LEADER-FOLLOWER GAMES WITH POLICY RECOMMENDATION: REINFORCEMENT LEARNING UNDER A PRESERVATION CONSTRAINT

Anonymous authors
Paper under double-blind review

ABSTRACT

In leader–follower games (LFGs), policy recommendations provide a self-interested leader with an additional channel, beyond her committed policy, to influence a boundedly rational follower who has a recommendation-sensitive response. This steering effect can lead to outcomes that depart from the quantal Stackelberg equilibrium benchmark and potentially harm the follower's interests. To protect the follower's interests, an external preservation constraint is designed to restrict the leader's recommendations so that the follower's executed response is close to her quantal response, without requiring disclosure of her reward function. We study online reinforcement learning (RL) under this constraint in episodic LFGs with linear function approximation, where the leader cannot observe the follower's reward function or her realized rewards. We propose an algorithm that efficiently infers the follower's response from observed actions and operates under an estimated preservation constraint. We establish complementary positive and negative results on follower protection with and without the preservation constraint. Under the preservation constraint, we establish simultaneous high-probability bounds on absolute leader regret and three measures of the recommendation's effect on the follower. For horizon length H, feature dimension d, and T episodes, the bounds are $\widetilde{O}(H^2 d^{3/2}\sqrt{T})$ for absolute leader regret, $\widetilde{O}(Hd\sqrt{T})$ for cumulative preservation violation and reward gap, and $O(Hd^2\,\mathrm{polylog}(T))$ for cumulative criterion gap, with other problem parameters fixed. Without the preservation constraint, we construct an explicit game in which any leader achieving sublinear regret against the unconstrained benchmark induces linear cumulative follower losses. Together, these results demonstrate the importance of the preservation constraint in protecting the follower from long-term losses caused by the leader's recommendations.

1 INTRODUCTION

Leader–follower games model strategic interactions in which a leader commits to a policy and a follower responds. A boundedly rational follower (Simon, 1955) can be modeled by a quantal response (QR) (McKelvey & Palfrey, 1995), which assigns higher probabilities to actions with higher expected rewards. Quantal Stackelberg equilibrium (QSE) optimizes the leader's commitment against this response (Yang et al., 2012; Černý et al., 2021). In applications, online shopping platforms can recommend actions to influence users' choices (Häubl & Trifts, 2000; Adomavicius et al., 2013). This additional channel can move outcomes away from QSE and harm the follower's interests.

Advice-taking and recommender-system experiments document recommendation sensitivity and anchoring (Bailey et al., 2023; Adomavicius et al., 2013). We model this influence through a Kullback–Leibler (KL) divergence relative to the recommendation, drawing on reference-sensitive stochastic choice and information-cost models (Ortega & Braun, 2013; Matějka & McKay, 2015).

Self-interested platforms can steer recommendations toward their own interests at consumers' expense (Aridor & Gonçalves, 2022; Chen & Tsai, 2024). Regulation of recommendation and ranking systems motivates constraining this influence (European Parliament and Council of the European Union, 2022b;a). We propose a *preservation constraint*, under which the leader continues to optimize her return. A preservation constraint is called *effective* if it satisfies the following three design

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.

requirements. First, it should protect the follower against long-term additional behavioral and reward changes caused by the recommendation. Second, it should retain the original QSE benchmark, so that regulating the recommendation does not reduce either player's reward at the corresponding full-information outcome. Third, implementation should not require disclosure of the follower's true reward function or parameter to either the leader or the regulator. This requirement calls for an estimated preservation constraint whose protection is evaluated under the true response model.

Two questions naturally arise: *Can we design an effective preservation constraint that preserves the QSE benchmark and protects the follower during learning? What losses can arise without this constraint?* We answer the first question affirmatively, and also investigate the second question.

Challenges and contributions. The main challenges and contributions are as follows.

1. **Designing an effective preservation constraint.** We translate the regulator's response target into equality between the recommendation and the QR, quantify approximate preservation by a global total-variation (TV) bound, and retain the full-information QSE benchmark. An estimated constraint implements this target without disclosure of the follower's true reward parameter.
2. **Joint online learning and follower protection.** Observed actions follow the executed response policy (ER), while the QR determines the protection target and benchmark. Likelihood guarantees for the ER therefore do not directly yield the QR error bounds needed for sublinear regret. We exploit invariance to constant reward shifts to center reward errors and use the equality of the fitted ER and QR to control both response errors through covariance bounds. This transfers statistical guarantees from ER data to the QR quantities governing regret and preservation. Combining conditional maximum likelihood estimation (MLE) with optimistic value iteration, we obtain simultaneous high-probability bounds on absolute leader regret, cumulative preservation violation (CPV), reward gap (RG), and criterion gap (CG). Over T episodes, the first three bounds scale as $\widetilde{O}(\sqrt{T})$, while CG grows polylogarithmically. Absolute regret also controls episodes in which recommendation-induced deviations raise the leader's return above the constrained benchmark.
3. **Long-term losses without preservation.** We construct a single game in which every leader with sublinear regret against the unconstrained benchmark induces linear CPV, RG, CG, and cumulative follower reward loss. The construction also covers fixed relaxations of the preservation constraint. A time-varying extension separates the first-order and squared effects of tolerance and gives sufficient conditions for vanishing average follower-side effects.

1.1 Related work

Learning and player protection. Our closest methodological reference is the linear-myopic QSE-learning setting of Chen et al. (2023), which combines learning a hidden response model from actions with optimistic RL. Donahue et al. (2024) study decentralized Stackelberg learning with utility benchmarks that tolerate learning errors. Recommendations change the follower's executed response, and we study leader learning together with protection relative to the QR under the same commitment.

Recommendations and steering. Alanqary et al. (2026) study a moderator who learns recommendations in unknown static games, privately suggests sampled actions, and measures compliance through incentive-to-deviate regret. In our setting, a self-interested leader announces a recommendation policy that enters the follower's response as a KL reference, and also chooses actions to optimize her expected return. Incentive-compatible exploration (Mansour et al., 2015) and payment-based steering (Zhang et al., 2024) provide related approaches to influencing behavior; our preservation constraint regulates the recommendation's additional effect while retaining the QSE benchmark.

Platform incentives and regulation. Studies of platforms that also act as producers or retailers identify conflicts between platform incentives and consumer interests (Aridor & Gonçalves, 2022; Chen & Tsai, 2024). The Digital Services Act and Digital Markets Act provide institutional motivation for external restrictions on recommendation and ranking systems (European Parliament and Council of the European Union, 2022b;a). We study a preservation requirement on the recommendation channel and its implications for online learning and follower protection. An extended discussion of related work is deferred to Section A.

2 PRELIMINARIES AND PROBLEM FORMULATION

For distributions p, q on a finite set, write $\mathrm{TV}(p,q) := \|p-q\|_1/2$ for total variation, $\mathcal{H}$ for Shannon entropy, and KL for Kullback–Leibler (KL) divergence.

2.1 LEADER–FOLLOWER GAMES

We consider an episodic leader–follower game (LFG) $M = (\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \rho_0, P, u, r)$. Here $\mathcal{S}$ is a nonempty finite or countably infinite state space with its discrete sigma-field, $\mathcal{A}$ and $\mathcal{B}$ are nonempty finite action sets, $H \in \mathbb{N}$ is the horizon length, and $\rho_0 \in \Delta(\mathcal{S})$ is the initial-state distribution. For each $h \in [H]$, the transition kernel is $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$, and the deterministic reward functions of the leader and follower are $u_h, r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0,1]$. At the beginning of each episode, the leader chooses a commitment policy $\pi \in \Pi$ and a recommendation policy $y \in \mathcal{Y}$, where the leader policy space and recommendation policy space are $\Pi := \{\pi = \{\pi_h\}_{h\in[H]} : \pi_h : \mathcal{S} \to \Delta(\mathcal{A}),\ \forall h \in [H]\}$ and $\mathcal{Y} := \{y = \{y_h\}_{h\in[H]} : y_h : \mathcal{S} \to \operatorname{int}\Delta(\mathcal{B}),\ \forall h \in [H]\}$, and $\Pi \times \mathcal{Y}$ is the set of all *admissible policy pairs*. The follower responds to the announced pair (π, y) through the model in Section 2.2; the interaction protocol and feedback are specified in Section 4. For any bounded function $f : \mathcal{S} \to \mathbb{R}$, define the transition operator by $(P_h f)(s,a,b) := \sum_{s'\in\mathcal{S}} P_h(s' \mid s,a,b) f(s')$.

Definition 2.1 (Linear LFG (Zhang et al., 2021; Chen et al., 2023))**.** An episodic LFG is linear if, for every $h \in [H]$, there exist a feature map $\phi_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$, vectors $\vartheta_h, \theta_h \in \mathbb{R}^d$, and a vector-valued finite signed measure $\mathbf{m}_h$ on the discrete sigma-field of $\mathcal{S}$ such that $u_h(s,a,b) = \langle \phi_h(s,a,b), \vartheta_h \rangle$, $r_h(s,a,b) = \langle \phi_h(s,a,b), \theta_h \rangle$ and $P_h(s' \mid s,a,b) = \langle \phi_h(s,a,b), \mathbf{m}_h(\{s'\}) \rangle$. We impose the normalization conditions: $\sup_{s,a,b} \|\phi_h(s,a,b)\|_2 \le 1$, $\sum_{s'\in\mathcal{S}} \|\mathbf{m}_h(\{s'\})\|_2 \le B_P$, $\|\vartheta_h\|_2 \le B_\vartheta$, and $\operatorname{diam}(\Theta_h) \le B_\Theta$ for all $h \in [H]$. Here, $\Theta_h \subset \mathbb{R}^d$ is the set of all admissible parameters of the follower's step-h reward function, and it is nonempty, compact, and convex. For every $\theta \in \Theta_h$ and $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, $\langle \phi_h(s,a,b), \theta \rangle \in [0,1]$.

