta_2\|^2_{\Sigma^{p,\theta_2}_{x,s}} ,$$

with $C_{\mathrm{JDE}} = \eta^2 \exp(\eta B_A^\Theta)$. We then convert the per-state $\Sigma^{p,\theta}_{x,s}$-norm into a Ψ-norm form for an arbitrary positive definite matrix $\Psi \succ 0$ via the matrix Cauchy–Schwarz inequality (G.65).

(b) We then exploit the version-space identity (H.29) to show that the population-aggregated parameter error vanishes in the CO setting (Step 1, then averaged over μ in Step 2).

(c) We aggregate the per-state TV bound by a μ-weighted Jensen+linearity argument that commutes with the trace and the elliptical potential (Step 3). Martingale concentration converts conditional expectations to on-trajectory sums (Step 4), the matrix elliptical-potential lemma controls the trajectory sum (Step 5), and the union bound across $h \in [H]$ yields the final result (Step 6).

We are now ready to use the tools established above to derive an upper bound on the JDE in the CO setting. For every (h, t, s, μ), let $(\pi_h^t(\cdot \mid s, \mu), \theta_h^t(s, \mu))$, with $\theta_h^t(s, \mu) \in \Theta_h^{t-1}$, denote the fixed measurable joint maximizing selector underlying (3.2). Along the episode-t trajectory, abbreviate $\theta_h^t = \theta_h^t(s_h, \mu_h)$ and write $\widehat{\nu}_h^t = \nu^{\theta_h^t}$, with the local-policy and query arguments suppressed. By $\theta_h^t \in \Theta_h^{t-1}$ and the definition (3.1), for every $i \in [t-1]$, $h \in [H]$, $x \in \mathrm{supp}(\mu_h^i)$, and $b \in \mathcal{B}$,

$$\nu^{\theta_h^*}(b \,|\, x; s_h^i, \mu_h^i, \pi_h^i) = \widehat{\nu}_h^t(b \,|\, x; s_h^i, \mu_h^i, \pi_h^i). \tag{H.29}$$

The CO version space (3.1) constrains the follower response only on $\mathrm{supp}(\mu_h^i)$, which is precisely the observed part of the population under $\mathcal{P}_h^i = \mu_h^i \otimes \nu_h^i$. For x satisfying $\mu_h^i(x) = 0$, the conditional response $\nu_h^i(\cdot \mid x)$ is not identified from $\mathcal{P}_h^i$.

Recall in (H.24) that the covariance matrix $\Sigma_{x,s_h,\mu_h}^{\pi_h,\theta_h}$ is defined as

$$\begin{aligned}\Sigma_{x,s_h,\mu_h}^{\pi_h,\theta_h} :=& \mathbb{C}_{x,s_h,\mu_h}^{\pi_h,\theta_h}[\varphi^{\pi_h}(x, b, s_h, \mu_h)] \\ =& \mathbb{E}_{x,s_h,\mu_h}^{\pi_h,\theta_h}\big[\varphi^{\pi_h}(x, b, s_h, \mu_h)\varphi^{\pi_h}(x, b, s_h, \mu_h)^\top\big] \\ & - \mathbb{E}_{x,s_h,\mu_h}^{\pi_h,\theta_h}[\varphi^{\pi_h}(x, b, s_h, \mu_h)]\,\mathbb{E}_{x,s_h,\mu_h}^{\pi_h,\theta_h}[\varphi^{\pi_h}(x, b, s_h, \mu_h)^\top],\end{aligned}$$

where the expectation is taken with respect to $b \sim \nu^{\theta_h}(\cdot \mid x; s_h, \mu_h, \pi_h)$.