The leader knows the feature maps, the parameter sets, and the bounds B_ϑ, B_Θ, and B_P, while the true parameters ϑ_h^*, θ_h^*, $\mathbf{m}_h^*$, and hence the reward and transition functions are unknown to her.

For a leader policy π, define the expected feature vector $\phi_h^\pi(s,b) := \sum_{a\in\mathcal{A}} \pi_h(a \mid s)\phi_h(s,a,b)$ and the follower's expected reward $r_{\theta,h}^\pi(s,b) := \langle \phi_h^\pi(s,b), \theta \rangle$. The feature normalization in Definition 2.1 implies $\|\phi_h^\pi(s,b)\|_2 \le 1$. For a statewise mixture $p \in \Delta(\mathcal{A})$ at (h,s), define the expected feature vector $\phi_h^p(s,b) := \sum_{a\in\mathcal{A}} p(a)\phi_h(s,a,b)$ and the follower's expected reward $r_{\theta,h}^p(s,b) := \langle \phi_h^p(s,b), \theta \rangle$. The response policies $\mu_{\theta,h}^p$ and $\nu_{\theta,h}^{p,y}$ below likewise replace $\pi_h(\cdot \mid s)$ with p.

2.2 QUANTAL RESPONSE AND EXECUTED RESPONSE

Quantal response. The follower is myopic and boundedly rational (Simon, 1955; Conlisk, 1996; Jones, 1999). With rationality parameter $\eta > 0$, her *quantal response* (QR) maximizes the recommendation-free decision criterion F:

$$\mu_{\theta,h}^\pi(\cdot \mid s) \in \operatorname*{argmax}_{q\in\Delta(\mathcal{B})} F_{\theta,h}^\pi(s,q), \quad \text{where } F_{\theta,h}^\pi(s,q) := \mathbb{E}_{b\sim q}\big[r_{\theta,h}^\pi(s,b)\big] + \eta^{-1}\mathcal{H}(q). \tag{1}$$

Strict concavity gives the unique optimizer

$$\mu_{\theta,h}^\pi(b \mid s) = (\exp\{\eta r_{\theta,h}^\pi(s,b)\})/(\textstyle\sum_{b'\in\mathcal{B}} \exp\{\eta r_{\theta,h}^\pi(s,b')\}) \tag{2}$$

Executed response. For recommendation sensitivity $\tau > 0$, the follower's *executed response* (ER) maximizes the recommendation-dependent criterion G:

$$\nu_{\theta,h}^{\pi,y}(\cdot \mid s) \in \operatorname*{argmax}_{q\in\Delta(\mathcal{B})} G_{\theta,h}^{\pi,y}(s,q), \quad \text{where } G_{\theta,h}^{\pi,y}(s,q) := F_{\theta,h}^\pi(s,q) - \tau \mathrm{KL}\big(q \| y_h(\cdot \mid s)\big). \tag{3}$$

Its unique optimizer is

$$\nu_{\theta,h}^{\pi,y}(b \mid s) = (y_h(b \mid s)^\kappa \exp\{\varepsilon r_{\theta,h}^\pi(s,b)\})/(\textstyle\sum_{b'\in\mathcal{B}} y_h(b' \mid s)^\kappa \exp\{\varepsilon r_{\theta,h}^\pi(s,b')\}), \tag{4}$$

where $\varepsilon := \frac{\eta}{1+\eta\tau}$ and $\kappa := \tau\varepsilon = \frac{\eta\tau}{1+\eta\tau}$. Here κ controls recommendation influence and ε is the effective inverse temperature. Because the KL term restricts the response to the recommendation's

support, exact preservation of the QR requires a full-support recommendation. The main analysis assumes η and τ are known to the leader, while the misspecification setting is studied in Section R.

For a parameter sequence $\theta = \{\theta_h\}_{h\in[H]}$ in the product parameter space $\Theta := \prod_{h\in[H]} \Theta_h$, write the ER and QR policies as $\nu_\theta^{\pi,y} := \{\nu_{\theta_h,h}^{\pi,y}\}_{h\in[H]}$ and $\mu_\theta^\pi := \{\mu_{\theta_h,h}^\pi\}_{h\in[H]}$, respectively. Under the follower's true reward parameter sequence $\theta^* := \{\theta_h^*\}_{h\in[H]} \in \Theta$, the ER corresponding to a policy pair is $\nu_{\theta^*}^{\pi,y}$. The expectation $\mathbb{E}^{\pi,\nu_\theta^{\pi,y}}$ is over trajectories generated by ρ_0, P, π, and $\nu_\theta^{\pi,y}$.

2.3 Leader's value functions

For a leader policy π and follower policy ν, define the leader's action-value and value functions by

$$U_h^{\pi,\nu}(s,a,b) := u_h(s,a,b) + (P_h W_{h+1}^{\pi,\nu})(s,a,b),$$
$$W_h^{\pi,\nu}(s) := \sum_{a\in\mathcal{A},b\in\mathcal{B}} \pi_h(a \mid s)\nu_h(b \mid s) U_h^{\pi,\nu}(s,a,b), \qquad W_{H+1}^{\pi,\nu} \equiv 0.$$

To distinguish the return obtained by directly executing an arbitrary follower policy from the return induced by a recommendation through the follower's response model, we define the leader's expected cumulative returns for the first case and the second case, respectively, by

$$J^{\mathrm{pol}}(\pi,\nu) := \mathbb{E}^{\pi,\nu}\left[\sum_{h=1}^H u_h(s_h,a_h,b_h)\right], \qquad J(\pi,y) := J^{\mathrm{pol}}\big(\pi,\nu_{\theta^*}^{\pi,y}\big).$$

3 Preservation Constraint Design and the QSE Benchmark

3.1 Preservation violation and the preservation constraint

A self-interested leader can use the recommendation to influence the follower while holding her commitment fixed. The regulator evaluates this influence by comparing the ER with the QR (Figure 1).

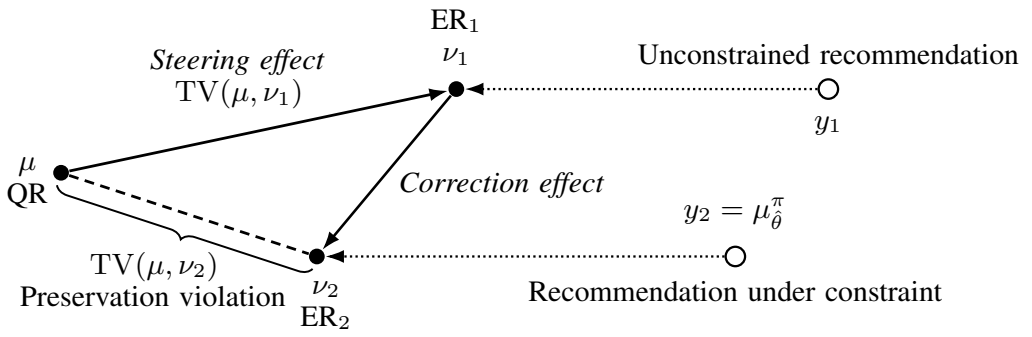

Figure 1: The *steering effect* of recommendations and the *correction effect* of the preservation constraint at fixed s,h,π. Here, the QR is $\mu := \mu_{\theta_h^*,h}^\pi(\cdot \mid s)$ and the ERs are $\nu_i := \nu_{\theta_h^*,h}^{\pi,y_i}(\cdot \mid s)$. The lower branch uses the estimated constraint, with $\hat\theta$ fitted from history. If $\hat\theta = \theta^*$, then $\nu_2 = y_2 = \mu$.

Definition 3.1 (Preservation violation)**.** For $h \in [H]$, $s \in \mathcal{S}$, and an admissible policy pair (π, y), define the preservation violation as $\mathrm{PV}_h(s,\pi,y;\theta_h) := \mathrm{TV}\Big(\nu_{\theta_h,h}^{\pi,y}(\cdot \mid s), \mu_{\theta_h,h}^\pi(\cdot \mid s)\Big)$.

Zero preservation violation is the full-information regulatory target at the given context. It removes the recommendation channel's additional effect while preserving the follower's QR and the leader's freedom to choose her commitment.

The following proposition expresses the ER in terms of the QR and the recommendation.

Proposition 3.2 (Geometric response and KL barycenter)**.** *Fix admissible h,s,π,θ_h and a recommendation $y_h(\cdot \mid s) \in \operatorname{int}\Delta(\mathcal{B})$. With $\kappa = \eta\tau/(1+\eta\tau)$, the response satisfies*

$$\nu_{\theta_h,h}^{\pi,y}(b \mid s) = (\mu_{\theta_h,h}^\pi(b \mid s)^{1-\kappa} y_h(b \mid s)^\kappa)/(\textstyle\sum_{b'\in\mathcal{B}} \mu_{\theta_h,h}^\pi(b' \mid s)^{1-\kappa} y_h(b' \mid s)^\kappa). \tag{5}$$

Equivalently, it is the unique KL barycenter:

$$\nu^{\pi,y}_{\theta_h,h}(\cdot \mid s) = \operatorname*{argmin}_{q\in\Delta(\mathcal{B})} \left\{(1-\kappa)\mathrm{KL}\big(q\|\mu^{\pi}_{\theta_h,h}(\cdot \mid s)\big) + \kappa\mathrm{KL}(q\|y_h(\cdot \mid s))\right\}. \tag{6}$$

Theorem 3.3 (Alignment of three policies)**.** *For every admissible h, s, π, θ_h and $y_h(\cdot \mid s) \in \operatorname{int}\Delta(\mathcal{B})$, we have that $\nu^{\pi,y}_{\theta_h,h}(\cdot \mid s) = \mu^{\pi}_{\theta_h,h}(\cdot \mid s) \iff y_h(\cdot \mid s) = \mu^{\pi}_{\theta_h,h}(\cdot \mid s) \iff y_h(\cdot \mid s) = \nu^{\pi,y}_{\theta_h,h}(\cdot \mid s)$.*

By Theorem 3.3, the QR is the unique fixed point of the recommendation-to-response map on the interior simplex. Theorem 3.3 also converts the regulator's outcome target $\mathrm{PV}_h(s,\pi,y;\theta_h) = 0$ into a direct condition on the leader's choice $y_h(\cdot \mid s) = \mu^{\pi}_{\theta_h,h}(\cdot \mid s)$.

Definition 3.4 (Preservation constraint)**.** For a follower parameter sequence $\theta = \{\theta_h\}_{h\in[H]}$, an admissible policy pair (π, y) satisfies the *preservation constraint* with respect to θ if $y_h(\cdot \mid s) = \mu^{\pi}_{\theta_h,h}(\cdot \mid s)$ for every $h \in [H]$ and $s \in \mathcal{S}$. Let $\mathcal{Z}_\theta$ be the set of all admissible policy pairs satisfying this constraint, i.e., $\mathcal{Z}_\theta := \{(\pi, y) \in \Pi \times \mathcal{Y} : y_h(\cdot \mid s) = \mu^{\pi}_{\theta_h,h}(\cdot \mid s), \forall h \in [H], s \in \mathcal{S}\}$.

The preservation constraint restricts the policy pair (π, y), since the QR on its right-hand side depends on the commitment π. It is equivalent to zero preservation violation at every state and step, including states not reached by a particular policy. The next proposition bounds preservation violation in terms of the recommendation's distance from the QR.

Proposition 3.5 (Recommendation error and preservation violation)**.** *Fix admissible h, s, π, θ_h and a recommendation $y_h(\cdot \mid s) \in \operatorname{int}\Delta(\mathcal{B})$. Abbreviate the QR, ER, and recommendation as $\mu := \mu^{\pi}_{\theta_h,h}(\cdot \mid s)$, $\nu := \nu^{\pi,y}_{\theta_h,h}(\cdot \mid s)$, and $y := y_h(\cdot \mid s)$, respectively. Let the minimum QR probability be $m := \min_{b\in\mathcal{B}} \mu(b)$ and the comparison constant be $C_m := m^{-1}\max\{2, \kappa^{-1}\}$. Then,*

$$C_m^{-1}\,\mathrm{TV}(y,\mu) \le \mathrm{TV}(\nu,\mu) \le \mathrm{TV}(y,\mu).$$

By Lemma C.2, one may replace m by $|\mathcal{B}|^{-1}e^{-\eta}$ to obtain a constant uniform in all h, s, π, θ.

The upper bound connects the recommendation that the leader can choose to the preservation violation that the regulator evaluates. It holds even when y approaches the boundary of the simplex.

3.2 Follower Reward and Decision Criterion

The next proposition quantifies the loss in the recommendation-free decision criterion defined in (1).

Proposition 3.6 (Criterion gap)**.** *Fix admissible h, s, π, θ_h and a recommendation $y_h(\cdot \mid s) \in \operatorname{int}\Delta(\mathcal{B})$. Let $\mu = \mu^{\pi}_{\theta_h,h}(\cdot \mid s)$ and $\nu = \nu^{\pi,y}_{\theta_h,h}(\cdot \mid s)$. Then*

$$F^{\pi}_{\theta_h,h}(s,\mu) - F^{\pi}_{\theta_h,h}(s,\nu) = \eta^{-1}\mathrm{KL}(\nu\|\mu) \ge 2\eta^{-1}\,\mathrm{TV}(\nu,\mu)^2 \ge 0. \tag{7}$$

Thus, preservation is equivalent to zero loss in the criterion.

Remark 3.7 (Why raw-reward preservation is insufficient). The preservation constraint targets the follower's response policy. It excludes recommendation-induced deviations from the QR even when they leave the follower's raw expected reward unchanged or increase it. For example, in a two-action context with equal follower rewards, every mixture preserves the follower's raw expected reward even though the recommendation can keep the ER a positive total-variation distance away from $\mu = (1/2, 1/2)$. Thus, preserving raw expected reward alone is insufficient. By (7), the criterion gap is zero if and only if the ER equals the QR.

3.3 The Full-Information Benchmark

A *quantal Stackelberg equilibrium* (QSE) is a leader–follower policy pair $(\pi^*, \mu^{\pi^*}_{\theta^*})$ satisfying $\pi^* \in \operatorname{argmax}_{\pi\in\Pi} J^{\mathrm{pol}}(\pi, \mu^{\pi}_{\theta^*})$. The lower-level optimization in this bilevel problem determines the follower's QR through (1) at each state and step. A full-information optimum maximizes $J(\pi, y)$ over $\mathcal{Z}_{\theta^*}$. By Theorem 3.3, the benchmark value is

$$J^* := \max_{(\pi,y)\in\mathcal{Z}_{\theta^*}} J(\pi,y) = \max_{\pi\in\Pi} J^{\mathrm{pol}}(\pi, \mu^{\pi}_{\theta^*}). \tag{8}$$

Within the leader policy class specified above, every QSE commitment has the preserving recommendation $y^* = \mu^{\pi^*}_{\theta^*}$, and every optimal preserving pair projects to a QSE. For each fixed commitment,

the induced joint-action and transition laws agree with those of the QR game, so both players' expected rewards agree at the corresponding outcome.

The Bellman characterization below also proves that a maximizing pair exists.

Lemma 3.8 (Bellman characterization of the benchmark)**.** *Let the terminal value be $W^*_{H+1} \equiv 0$, and recursively define the leader's optimal action-value and value functions by*

$$U^*_h(s,a,b) := u_h(s,a,b) + (P_h W^*_{h+1})(s,a,b), \tag{9}$$

$$W^*_h(s) := \max_{p\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} p(a)\mu^p_{\theta^*_h,h}(b \mid s) U^*_h(s,a,b). \tag{10}$$

Then, $J^* = \mathbb{E}_{s_1\sim\rho_0}[W^*_1(s_1)]$.

The myopic follower's response at a state depends on the current leader mixture but not on future recommendations. Each statewise simplex is compact and the Bellman objective is continuous, so backward induction supplies a maximizer. On the countable discrete state space, the resulting policy is measurable. Because the follower's reward parameter is hidden, online implementation uses an estimated QR; the resulting protection is evaluated under the true response model.

4 Online Learning Protocol and Objectives

The leader interacts with the follower and environment over T episodes. Before the initial state in episode t is drawn, she selects and announces (π^t, y^t) from past observations, keeping both policies fixed throughout the episode. The pair induces the follower's ER $\nu^t := \nu^{\pi^t,y^t}_{\theta^*}$, and $s^t_1 \sim \rho_0$. At each step h, both players observe s^t_h and sample actions as $a^t_h \sim \pi^t_h(\cdot \mid s^t_h)$ and $b^t_h \sim \nu^t_h(\cdot \mid s^t_h)$. The leader observes her reward $u_h(s^t_h, a^t_h, b^t_h)$, the follower's action, and the next state $s^t_{h+1} \sim P_h(\cdot \mid s^t_h, a^t_h, b^t_h)$, and uses these data for the next episode's update. The follower knows her own reward function (Zhao et al., 2023); the leader learns from actions without observing follower rewards.

Let $\mathcal{F}^t_0$ denote the history before episode t; the announced policy pair is $\mathcal{F}^t_0$-measurable. Define the episode-start conditional expectation and executed state occupancy by

$$\mathbb{E}^t[\,\cdot\,] := \mathbb{E}[\,\cdot \mid \mathcal{F}^t_0], \qquad d^t_h(s) := \mathbb{P}(s^t_h = s \mid \mathcal{F}^t_0). \tag{11}$$

Estimated preservation constraint. The no-disclosure requirement motivates using an estimate $\hat{\theta}^t_h$ computed from the history. The leader selects (π^t, y^t) under the estimated preservation constraint

$$y^t_h(\cdot \mid s) = \mu^{\pi^t}_{\hat{\theta}^t_h,h}(\cdot \mid s) \quad \text{for every } h \in [H],\ s \in \mathcal{S}. \tag{12}$$

The regulator requires the same reward-parameter estimate to be used in this constraint and in the leader's planner. The leader retains the choice of commitment and estimation method.

Leader performance. For any policy pair (π, y), define its suboptimality by $\mathrm{SubOpt}(\pi, y) := J^* - J(\pi, y)$. The full-information comparator satisfies the preservation constraint, but a pair selected using an estimated constraint need not satisfy the true one. Consequently, $\mathrm{SubOpt}(\pi, y)$ may be negative when (π, y) gives the leader a return above the constrained benchmark. Define the standard and absolute leader regrets respectively by

$$\mathrm{Reg}(T) := \textstyle\sum_{t=1}^T \mathrm{SubOpt}(\pi^t, y^t), \qquad \mathrm{Reg}^{\mathrm{abs}}(T) := \textstyle\sum_{t=1}^T |\,\mathrm{SubOpt}(\pi^t, y^t)|.$$

Absolute regret controls deviations on both sides of the benchmark, whereas an upper bound on standard regret alone does not exclude persistently negative suboptimality.

Follower-side protection. Let $\mu_h^t(\cdot \mid s) := \mu_{\theta_h^*,h}^{\pi^t}(\cdot \mid s)$ denote the true QR under the episode-t commitment. Define cumulative preservation violation, reward gap, and criterion gap by

$$\mathrm{CPV}(T) := \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\mathrm{TV}\big(\nu_h^t(\cdot \mid s_h^t), \mu_h^t(\cdot \mid s_h^t)\big)\right],$$

$$\mathrm{RG}(T) := \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\left|\sum_{b\in\mathcal{B}} (\mu_h^t(b \mid s_h^t) - \nu_h^t(b \mid s_h^t)) r_{\theta_h^*,h}^{\pi^t}(s_h^t, b)\right|\right], \tag{13}$$

$$\mathrm{CG}(T) := \frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t\left[\mathrm{KL}\big(\nu_h^t(\cdot \mid s_h^t)\big\|\mu_h^t(\cdot \mid s_h^t)\big)\right]. \tag{14}$$

All three compare the ER and QR at the same visited state, step, and commitment. CPV measures behavioral deviation, and RG measures the magnitude of the expected-reward difference. By Proposition 3.6, each CG summand is the loss in the recommendation-free criterion F from executing the ER instead of the QR. Moreover, $0 \le \mathrm{RG}(T) \le \mathrm{CPV}(T)$ by Proposition F.2. The objective is vanishing average absolute leader regret together with vanishing average CPV, RG, and CG.