Step 1: Per-state version-space identity on the support. By the version-space identity (H.29), for every $i \in [t-1]$ and every $x \in \mathrm{supp}(\mu_h^i)$,

$$\nu^{\theta_h^*}(\cdot \mid x; s_h^i, \mu_h^i, \pi_h^i) = \widehat{\nu}_h^t(\cdot \mid x; s_h^i, \mu_h^i, \pi_h^i).$$

Step 1a: distribution equality implies advantage equality. Recall the softmax representation

$$\nu^\theta(b) = \exp(\eta A_\theta(b)), \qquad A_\theta(b) = Q_\theta(b) - V_\theta, \qquad V_\theta = \eta^{-1} \log \sum_{b' \in \mathcal{B}} \exp(\eta Q_\theta(b')).$$

The equality of response distributions for every $b \in \mathcal{B}$ implies $A_{\theta_h^*}(b) = A_{\theta_h^t}(b)$. Hence $A_{\theta_h^t}(b) - A_{\theta_h^*}(b) = 0$ pointwise on $\mathcal{B}$; equivalently, $Q_{\theta_h^t} - Q_{\theta_h^*}$ is constant in b, and the normalization identifies this constant with $V_{\theta_h^t} - V_{\theta_h^*}$.

Step 1b: advantage equality implies zero parameter-error covariance. By the variance identity (H.25),

$$\big\|\theta_h^* - \theta_h^t\big\|^2_{\Sigma^{\pi_h^i,\theta_h^*}_{x,s_h^i,\mu_h^i}} = \mathbb{V}_{b\sim\nu^{\theta_h^*}}\big[(Q_{\theta_h^t} - Q_{\theta_h^*})(b)\big] = \mathbb{V}_{b\sim\nu^{\theta_h^*}}\big[A_{\theta_h^t}(b) - A_{\theta_h^*}(b)\big] = 0, \tag{H.30}$$

which is the per-state CO version-space identity at the parameter-error level. For $x \notin \mathrm{supp}(\mu_h^i)$, the identity need not hold because (3.1) imposes no constraint there. Every such x has weight $\mu_h^i(x) = 0$ in the μ-weighted aggregate below and therefore contributes nothing.

Step 2: Population-aggregated covariance. A direct per-x analysis of the joint distribution error $\sum_x \mu_h(x)\,\big\|(\widehat{\nu}_h^t - \nu^{\theta_h^*})(\cdot|x)\big\|_1$, while feasible, would lead to a spurious $|\mathcal{X}|$ factor in the regret because the elliptical-potential argument is applied separately to each x before summation. To avoid this, we introduce the *population-weighted aggregated covariance* at step h, episode i:

$$\bar{\Sigma}_{s,\mu}^{\pi,\theta} := \sum_{x\in\mathcal{X}} \mu(x)\,\Sigma_{x,s,\mu}^{\pi,\theta},$$

and the cumulative aggregated covariance

$$\Omega_{h,t}^{\theta_h} := \sum_{i=1}^{t-1} \bar{\Sigma}_{s_h^i,\mu_h^i}^{\pi_h^i,\theta_h} = \sum_{i=1}^{t-1}\sum_{x\in\mathcal{X}} \mu_h^i(x)\,\Sigma_{x,s_h^i,\mu_h^i}^{\pi_h^i,\theta_h}. \tag{H.31}$$

Weighting (H.30) by $\mu_h^i(x)$ and summing over $x \in \mathcal{X}$ and $i \in [t-1]$: every term with $x \in \mathrm{supp}(\mu_h^i)$ vanishes by (H.30), while every term with $x \notin \mathrm{supp}(\mu_h^i)$ vanishes because $\mu_h^i(x) = 0$. Hence

$$\big\|\theta_h^* - \theta_h^t\big\|^2_{\Omega_{h,t}^{\theta_h^*}} = \sum_{i=1}^{t-1}\sum_{x\in\mathcal{X}} \mu_h^i(x)\big\|\theta_h^* - \theta_h^t\big\|^2_{\Sigma^{\pi_h^i,\theta_h^*}_{x,s_h^i,\mu_h^i}} = 0. \tag{H.32}$$