5 Learning under the Estimated Preservation Constraint

We learn the leader's commitment and recommendation under the estimated preservation constraint by combining conditional MLE of the follower's response with optimistic least-squares value iteration (Jin et al., 2020). For each candidate commitment, the recommendation equals the estimated QR. An additional response-uncertainty bonus accounts for estimation error in the leader's planning objective.

5.1 Maximum likelihood estimation (MLE)

At the beginning of episode t, for each step h, define the negative log-likelihood and choose an exact maximum likelihood estimate by

$$\ell_h^t(\theta) := -\textstyle\sum_{i=1}^{t-1} \log \nu_{\theta,h}^{\pi^i,y^i}(b_h^i \mid s_h^i), \qquad \hat{\theta}_h^t \in \operatorname{argmin}_{\theta\in\Theta_h} \ell_h^t(\theta). \tag{15}$$

The negative log-likelihood is convex over Θ_h by Proposition M.1.

For context (s, π, y) and parameter θ, define the ER feature covariance matrix by $\Sigma_h^{\pi,y,\theta}(s) := \mathrm{Cov}_{b\sim\nu_{\theta,h}^{\pi,y}(\cdot|s)}[\phi_h^\pi(s,b)]$. The historical covariance matrix and the fitted feature covariance matrices used to construct it are

$$V_h^t := I_d + \textstyle\sum_{i=1}^{t-1} \hat{\Sigma}_h^i, \qquad \hat{\Sigma}_h^i := \Sigma_h^{\pi^i,y^i,\hat{\theta}_h^i}(s_h^i). \tag{16}$$

Each covariance uses its own episode's estimate and is stored through $V_h^{t+1} = V_h^t + \hat{\Sigma}_h^t$.

For a candidate leader action distribution $p \in \Delta(\mathcal{A})$ at state s, define the estimated QR recommendation and its feature covariance as $y_h^{t,p}(\cdot \mid s) := \mu_{\hat{\theta}_h^t,h}^{p}(\cdot \mid s)$ and $\Gamma_h^t(s,p) := \mathrm{Cov}_{b\sim y_h^{t,p}(\cdot|s)}[\phi_h^p(s,b)]$. By Theorem 3.3,

$$\nu_{\hat{\theta}_h^t,h}^{p,y_h^{t,p}}(\cdot \mid s) = y_h^{t,p}(\cdot \mid s). \tag{17}$$

Let R_Θ be the response confidence radius in (30), and set $L_\eta := \eta e^\eta/2$. Define the response width and QR bonus by $\rho_h^t(s,p) := \sqrt{\mathrm{Tr}((V_h^t)^{-1}\Gamma_h^t(s,p))}$ and $b_h^{t,\mathrm{Q}}(s,p) := \min\{1, L_\eta R_\Theta \rho_h^t(s,p)\}$. The bonus bounds the TV distance between the estimated and true QR on the response-confidence event, as established in Proposition M.4.

5.2 Optimistic least-squares value iteration

Set the remaining-horizon weight $M_h := H - h + 1$ and initialize $\hat{W}_{H+1}^t \equiv 0$. At step h, write $\phi_h^i := \phi_h(s_h^i, a_h^i, b_h^i)$ for the observed feature vector and define the Gram matrix $\Lambda_h^t := I_d + \sum_{i=1}^{t-1} \phi_h^i (\phi_h^i)^\top$.

$$\hat{\omega}_h^t := (\Lambda_h^t)^{-1} \textstyle\sum_{i=1}^{t-1} \phi_h^i \left(u_h(s_h^i, a_h^i, b_h^i) + \hat{W}_{h+1}^t(s_{h+1}^i)\right) \tag{18}$$

is the ridge-regression estimator. The Bellman confidence radii β_h^{U} in (33)–(34) are calibrated backward with deterministic value-function covers to handle the data-dependent targets in (18). For $M > 0$ and $x \in \mathbb{R}$, define the truncation operator by $\mathrm{Trun}_{[0,M]}(x) := \min\{M, \max\{0, x\}\}$. The leader bonus and optimistic U-function are

$$b_h^{t,\mathrm{U}}(s,a,b) := \beta_h^{\mathrm{U}} \sqrt{\phi_h(s,a,b)^\top (\Lambda_h^t)^{-1} \phi_h(s,a,b)},$$
$$\hat{U}_h^t(s,a,b) := \mathrm{Trun}_{[0,M_h]}\Big(\phi_h(s,a,b)^\top \hat{\omega}_h^t + b_h^{t,\mathrm{U}}(s,a,b)\Big). \tag{19}$$

For a candidate $p \in \Delta(\mathcal{A})$, define the optimistic planning objective by

$$\mathcal{Q}_h^t(s,p) := \textstyle\sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} p(a) y_h^{t,p}(b \mid s) \hat{U}_h^t(s,a,b) + M_h b_h^{t,\mathrm{Q}}(s,p).$$

The term $M_h b_h^{t,\mathrm{Q}}(s,p)$ accounts for response uncertainty in planning.

For every t, h, and s, select the leader policy, recommendation policy, and optimistic value function:

$$\pi_h^t(\cdot \mid s) \in \operatorname*{argmax}_{p\in\Delta(\mathcal{A})} \mathcal{Q}_h^t(s,p), \qquad y_h^t(\cdot \mid s) := y_h^{t,\pi_h^t}(\cdot \mid s), \tag{20}$$

$$\hat{W}_h^t(s) := \mathrm{Trun}_{[0,M_h]}\big(\mathcal{Q}_h^t(s, \pi_h^t(\cdot \mid s))\big). \tag{21}$$

We assume access to an exact statewise planning oracle that returns a global maximizer of the objective in (20). Fixed measurable selections are specified in Section D.2, and Corollary E.5 extends the analysis to certified approximate MLEs. Algorithm 1 summarizes our proposed algorithm, **P**olicy **R**ecommendation with **O**ptimistic Value Iteration and **M**aximum Likelihood Estimation (**PROM**).

Algorithm 1 Policy Recommendation with Optimistic Value Iteration and MLE (PROM)

Input: horizon length H, number of episodes T, behavioral parameters (η, τ), feature maps $\{\phi_h\}_{h\in[H]}$, parameter sets $\{\Theta_h\}_{h\in[H]}$, valid bounds $(B_\Theta, B_\vartheta, B_P)$, and confidence level δ.
Initialize an empty data set $\mathcal{D}^0 = \emptyset$.
for episode $t = 1, \ldots, T$ **do**
  For every $h \in [H]$, use $\mathcal{D}^{t-1}$ to compute $\hat{\theta}_h^t$ from (15) and compute V_h^t from (16).
  Set the terminal value function $\hat{W}_{H+1}^t \equiv 0$.
  for step $h = H, H-1, \ldots, 1$ **do**
    Compute $\hat{\omega}_h^t$ and $\hat{U}_h^t$ from (18) and (19) using $\mathcal{D}^{t-1}$.
    Obtain the policy pair (π_h^t, y_h^t) and optimistic value function $\hat{W}_h^t$ by (20)–(21).
  end for
  Announce (π^t, y^t), commit to π^t, and recommend y^t to the follower.
  Execute one episode under (π^t, y^t) according to Section 4.
  Obtain $\mathcal{D}^t$ by appending the trajectory $\{(s_h^t, a_h^t, b_h^t, u_h(s_h^t, a_h^t, b_h^t), s_{h+1}^t)\}_{h\in[H]}$ to $\mathcal{D}^{t-1}$.
  Store $\{(\pi_h^t(\cdot \mid s_h^t), y_h^t(\cdot \mid s_h^t), \hat{\theta}_h^t)\}_{h\in[H]}$ with the corresponding trajectory entries in $\mathcal{D}^t$.
end for

6 Theoretical Results

6.1 Learning and Preservation Guarantees

The following theorem establishes joint guarantees for leader learning and follower protection.

Theorem 6.1 (Joint learning and follower-side protection)**.** *Under Definition 2.1, let $T \geq 1$ and $\delta \in (0,1)$, and run PROM (Algorithm 1) with the response confidence radius in (30) and Bellman confidence radii in (33)–(34). For fixed norm bounds $B_\vartheta, B_P, B_\Theta$, behavioral coefficients η, τ, and action-set cardinalities $|\mathcal{A}|, |\mathcal{B}|$, with probability at least $1-\delta$, simultaneously,*

$$\mathrm{Reg}^{\mathrm{abs}}(T) = \widetilde{O}(H^2 d^{\frac{3}{2}} \sqrt{T}),\ \mathrm{RG}(T) \leq \mathrm{CPV}(T) = \widetilde{O}(Hd\sqrt{T}),\ \mathrm{CG}(T) = O\big(Hd^2 \log^2(eTH/\delta)\big).$$

Here, $\widetilde{O}$ suppresses logarithmic factors in T, d, H, and δ^{-1}. The complete bounds are given in Theorem E.1. Thus PROM learns the constrained benchmark while making the average behavioral deviation, reward gap, and criterion loss vanish.

Under (12), the actual ER is the QR at the interpolated reward parameter $(1-\kappa)\theta_h^* + \kappa\hat{\theta}_h^t$, as shown in Proposition M.4. This identity connects preservation to the response-model error controlled by the conditional likelihood. TV and reward differences are controlled by first-order confidence widths, whereas the criterion gap is a KL divergence controlled by squared widths. The preservation and leader guarantees are also connected beyond this algorithm. For cumulative-return comparisons, define the weighted cumulative preservation violation by $\mathrm{CPV}_{\mathrm{w}}(T) := \sum_{t=1}^{T}\sum_{h=1}^{H} M_h \mathbb{E}^t[\mathrm{PV}_h(s_h^t, \pi^t, y^t; \theta_h^*)]$. For any admissible policy-pair sequence, the follower-policy performance-difference identity gives $\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2\,\mathrm{CPV}_{\mathrm{w}}(T) \le \mathrm{Reg}(T) + 2H\,\mathrm{CPV}(T)$.

6.2 Long-term losses without preservation

We construct a single game in which every leader with sublinear regret against the unconstrained benchmark induces linear growth in CPV, RG, and CG, together with linear cumulative follower reward loss. The same construction applies under a relaxed preservation constraint, so we state the result for the corresponding family of feasible sets. Define the relaxed full-information feasible set by

$$\mathcal{Z}_\theta(\alpha) := \left\{(\pi, y) \in \Pi \times \mathcal{Y} : \mathrm{TV}\Big(y_h(\cdot \mid s), \mu^{\pi}_{\theta_h,h}(\cdot \mid s)\Big) \le \alpha \text{ for every } h, s\right\}, \quad \alpha \in [0,1]. \tag{22}$$

Write the relaxed benchmark value as $J_\alpha^* := \sup_{(\pi,y)\in\mathcal{Z}_{\theta^*}(\alpha)} J(\pi, y)$. For a sequence feasible for this relaxation, define the leader's regret against its relaxed benchmark by $\mathrm{Reg}_\alpha(T) := \sum_{t=1}^{T}\big[J_\alpha^* - J(\pi^t, y^t)\big]$. The comparator J_α^* is the leader's optimal return under the permitted tolerance.

Theorem 6.2 (Long-term follower-side loss under relaxed preservation)**.** *Define $[x]_+ := \max\{x, 0\}$. For fixed $\eta, \tau > 0$, there is a one-state, one-step linear game with two actions for each player and feature dimension four, with $J_1^* = 1$, such that the following holds. For every fixed $\alpha \in (0, 1]$, there is a constant $c_\alpha > 0$ for which every policy-pair sequence in $\mathcal{Z}_{\theta^*}(\alpha)$ satisfies, pathwise,*

$$\mathrm{CPV}(T) = \mathrm{RG}(T) \ge [c_\alpha T - \mathrm{Reg}_\alpha(T)]_+, \quad \mathrm{CG}(T) \ge \frac{2}{\eta\tau}[c_\alpha T - \mathrm{Reg}_\alpha(T)]_+^2.$$

The cumulative signed follower reward loss relative to the QR is at least $c_\alpha T - \mathrm{Reg}_\alpha(T)$. Consequently, $\mathrm{Reg}_\alpha(T) = o(T)$ implies linear CPV, RG, CG, and cumulative follower reward loss.

Corollary 6.3 (Removing the preservation constraint)**.** *In the game of Theorem 6.2, $\mathcal{Z}_{\theta^*}(1)$ is the entire admissible policy-pair class and $J_1^* = 1$. Every unconstrained leader with $\mathrm{Reg}_1(T) = o(T)$ incurs linear $\mathrm{CPV}(T)$, $\mathrm{RG}(T)$, and $\mathrm{CG}(T)$. The value one is a supremum on the full-support recommendation domain; the conclusion applies to learning rules approaching that value and does not require an interior maximizer.*

In this game, the follower's rewards are independent of the leader's action and favor action two, whereas the leader earns a reward only when the follower takes action one. Low regret against the relaxed benchmark therefore forces a nonvanishing average shift toward the follower's lower-reward action. The construction and exact relaxed benchmark values are given in Theorem E.18. A comparison of the protection guarantees with and without preservation is summarized in Table 3.

Tolerance in the estimated constraint. For predictable policy pairs satisfying $\mathrm{TV}(y_h^t(\cdot \mid s), \mu^{\pi^t}_{\hat{\theta}_h^t,h}(\cdot \mid s)) \le \alpha_t$ at every state and step, with $\alpha_t \in [0,1]$ and estimates computed by exact conditional MLE, Theorem E.19 gives protection bounds with additional terms proportional to $\sum_{t=1}^{T}\alpha_t$ for CPV and RG and $\sum_{t=1}^{T}\alpha_t^2$ for CG. For a single learning process with fixed problem parameters, vanishing average tolerance yields vanishing average follower-side effects on a common high-probability event. The complete guarantees are provided in Section E.

7 Conclusion

We study reinforcement learning under an external preservation constraint on the leader-controlled recommendation channel. The constraint retains the QSE benchmark, and its estimated implementation supports joint guarantees for leader learning and follower protection. Our loss construction shows sublinear leader regret can coexist with linear follower-side losses when preservation is relaxed or removed. These results connect regulation of recommendation-induced behavior with learning from hidden follower rewards. Experiments on leader learning and follower protection are in Section T.

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. All AI-assisted edits were reviewed by the authors, who take responsibility for the final content of the paper.

REFERENCES

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

Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Information Systems Research*, 24(4):956–975, 2013.

Arwa Alanqary, Zakaria Baba, Manxi Wu, and Alexandre M. Bayen. Learning to recommend in unknown games. *arXiv preprint arXiv:2602.16998*, 2026. URL https://arxiv.org/abs/2602.16998.

Guy Aridor and Duarte Gonçalves. Recommenders' originals: The welfare effects of the dual role of platforms as producers and recommender systems. *International Journal of Industrial Organization*, 83:102845, 2022.

Francesco Bacchiocchi, Jiarui Gan, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Contract design under approximate best responses. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2203–2220. PMLR, 2025. URL https://proceedings.mlr.press/v267/bacchiocchi25a.html.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

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.

Phoebe E. Bailey, Tarren Leon, Natalie C. Ebner, Ahmed A. Moustafa, and Gabrielle Weidemann. A meta-analysis of the weight of advice in decision-making. *Current Psychology*, 42(28):24516–24541, 2023.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Jakub Černý, Viliam Lisý, Branislav Bošanský, and Bo An. Computing quantal Stackelberg equilibrium in extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5260–5268, 2021.

Nan Chen and Hsin-Tien Tsai. Steering via algorithmic recommendations. *The RAND Journal of Economics*, 55(4):501–518, 2024.

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

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player Markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021.

Jaedeug Choi and Kee-Eung Kim. Nonparametric bayesian inverse reinforcement learning for multiple reward functions. *Advances in neural information processing systems*, 25, 2012.

John Conlisk. Why bounded rationality? *Journal of economic literature*, 34(2):669–700, 1996.

Qiwen Cui, Kaiqing Zhang, and Simon S Du. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. *arXiv preprint arXiv:2302.03673*, 2023.

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. *Journal of Political Economy*, 131(12):3415–3461, 2023.

Kate Donahue, Nicole Immorlica, Meena Jagadeesan, Brendan Lucier, and Aleksandrs Slivkins. Impact of decentralized learning on player utilities in Stackelberg games. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11253–11310. PMLR, 2024.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum markov games. *arXiv preprint arXiv:2207.14211*, 2022.

European Parliament and Council of the European Union. Regulation (EU) 2022/1925 of the european parliament and of the council of 14 september 2022 on contestable and fair markets in the digital sector and amending directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). Official Journal of the European Union, L 265, 2022a. ELI: https://data.europa.eu/eli/reg/2022/1925/oj.

European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec (Digital Services Act). Official Journal of the European Union, L 277, 2022b. ELI: https://data.europa.eu/eli/reg/2022/2065/oj.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Jiarui Gan, Haifeng Xu, Qingyu Guo, Long Tran-Thanh, Zinovi Rabinovich, and Michael Wooldridge. Imitative follower deception in stackelberg games. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 639–657, 2019.

Adam Gleave and Sam Toyer. A primer on maximum causal entropy inverse reinforcement learning. *arXiv preprint arXiv:2203.11409*, 2022.

Nika Haghtalab, Fei Fang, Thanh H Nguyen, Arunesh Sinha, Ariel D Procaccia, and Milind Tambe. Three strategies to success: learning adversary models in security games. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 308–314, 2016.

Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36: 61645–61677, 2023.

Keegan Harris, Zhiwei Steven Wu, and Maria-Florina Balcan. Regret minimization in Stackelberg games with side information. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12944–12976, 2024. URL https://arxiv.org/abs/2402.08576.

Gerald Häubl and Valerie Trifts. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, 19(1):4–21, 2000.

Vijeth Hebbar and Cedric Langbort. Responding to promises: No-regret learning against followers with memory. In *Proceedings of the 7th Annual Learning for Dynamics and Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 647–659. PMLR, 2025. URL https://proceedings.mlr.press/v283/hebbar25a.html.

Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

Dima Ivanov, Paul Dütting, Inbal Talgam-Cohen, Tonghan Wang, and David C. Parkes. Principal-agent reinforcement learning: Orchestrating ai agents with contracts. *arXiv preprint arXiv:2407.18074*, 2024. URL https://arxiv.org/abs/2407.18074.

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.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021a.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021b.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021c.

Bryan D Jones. Bounded rationality. *Annual review of political science*, 2(1):297–321, 1999.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

Hsu Kao, Chen-Yu Wei, and Vijay Subramanian. Decentralized cooperative reinforcement learning with hierarchical information structure. In *International Conference on Algorithmic Learning Theory*, pp. 573–605. PMLR, 2022.

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. doi: 10.1109/TAC.2023.3330797.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International symposium on algorithmic game theory*, pp. 250–262. Springer, 2009.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Junyi Liao, Zihan Zhu, Ethan X. Fang, Zhuoran Yang, and Vahid Tarokh. Decoding rewards in competitive games: Inverse game theory with entropy regularization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37610–37622. PMLR, 2025.