Remark H.10 (Why the version-space identity alone is insufficient). The identity (H.29) immediately implies

$$D_{\mathrm{TV}}\big(\nu^{\theta_h^*}(\cdot \,|\, x; s_h^i, \mu_h^i, \pi_h^i),\, \widehat{\nu}_h^t(\cdot \,|\, x; s_h^i, \mu_h^i, \pi_h^i)\big) \;=\; 0,$$

for every $x \in \mathrm{supp}(\mu_h^i)$, $i \in [t-1]$, and $h \in [H]$. This fact does not directly bound the JDE because the JDE is evaluated at the future trajectory points (s_h, μ_h, π_h^t) generated in episode t, which generally differ from the historical points defining the version space from which θ_h^t is selected. The softmax response bound and matrix Cauchy–Schwarz step convert the parameter gap into a bound at the future trajectory points required by the JDE.

Step 3: Bounding the μ-weighted JDE. We now use the bounds established above to control the JDE appearing in Lemma H.5, i.e., we control

$$\mathrm{JDE} = \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \mu_h(x)\left\|\left(\widehat{\nu}_h^t - \nu^{\theta_h^*}\right)(\cdot \,|\, x; s_h, \mu_h, \pi_h^t)\right\|_1\right],$$

where the expectation $\mathbb{E}^t$ is taken with respect to the randomness of the trajectory $\{s_h, \mu_h\}_{h\in[H]}$ generated by the leader's policy π^t under the true model.

Set $\Psi = \Omega_{h,t}^{\theta_h^*} + I_{d_F}$. By (H.32),

$$\big\|\theta_h^* - \theta_h^t\big\|_\Psi^2 \;=\; \big\|\theta_h^* - \theta_h^t\big\|_{\Omega_{h,t}^{\theta_h^*}}^2 + \big\|\theta_h^* - \theta_h^t\big\|_{I_{d_F}}^2 \;=\; \big\|\theta_h^* - \theta_h^t\big\|_2^2 \;\le\; (2B_\Theta)^2, \tag{H.33}$$

where the second equality uses (H.32) and the inequality uses $\Theta = \overline{\mathbb{B}(0, B_\Theta)}$.

Applying (G.62) of Lemma G.17 (centered L_1 form; cf. Chen et al., 2023a, Lemma G.1) to $(\theta_1, \theta_2) = (\theta_h^t, \theta_h^*)$ and any $(p, x, s, \mu) = (\pi_h^t, x, s_h, \mu_h)$, then upper-bounding the per-state Σ-norm via matrix Cauchy–Schwarz with $\|z\|_\Sigma^2 \le \mathrm{Tr}(\Psi^{-1}\Sigma)\cdot\|z\|_\Psi^2$ (which gives $\sqrt{\|z\|_\Sigma^2} \le \sqrt{\mathrm{Tr}(\Psi^{-1}\Sigma)}\cdot\|z\|_\Psi$), we obtain for each fixed $x \in \mathcal{X}$:

$$\left\|(\widehat{\nu}_h^t - \nu^{\theta_h^*})(\cdot \,|\, x; s_h, \mu_h, \pi_h^t)\right\|_1 \;\le\; g\Big(\big\|\theta_h^* - \theta_h^t\big\|_\Psi \cdot \sqrt{\mathrm{Tr}\left(\Psi^{-1}\Sigma_{x,s_h,\mu_h}^{\pi_h^t,\theta_h^*}\right)}\Big),$$

where $g(y) := 2\eta\, y + C_{\mathrm{JDE}}\, y^2$. Then, the aggregation $\sum_x \mu_h(x)\, g\big(\|\theta_h^* - \theta_h^t\|_\Psi \sqrt{\mathrm{Tr}(\Psi^{-1}\Sigma_x)}\big)$ requires different tools for the linear and the quadratic terms:

- **(i)** For the *linear* term $2\eta\, \|\theta_h^* - \theta_h^t\|_\Psi \sqrt{\mathrm{Tr}(\Psi^{-1}\Sigma_x)}$, we take the μ_h-weighted sum over x and apply Jensen's inequality to the concave function $\sqrt{\cdot}$: $\sum_x \mu_h(x)\sqrt{\mathrm{Tr}(\Psi^{-1}\Sigma_x)} \le \sqrt{\sum_x \mu_h(x)\,\mathrm{Tr}(\Psi^{-1}\Sigma_x)} = \sqrt{\mathrm{Tr}(\Psi^{-1}\bar{\Sigma})}$.
- **(ii)** For the *quadratic* term $C_{\mathrm{JDE}}\, \|\theta_h^* - \theta_h^t\|_\Psi^2\, \mathrm{Tr}(\Psi^{-1}\Sigma_x)$, we exchange $\sum_x \mu_h(x)$ with the trace by the linearity of the trace (and the fact that $\|\theta_h^* - \theta_h^t\|_\Psi^2$ is independent of x): $\sum_x \mu_h(x)\,\mathrm{Tr}(\Psi^{-1}\Sigma_x) = \mathrm{Tr}(\Psi^{-1}\bar{\Sigma})$.

Combining (i) and (ii):

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} \mu_h(x)\left\|(\widehat{\nu}_h^t - \nu^{\theta_h^*})(\cdot \,|\, x; s_h, \mu_h, \pi_h^t)\right\|_1 \\
&\le\; 2\eta\,\big\|\theta_h^* - \theta_h^t\big\|_\Psi \sum_{x\in\mathcal{X}} \mu_h(x)\sqrt{\mathrm{Tr}\left(\Psi^{-1}\Sigma_{x,s_h,\mu_h}^{\pi_h^t,\theta_h^*}\right)} + C_{\mathrm{JDE}}\big\|\theta_h^* - \theta_h^t\big\|_\Psi^2 \sum_{x\in\mathcal{X}} \mu_h(x)\,\mathrm{Tr}\left(\Psi^{-1}\Sigma_{x,s_h,\mu_h}^{\pi_h^t,\theta_h^*}\right) \\
&\le\; 2\eta\,\big\|\theta_h^* - \theta_h^t\big\|_\Psi \sqrt{\mathrm{Tr}\left(\Psi^{-1}\bar{\Sigma}_{s_h,\mu_h}^{\pi_h^t,\theta_h^*}\right)} + C_{\mathrm{JDE}}\big\|\theta_h^* - \theta_h^t\big\|_\Psi^2\; \mathrm{Tr}\left(\Psi^{-1}\bar{\Sigma}_{s_h,\mu_h}^{\pi_h^t,\theta_h^*}\right) \\
&=\; g\left(\big\|\theta_h^* - \theta_h^t\big\|_\Psi \cdot \Upsilon_{h,t}\right),
\end{aligned} \tag{H.34}$$

where the second inequality uses Jensen's inequality $\sum_x \mu_h(x)\sqrt{a_x} \le \sqrt{\sum_x \mu_h(x) a_x}$ applied to $a_x = \mathrm{Tr}(\Psi^{-1}\Sigma_x)$, and we have defined the *population-aggregated potential*:

$$\Upsilon_{h,t} := \sqrt{\mathrm{Tr}\left(\Psi^{-1}\bar{\Sigma}^{\pi_h^t,\theta_h^*}_{s_h,\mu_h}\right)} \;=\; \sqrt{\mathrm{Tr}\left((\Omega^{\theta_h^*}_{h,t} + I_{d_F})^{-1}\bar{\Sigma}^{\pi_h^t,\theta_h^*}_{s_h,\mu_h}\right)}.$$

Note crucially that $\Upsilon_{h,t}$ is *not* indexed by x: the μ-weighted aggregation has commuted with both the trace and the elliptical potential.