Junyan Liu and Lillian J. Ratliff. Principal-agent bandit games with self-interested and exploratory learning agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39318–39365. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25bf.html.

Junyan Liu, Arnab Maiti, Artin Tajdini, Kevin Jamieson, and Lillian J. Ratliff. Learning to incentivize in repeated principal-agent problems with adversarial agent arrivals. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39085–39105. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25at.html.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

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

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23642–23680. PMLR, 2023.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 565–582, 2015.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, 13(1):165–186, 2023.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

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

John P. Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 7710–7720. PMLR, 2021.

Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pp. 61–73, 1979.

Roger B Myerson. Mechanism design by an informed principal. *Econometrica: Journal of the Econometric Society*, pp. 1767–1797, 1983.

Gergely Neu and Csaba Szepesvári. Training parsers by inverse reinforcement learning. *Machine Learning*, 77:303–337, 2009.

Pedro A. Ortega and Daniel A. Braun. Thermodynamics as a theory of decision-making with information processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153), 2013.

Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2149–2156, 2019.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Antoine Scheid, Daniil Tiapkin, Etienne Boursier, Aymeric Capitaine, Eric Moulines, Michael Jordan, El-Mahdi El-Mhamdi, and Alain Oliviero Durmus. Incentivized learning in principal-agent bandit games. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43608–43631. PMLR, 2024. URL https://proceedings.mlr.press/v235/scheid24a.html.

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

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In *International conference on machine learning*, pp. 10279–10288. PMLR, 2021.

Hamza Virk, Sandro Amaglobeli, and Zuhayr Syed. Blind inverse game theory: Jointly decoding rewards and rationality in entropy-regularized competitive games. *arXiv preprint arXiv:2511.05640*, 2025.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. *arXiv preprint arXiv:2302.06606*, 2023.

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

Jibang Wu, Siyu Chen, Mengdi Wang, Huazheng Wang, and Haifeng Xu. Contractual reinforcement learning: Pulling arms with invisible hands. *arXiv preprint arXiv:2407.01458*, 2024. URL https://arxiv.org/abs/2407.01458.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pp. 3674–3682. PMLR, 2020.

Rong Yang, Fernando Ordonez, and Milind Tambe. Computing optimal strategy against quantal response in security games. In *AAMAS*, pp. 847–854, 2012.

Wenhao Zhan, Jason D Lee, and Zhuoran Yang. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. *arXiv preprint arXiv:2206.01588*, 2022.

Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen McAleer, Andreas Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Steering no-regret learners to a desired equilibrium. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 73–74, 2024.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

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.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067. PMLR, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pp. 1433–1438, 2008.

A Extended Related Work

Learning in leader-follower games. Our work builds on the literature on learning in LFGs, also known as Stackelberg games, under online, offline, tabular, and function-approximation models (Letchford et al., 2009; Peng et al., 2019; Bai et al., 2021; Kao et al., 2022; Zhao et al., 2023; Zhong et al., 2023). Several works study learned follower models, strategic deception, or calibrated best responses (Haghtalab et al., 2016; Gan et al., 2019; Haghtalab et al., 2023).

Our closest methodological reference is the online linear-myopic QSE-learning setting of Chen et al. (2023), which learns a hidden quantal-response model from actions and combines response uncertainty with optimistic value learning. The present paper studies the design and protective effect of an external restriction on a recommendation channel controlled by a self-interested leader. In the deployed interface, actions follow the recommendation-dependent executed response (ER), whereas preservation and the full-information comparator are defined by the true quantal response (QR). Fitted-response interpolation connects these objects along the executed occupancies. Exact preservation retains the full-information QSE benchmark, while the online analysis gives joint guarantees for leader performance, cumulative preservation violation, reward gap, and criterion gap under the estimated constraint. A complementary construction shows that, without preservation, every leader with sublinear regret against the unconstrained benchmark induces linear follower-side losses on one explicit game. Related dynamic work learns an unknown linearly parameterized follower utility under best-response feedback rather than recommendation-sensitive quantal-response feedback (Lauffer et al., 2024).

With external side information, Harris et al. (2024) study online Stackelberg learning with changing contexts and follower types, obtaining no-regret guarantees under restrictions on their generation. In decentralized Stackelberg learning, Donahue et al. (2024) construct utility benchmarks tolerant to learning errors and obtain simultaneous sublinear regret for both learning players. Their tolerances concern approximate best responses, whereas our recommendation tolerance restricts departure from the QR; our follower-side quantities measure ER–QR differences under the same leader commitment.

Learning in Markov games and linear reinforcement learning. Equilibrium learning in Markov games has been studied through centralized optimism and decentralized online-learning methods (Bai et al., 2020; Bai & Jin, 2020; Liu et al., 2020; Jin et al., 2021b; Huang et al., 2021; Xie et al., 2020; Chen et al., 2021; Jin et al., 2021a; Tian et al., 2021; Erez et al., 2022; Liu et al., 2022; Zhan et al., 2022; Mao & Başar, 2023; Wang et al., 2023; Cui et al., 2023). These works primarily target zero-sum or general-sum equilibrium concepts, whereas we study a committing leader and a myopic boundedly rational follower whose actions define a conditional statistical likelihood. On the leader side, our analysis builds on self-normalized linear-bandit concentration (Abbasi-Yadkori et al., 2011) and optimistic least-squares value iteration for linear MDPs (Jin et al., 2020). As in prior QSE learning, fitted response uncertainty makes the continuation values data dependent. We adapt the generated-value-class analysis to the recommendation-conditioned likelihood and estimated-preservation planner, using a deterministic, noncircular cover to support the Bellman confidence bound.

Performative prediction and reinforcement learning. Our recommendation is performative in the sense that the deployed distribution changes the follower-action distribution against which it is subsequently evaluated and learned (Perdomo et al., 2020). Requiring the recommendation to equal its induced response is a fixed-point condition for the recommendation-to-response map, but general performative stability need not coincide with performative optimality or preservation of a designated baseline (Miller et al., 2021). Here the KL geometry makes the QR the unique recommendation satisfying this condition for positive sensitivity, while the learner jointly chooses a leader commitment and a recommendation and estimates a hidden reward-parameterized response from actions. Performative reinforcement learning studies learning in policy-dependent environments (Mandal et al., 2023). In our game, the primitive rewards and transition kernel are fixed, but averaging over the recommendation-sensitive follower yields effective rewards and transitions that depend on the announced policy pair. The resulting joint-action and state-occupancy distributions are endogenous, and our objective is to learn the preservation-constrained QSE benchmark while controlling the recommendation's cumulative effects.

Quantal response, KL references, and inverse learning. We model limited rationality by an entropy-regularized quantal response in the tradition of Simon (1955) and subsequent work (Conlisk, 1996; Jones, 1999; Yang et al., 2012). Advice-taking and recommender-system experiments support systematic but incomplete recommendation sensitivity and anchoring (Bailey et al., 2023; Adomavicius et al., 2013). Free-energy and rational-inattention models provide related foundations for using information costs in stochastic choice (Ortega & Braun, 2013; Matějka & McKay, 2015). We therefore use the KL response as a theoretically grounded and analytically tractable reduced form; experimental work also documents departures from restrictive linear-Shannon specifications (Dean & Neligh, 2023).

The follower's conditional likelihood is a multinomial-logit model closely related to maximum-entropy inverse reinforcement learning and energy-based reward learning (Ziebart et al., 2008; Neu & Szepesvári, 2009; Choi & Kim, 2012; Gleave & Toyer, 2022; Zhu et al., 2023). Recent inverse-game work studies reward identifiability under entropy-regularized equilibrium observations. Liao et al. (2025) characterize reward recovery and identifiability in entropy-regularized two-player zero-sum matrix and Markov games, while Virk et al. (2025) study joint recovery of rewards and rationality parameters under normalization-based identification conditions and extend the analysis to Markov games. Those works concern inverse recovery from equilibrium behavior, whereas our likelihood estimator is embedded in an online commitment-recommendation problem and its predictive uncertainty must be propagated through long-horizon planning. Our regret guarantees require predictive control in response-relevant covariance directions and do not require unique Euclidean recovery of the follower reward vector.

Recommendations, information design, and steering. The requirement that a recommendation equal its induced response is a distributional condition within the specified behavioral family. It is not Bayesian incentive compatibility, truthfulness, or the general collection of obedience inequalities studied in mechanism and information design (Myerson, 1979; 1983; Kamenica & Gentzkow, 2011; Bergemann & Morris, 2019). Bayesian incentive-compatible exploration (Mansour et al., 2015), principal-agent bandits and repeated principal-agent learning with explicit incentives (Scheid et al., 2024; Liu et al., 2025; Liu & Ratliff, 2025), contract design under approximate best responses (Bacchiocchi et al., 2025), sequential principal-agent reinforcement learning with contracts (Wu et al., 2024; Ivanov et al., 2024), inverse Stackelberg learning (Wu et al., 2022), and steering through monetary payments (Zhang et al., 2024) study adjacent mechanisms and feedback structures.

Most closely on the recommendation axis, Alanqary et al. (2026) study a moderator in an unknown static normal-form multi-agent game who commits to a joint recommendation mechanism, privately recommends sampled actions, and observes realized actions. Their quantal-response analysis uses sets of deviations with nonnegative incentive rather than relative action frequencies, and their objective is correlated-equilibrium compliance measured by incentive-to-deviate regret. Utility identification in their analysis is driven by the resulting deviation-support information. Our public distribution-valued recommendation enters a single follower's full-support logit response as a KL reference. The leader also acts and optimizes her own long-horizon return under unknown Markov dynamics. Our criteria compare the recommendation, ER, and same-commitment QR under the induced occupancy. Thus the differences concern the response map, normative target, decision maker, feedback law, and dynamic objective.

In a different repeated-game model, Hebbar & Langbort (2025) study quantal followers whose response to the leader's commitment may depend on an aggregate of past commitments, modeling reputation and memory. Their feedback loop acts through commitment history rather than a separately announced distribution-valued recommendation, and their objective is regret against followers of unknown types rather than preservation or equality between recommendations and responses under hidden reward feedback. Our leader makes no transfer, the recommendation enters the current response as a public distribution-valued reference, and the leader receives no follower-reward feedback. Our analysis propagates response-prediction uncertainty through the induced state occupancies to control long-horizon leader regret together with CPV, RG, and CG.

B Notation and Assumption Ledger

For distributions p and q over a finite set X, we write the total-variation (TV) distance, Shannon entropy and KL divergence as

$$\mathrm{TV}(p,q) := \frac{1}{2}\sum_{x\in X}|p(x)-q(x)|\,,\ \mathcal{H}(p) := -\sum_{x\in X}p(x)\log p(x),\ \mathrm{KL}(p\|q) := \sum_{x\in X}p(x)\log\frac{p(x)}{q(x)}.$$

The normalization $\mathrm{TV}(p,q) = \|p-q\|_1/2$ is the standard total-variation convention on a finite discrete space (Levin & Peres, 2017). We adopt the convention $0\log 0 = 0$, and whenever a Kullback–Leibler (KL) divergence appears, its second argument has full support.

Table 1 summarizes the notation used throughout the paper. Table 2 states exactly which results use which assumptions.

Table 1: Core notation for the episodic Markov model.

Symbol	Meaning
$\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \rho_0$	State space, leader and follower action sets, horizon, and initial-state distribution.
$\phi_h, \vartheta_h^*, \theta_h^*, \mathbf{m}_h^*$	Known feature map and unknown leader-reward vector, follower-reward vector, and vector-valued transition measure; B_P bounds the total variation of the latter.
$u_h(s,a,b)$	Deterministic leader step reward, observed after the joint action and distinct from the leader action-value function U_h.
π_h, y_h	Leader's commitment policy and public, state-dependent recommendation policy at step h.
$\nu_{\theta,h}^{\pi,y}$	Follower's executed response policy (ER).
$\mu_{\theta,h}^{\pi}$	Follower's quantal response policy (QR) under the same commitment.
Preservation constraint	$y = \mu_\theta^\pi$; equivalent to $\nu = \mu_\theta^\pi$ and $y = \nu$ under the main model.
$\eta, \tau, \varepsilon, \kappa, \lambda_{\mathrm{inv}}$	Rationality, recommendation sensitivity, effective inverse temperature, recommendation exponent, and inverse-map exponent; see Section 2.2 and Proposition C.3.
$U_h^{\pi,\nu}, W_h^{\pi,\nu}, J^{\mathrm{pol}}(\pi,\nu)$	Leader action value, state value, and episodic return.
Λ_h^t, V_h^t	Leader least-squares Gram matrix and historical covariance matrix for the follower response.
$V_h^{t,\gamma}, \Gamma_h^{t,\gamma}, \rho_h^{t,\gamma}$	Fitted historical covariance matrix, current covariance, and response width produced by a certified negative-log-likelihood-gap solver. Barred versions denote the behaviorally misspecified construction.
$R_{\Theta,\gamma}, R_{\mathrm{mis}}^{t,\gamma}$	Confidence radii inflated by the well-specified or fitted negative-log-likelihood-optimization certificate.
$b^{\mathrm{U}}, b^{\mathrm{F}}, b^{\mathrm{Q}}, b^{\mathrm{P}}$	Bellman bonus, follower-response bonus, QR bonus, and analysis-only preservation certificate; β_h^{U} is the Bellman confidence radius and $\chi_{T,h}$ is its uniform net-extension term.
$\mathrm{RRD}, \mathrm{RRD}_{\mathrm{w}}$	Recommendation–response distance and its remaining-horizon-weighted version.
M_h, S_H	Remaining-horizon weight $M_h = H-h+1$ and total weight $S_H = H(H+1)/2$.
$\mathrm{Reg}, \mathrm{Reg}^{\mathrm{abs}}$	Standard and absolute leader regret.
$\mathrm{RRD}_{\mathrm{cum}}, \mathrm{RRD}_{\mathrm{w,cum}}$	Cumulative ordinary and remaining-horizon-weighted RRD.
$\mathrm{CPV}, \mathrm{CPV}_{\mathrm{w}}$	Actual-occupancy cumulative preservation violation and its remaining-horizon-weighted version.
RG	Reward gap: cumulative absolute ER–QR difference in expected follower reward at the same visited context; see (13).
CG	Cumulative loss in the entropy-regularized decision criterion.

Table 2: Assumption ledger. "Robust only" means that the assumption is not needed for the well-specified theorems.

Scope	Assumption
Main results	$\mathcal{S}$ is a nonempty finite set or a countably infinite set with its discrete sigma-field; $\mathcal{A}$ and $\mathcal{B}$ are nonempty finite sets; the step reward functions take values in $[0, 1]$; the follower is myopic; and $\eta > 0, \tau > 0$.
Structural results	The response objective is (3), with the positive coefficients of the main model; no linear transition model is needed. Preservation is equivalent to $y = \mu$ and $y = \nu$ on the full-support domain. The inactive-channel case is treated separately in Section G.2.
Response-model scope	At each visited state, the follower conditions on the publicly specified distribution $y_h(\cdot \mid s)$ rather than only on one sampled action. The KL response is a reduced-form behavioral specification, the follower is myopic, and the well-specified theorem treats η and τ as known.
Well-specified regret	The environment satisfies Definition 2.1; (η, τ) are known; the leader observes states, joint actions, her own rewards, and next states, but no follower-reward feedback; a numerical description of ρ_0 is not an algorithmic input; the main theorem uses the fixed measurable exact-MLE and exact statewise-planning conventions in Section 5, while Corollary E.5 exposes a certified negative-log-likelihood-optimization gap.
Auxiliary RRD-based absolute-regret bridge	For the conversion from RRD to absolute regret in Theorem E.9, arbitrary recommendation sequences require a common positive coordinate lower bound.
Algorithm-specific preservation and absolute regret	Theorem E.1 combines the interpolation identity, the direct preservation certificate, and the bounded-reward inequality $\mathrm{RG}(T) \le \mathrm{CPV}(T)$; no uniform recommendation floor is required beyond the main well-specified model.
Preservation-guarantee scope	CPV, RG, CG, and auxiliary cumulative RRD use episode-start conditional expectations and actual executed occupancies. RG compares rewards under the same visited state and commitment; it does not compare total rewards under different trajectory laws or QSE commitments. The guarantees are cumulative rather than pointwise in every episode or state.
Robust only	Assumption R.1, a supplied upper bound ζ_0 on calibration error, and the local condition $C_{\mathrm{beh}}\zeta_0 \le 1$.
Computational scope	The statistical theorems treat fixed Borel measurable exact negative-log-likelihood minimizers or certified negative-log-likelihood-gap solvers and fixed Borel measurable exact statewise maximizers as optimization interfaces, with the corresponding fitted-objective conventions used in the robust branch. A numerical implementation must additionally specify the representation of Θ_h, the projection or separation procedures, state and feature queries, arithmetic precision, and a global statewise optimization method.

C Auxiliary Protection Quantities and Recommendation–Response Distance

C.1 Occupancy representations and weighted preservation

For cumulative-return comparisons, set the remaining-horizon weight $M_h := H - h + 1$ and total weight $S_H := \sum_h M_h = H(H+1)/2$. The weighted cumulative preservation violation is

$$\mathrm{CPV}_{\mathrm{w}}(T) := \sum_{t=1}^{T} \sum_{h=1}^{H} M_h \mathbb{E}^t \left[\mathrm{PV}_h(s_h^t, \pi^t, y^t; \theta_h^*) \right].$$

For RG in (13), the absolute value prevents reward increases in some contexts from cancelling losses in others.

In particular, RG also bounds the cumulative positive part of the follower's reward loss caused by the recommendation.

The corresponding loss in the QR decision criterion is

$$\begin{aligned} \mathrm{CG}(T) &= \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}^t \left[F^{\pi^t}_{\theta_h^*,h}\left(s_h^t, \mu^{\pi^t}_{\theta_h^*,h}(\cdot \mid s_h^t)\right) - F^{\pi^t}_{\theta_h^*,h}\left(s_h^t, \nu^{\pi^t,y^t}_{\theta_h^*,h}(\cdot \mid s_h^t)\right)\right] \\ &= \frac{1}{\eta} \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}^t \mathrm{KL}\left(\nu^{\pi^t,y^t}_{\theta_h^*,h}(\cdot \mid s_h^t) \middle\| \mu^{\pi^t}_{\theta_h^*,h}(\cdot \mid s_h^t)\right), \end{aligned}$$

where the equality follows from Proposition 3.6.

Thus CPV measures behavioral deviation, RG measures the absolute difference in expected reward, and CG measures loss in the entropy-regularized decision criterion. They are distinct quantities, all comparing the ER with the QR under the same state, step, and commitment. In particular, RG and CG are not differences between returns along two different state-occupancy processes.

The cumulative quantities use episode-start conditional expectations and are random through the history-dependent announced policies. For example, CPV can be written as

$$\mathrm{CPV}(T) = \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^t(s)\, \mathrm{PV}_h(s, \pi^t, y^t; \theta_h^*).$$

The nonnegative per-episode sums for CPV and RG are bounded by H, and the corresponding weighted sum is bounded by S_H. The criterion gap is at most $H\eta^{-1}\log(1/m_0)$ per episode, where $m_0 = B^{-1}e^{-\eta}$ and $B = |\mathcal{B}|$. These bounds ensure that the countable-state expectations are well defined. The online objective is vanishing average preservation violation, $\mathrm{CPV}(T)/T \to 0$ for fixed problem parameters, together with vanishing average absolute leader regret and follower-side gaps. Unlike the full-information target, it does not require true preservation in every episode or accurate prediction at unvisited states.