By (H.33), $\left\|\theta_h^* - \theta_h^t\right\|_\Psi \le 2B_\Theta$. Substituting into (H.34):

$$\sum_{x\in\mathcal{X}} \mu_h(x) \left\|(\widehat{\nu}_h^t - \nu^{\theta_h^*})(\cdot \mid x; s_h, \mu_h, \pi_h^t)\right\|_1 \;\le\; 4\eta B_\Theta \cdot \Upsilon_{h,t} + 4C_{\mathrm{JDE}} B_\Theta^2 \cdot \Upsilon_{h,t}^2. \tag{H.35}$$

Step 4: Trace bound and martingale concentration. Since $\varphi^\pi(x,b,s,\mu)$ is a convex combination of the primitive follower features and $\|\varphi(s,a,x,b)\|_2 \le 1$ under Assumption 2.1, $\|\varphi^\pi(x,b,s,\mu)\|_2 \le 1$. The centered feature satisfies $\|\Delta\varphi^\pi\|_2 \le 2$, hence $\mathrm{Tr}(\Sigma^{\pi,\theta}_{x,s,\mu}) = \mathbb{E}\,\|\Delta\varphi^\pi\|_2^2 \le 4$, and $\mathrm{Tr}(\bar{\Sigma}^{\pi_h^t,\theta_h^*}_{s_h,\mu_h}) = \sum_x \mu_h(x)\,\mathrm{Tr}(\Sigma^{\pi_h^t,\theta_h^*}_{x,s_h,\mu_h}) \le 4$. We apply Lemma E.6 with $L = 4$; the resulting constant factors are absorbed into the implicit constants in (H.40).

As $\Omega^{\theta_h^*}_{h,t} \succeq 0$, we have

$$\Upsilon_{h,t}^2 \;=\; \mathrm{Tr}\left((\Omega^{\theta_h^*}_{h,t} + I_{d_F})^{-1}\bar{\Sigma}^{\pi_h^t,\theta_h^*}_{s_h,\mu_h}\right) \;\le\; \mathrm{Tr}\left(\bar{\Sigma}^{\pi_h^t,\theta_h^*}_{s_h,\mu_h}\right) \le 4.$$

Therefore $\Upsilon_{h,t} \in [0,2]$ for all (h,t), and hence $|\Upsilon_{h,t} - \mathbb{E}^t[\Upsilon_{h,t}]| \le 2$ and $\left|\Upsilon_{h,t}^2 - \mathbb{E}^t[\Upsilon_{h,t}^2]\right| \le 4$.

For each fixed h, conditional on the history before episode t, the martingale difference $\mathbb{E}^t[\Upsilon_{h,t}] - \Upsilon_{h,t}$ lies in $[\mathbb{E}^t[\Upsilon_{h,t}] - 2, \mathbb{E}^t[\Upsilon_{h,t}]]$, an interval of length 2. Applying the one-sided Azuma–Hoeffding inequality with these conditional ranges gives, with probability at least $1 - \delta/(2H)$,

$$\sum_{t=1}^T \mathbb{E}^t[\Upsilon_{h,t}] \;\le\; \sum_{t=1}^T \Upsilon_{h,t} + \sqrt{2T\log(2H/\delta)}. \tag{H.36}$$

Similarly, by (E.11) applied to $\Upsilon_{h,t}^2 \in [0,4]$ (so the parameter R in (E.11) equals 4) with confidence $\delta/(2H)$,

$$\sum_{t=1}^T \mathbb{E}^t[\Upsilon_{h,t}^2] \;\le\; 2\sum_{t=1}^T \Upsilon_{h,t}^2 + 16\log(4H/\delta). \tag{H.37}$$

Combining (H.35) with (H.36) and (H.37), summing over $t \in [T]$, we have with probability at least $1 - \delta/H$:

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}^t\left[\sum_{x\in\mathcal{X}} \mu_h(x) \left\|(\widehat{\nu}_h^t - \nu^{\theta_h^*})(\cdot \mid x; s_h, \mu_h, \pi_h^t)\right\|_1\right] \\
&\le\; 4\eta B_\Theta \sum_{t=1}^T \Upsilon_{h,t} + 8C_{\mathrm{JDE}} B_\Theta^2 \sum_{t=1}^T \Upsilon_{h,t}^2 + 4\eta B_\Theta\sqrt{2T\log(2H/\delta)} + 64 C_{\mathrm{JDE}} B_\Theta^2 \log(4H/\delta).
\end{aligned} \tag{H.38}$$

Step 5: Elliptical potential reduction. We now apply the elliptical potential lemma (Lemma E.6) to the sequence of population-aggregated covariances $\{\bar{\Sigma}^{\pi_h^t,\theta_h^*}_{s_h^t,\mu_h^t}\}_{t=1}^T$. *No $|\mathcal{X}|$ factor appears* in this step: although the per-state covariance $\Sigma^{\pi,\theta}_{x,s,\mu}$ has $\mathrm{Tr}(\Sigma_x) \le 4$ for each x, the *population-aggregated* covariance $\bar{\Sigma} = \sum_x \mu(x)\Sigma_x$ still satisfies $\mathrm{Tr}(\bar{\Sigma}) = \sum_x \mu(x)\,\mathrm{Tr}(\Sigma_x) \le 4\sum_x \mu(x) = 4$ (since μ is a probability distribution). Therefore, we can apply Lemma E.6 directly with $L = 4$ to the μ-weighted sequence $\{\bar{\Sigma}^{\pi_h^t,\theta_h^*}_{s_h^t,\mu_h^t}\}_t$ *without* incurring an additional $|\mathcal{X}|$ factor, which would appear if we naively applied the lemma per x and then summed. Explicitly, with $\mathrm{Tr}(\bar{\Sigma}) \le 4 = L$, and choosing $\lambda = 1$:

$$\sum_{t=1}^T \Upsilon_{h,t} \;\le\; \sqrt{\frac{4T}{\log 5} \cdot d_F \log\left(1 + 4T/d_F\right)} \;\lesssim\; \sqrt{d_F T \log(1 + T/d_F)}, \tag{H.39}$$

and similarly

$$\sum_{t=1}^{T} \Upsilon_{h,t}^2 \;\le\; \frac{4}{\log 5}\, d_F\, \log(1 + 4T/d_F) \;\lesssim\; d_F \log(1 + T/d_F).$$

The multiplicative factors induced by $L = 4$ are absorbed into the implicit constants of the $\lesssim$ notation, and the final rate in (H.40) is unchanged.

Remark on adapted covariates. Lemma E.6 is a pathwise trace-form elliptical-potential inequality: it holds for any realized sequence $\{X_t\}_{t=1}^T$ of PSD matrices satisfying $\mathrm{Tr}(X_t) \le L$, deterministic or adapted. Its proof relies only on deterministic matrix identities, namely the determinant-trace inequality and the trace-AM–GM step; no self-normalized martingale concentration is invoked for the trace bound itself. In our application, $X_t = \bar{\Sigma}^{\pi_h^t, \theta_h^*}_{s_h^t, \mu_h^t}$ is generated adaptively by the algorithm, but the inequality applies conditionally on the realized trajectory. The only probabilistic step is the separate martingale concentration bound that converts realized sums to conditional-expectation sums.

Step 6: Final assembly. Combining (H.38) with (H.39) and summing over $h \in [H]$ via a union bound (so that each step $h \in [H]$ contributes a failure probability of at most δ/H, and the total failure across H steps is at most δ; choosing $\delta = \zeta$),

$$\begin{aligned} \mathrm{JDE} \;&=\; \sum_{h=1}^{H} \sum_{t=1}^{T} \mathbb{E}^t \left[\sum_{x \in \mathcal{X}} \mu_h(x) \left\| (\widehat{\nu}_h^t - \nu^{\theta_h^*})(\cdot \mid x; s_h, \mu_h, \pi_h^t) \right\|_1 \right] \\ &\lesssim\; \eta B_\Theta H \sqrt{d_F T \log(HT/\zeta)} + C_{\mathrm{JDE}} B_\Theta^2 d_F H \log(HT/\zeta), \end{aligned} \tag{H.40}$$

with probability at least $1 - \zeta$. The lower-order martingale terms are absorbed into the $\log(HT/\zeta)$ factor. This completes the proof of Lemma H.9, with no $|\mathcal{X}|$ factor appearing because the μ-weighted aggregation commutes with both the trace operation and the elliptical potential lemma. □