The recommendation–response distance below is an auxiliary quantity used in the regret analysis.

C.2 Recommendation–response distance

Preservation violation compares the ER with the QR. The recommendation–response distance is an auxiliary quantity in the statistical analysis. This appendix defines that auxiliary quantity and its cumulative and weighted versions.

Recall the remaining-horizon weight $M_h := H - h + 1$ and total weight $S_H := H(H+1)/2$. For departures from the preservation constraint, define the recommendation–response distance (RRD)

$$\mathrm{RRD}(\pi, y; \theta) := \mathbb{E}^{\pi, \nu_\theta^{\pi,y}} \left[\sum_{h=1}^{H} \mathrm{TV}\left(y_h(\cdot \mid s_h), \nu^{\pi,y}_{\theta_h,h}(\cdot \mid s_h)\right) \right].$$

The total-variation distance already aggregates over the follower's action space. The outer expectation additionally averages the statewise discrepancy over the complete trajectory induced by the announced

pair and the follower's response. This occupancy weighting measures recommendation–response distance where the pair is actually executed and does not require learning at unreachable states.

For comparisons of cumulative returns, define the remaining-horizon-weighted RRD

$$\mathrm{RRD}_{\mathrm{w}}(\pi, y; \theta) := \mathbb{E}^{\pi, \nu_\theta^{\pi,y}}\left[\sum_{h=1}^{H} M_h \,\mathrm{TV}\Big(y_h(\cdot \mid s_h), \nu_{\theta_h,h}^{\pi,y}(\cdot \mid s_h)\Big)\right].$$

The factor $M_h = H - h + 1$ is an upper bound on the remaining leader return from step h. A response discrepancy at that step can affect the current reward and every subsequent reward, so this weight appears in the follower-policy performance-difference bound (121). Consequently, $\mathrm{RRD}_{\mathrm{w}}$ is the remaining-horizon-weighted RRD used in the model-level bridge of Theorem E.9, while RRD omits the remaining-horizon weight. For PROM (Algorithm 1), the preservation certificate b^{P} controls preservation violation, yielding the absolute-regret bound in Theorem E.7. Since $1 \le M_h \le H$,

$$\mathrm{RRD}(\pi, y; \theta) \le \mathrm{RRD}_{\mathrm{w}}(\pi, y; \theta) \le H\,\mathrm{RRD}(\pi, y; \theta).$$

Both notions ignore unreachable states. Requiring the recommendation to equal the response at every state is stronger because it also constrains unreachable states.

Over T episodes, define the standard leader regret, cumulative RRD, remaining-horizon-weighted cumulative RRD, and absolute leader regret by

$$\mathrm{Reg}(T) := \sum_{t=1}^{T} \mathrm{SubOpt}(\pi^t, y^t), \qquad \mathrm{RRD}_{\mathrm{cum}}(T) := \sum_{t=1}^{T} \mathrm{RRD}(\pi^t, y^t; \theta^*), \tag{23}$$

$$\mathrm{RRD}_{\mathrm{w,cum}}(T) := \sum_{t=1}^{T} \mathrm{RRD}_{\mathrm{w}}(\pi^t, y^t; \theta^*), \qquad \mathrm{Reg}^{\mathrm{abs}}(T) := \sum_{t=1}^{T} \big|\mathrm{SubOpt}(\pi^t, y^t)\big|. \tag{24}$$

The regret criteria use conditional expected episodic returns, while the RRD criteria use conditional expected recommendation–response distances, all under policy pairs selected from the random history. Consequently, the cumulative quantities are random through the data-dependent announced policies. For the executed pair in episode t, the countable-state formulation is

$$\mathrm{RRD}(\pi^t, y^t; \theta^*) = \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^t(s)\,\mathrm{TV}\Big(y_h^t(\cdot \mid s), \nu_{\theta_h^*,h}^{\pi^t, y^t}(\cdot \mid s)\Big).$$

The weighted version inserts M_h inside the outer sum. The ordinary and weighted nonnegative series are bounded by H and S_H, respectively, so they converge absolutely; this establishes that the countable-state definitions are well posed and introduces no finite-state-cardinality factor.

C.3 The recommendation gap

The condition $y = \nu$ can also be expressed through a regularized optimality gap. For a fixed state and step, define the *recommendation gap* by

$$\mathrm{SubOpt}_{\mathrm{R}}(h, s; \pi, y, \theta_h) := \max_{q \in \Delta(\mathcal{B})} G_{\theta_h,h}^{\pi,y}(s, q) - G_{\theta_h,h}^{\pi,y}\big(s, y_h(\cdot \mid s)\big),$$

where $G_{\theta,h}^{\pi,y}$ is defined in (3). The recommendation gap measures the gain in the follower's recommendation-dependent objective from using the ER instead of the recommendation. Since $\mathrm{KL}(y_h(\cdot \mid s) \| y_h(\cdot \mid s)) = 0$, the second term equals the recommendation-free criterion $F_{\theta_h,h}^{\pi}\big(s, y_h(\cdot \mid s)\big)$.

Proposition C.1 (Recommendation gap)**.** *Fix admissible h, s, π, θ_h and a recommendation $y_h(\cdot \mid s) \in \mathrm{int}\,\Delta(\mathcal{B})$. For every $\tau \ge 0$,*

$$\mathrm{SubOpt}_{\mathrm{R}}(h, s; \pi, y, \theta_h) = (\eta^{-1} + \tau)\mathrm{KL}\Big(y_h(\cdot \mid s) \Big\| \nu_{\theta_h,h}^{\pi,y}(\cdot \mid s)\Big). \tag{25}$$

Consequently, the recommendation gap is zero if and only if the recommendation equals the response. Pinsker's inequality also gives

$$\mathrm{SubOpt}_{\mathrm{R}}(h, s; \pi, y, \theta_h) \ge 2(\eta^{-1} + \tau)\,\mathrm{TV}\Big(y_h, \nu_{\theta_h,h}^{\pi,y}\Big)^2,$$

linking the recommendation gap to the squared statewise recommendation–response distance.

C.4 Positivity and the auxiliary inverse-map bound

Recall the number of follower actions $B := |\mathcal{B}|$ and the uniform lower bound $m_0 := B^{-1}e^{-\eta}$ on QR probabilities.

Lemma C.2 (Uniform positivity)**.** *For every admissible h, s, π, θ_h,*

$$\mu^{\pi}_{\theta_h,h}(b \mid s) \geq m_0.$$

More generally, if $y_b \geq \underline{y} > 0$ for all b, then the uniform lower bound on ER probabilities is given by

$$\nu^{\pi,y}_{\theta_h,h}(b \mid s) \geq \underline{n}(\underline{y}) := B^{-1}\underline{y}^{\kappa}e^{-\varepsilon}.$$

Proposition C.3 (Preservation violation and recommendation–response distance)**.** *Let the inverse-map exponent be $\lambda_{\mathrm{inv}} := \eta/\varepsilon = 1 + \eta\tau$. Fix h, s, π, θ_h, let $\mu = \mu^{\pi}_{\theta_h,h}(\cdot \mid s)$, and let $\nu = \nu^{\pi,y}_{\theta_h,h}(\cdot \mid s)$. Suppose*

$$\min_{b\in\mathcal{B}} y_b \geq \underline{y} > 0, \qquad \min_{b\in\mathcal{B}} \nu_b \geq \underline{n} > 0,$$

and set the common lower bound on recommendation and ER probabilities to $\underline{p} := \min\{\underline{y}, \underline{n}\}$. With the comparison constant L_{dev} defined below,

$$\mathrm{TV}(\mu, y) \leq \frac{\lambda_{\mathrm{inv}}}{2\underline{p}}\,\mathrm{TV}(\nu, y), \tag{26}$$

$$\mathrm{TV}(\mu, \nu) \leq L_{\mathrm{dev}}\,\mathrm{TV}(y, \nu), \qquad L_{\mathrm{dev}} := 1 + \frac{\lambda_{\mathrm{inv}}}{2\underline{p}}. \tag{27}$$

These bounds remain valid when $\tau = 0$, in which case $\mu = \nu$ and the second inequality (27) is trivial.

D Optimization Interfaces and Confidence Calibration

This appendix records the measurable-selection arguments and the complete Bellman calibration used by PROM. The main algorithm specifies how the estimates, bonuses, and policies are combined; the constructions below make those statistical interfaces explicit.

D.1 Interaction history and filtration

In online learning, the leader interacts with the follower and the environment over T episodes. At the start of episode $t \in [T]$, the leader selects and announces the policy pair (π^t, y^t) based on past observations, keeping both policies fixed throughout the episode. The announced pair induces the follower's ER $\nu^t := \nu^{\pi^t,y^t}_{\theta^*}$ according to the response model in (4). The initial state is drawn as $s_1^t \sim \rho_0$. At each step $h \in [H]$, both players observe s_h^t and sample their actions conditionally independently as $a_h^t \sim \pi_h^t(\cdot \mid s_h^t)$ and $b_h^t \sim \nu_h^t(\cdot \mid s_h^t)$. They receive their respective rewards $u_h(s_h^t, a_h^t, b_h^t)$ and $r_h(s_h^t, a_h^t, b_h^t)$, and the next state is drawn from $P_h(\cdot \mid s_h^t, a_h^t, b_h^t)$. After the episode, the leader uses the observed trajectory to update the policy pair for the next episode.

Information structure and feedback. The leader has no prior knowledge of either player's reward function or the transition kernel, whereas the follower knows her own reward function (Zhao et al., 2023). At the end of step h in episode t, the leader observes $(s_h^t, a_h^t, b_h^t, u_h(s_h^t, a_h^t, b_h^t), s_{h+1}^t)$, but not the follower's realized reward. Thus, the leader must infer the follower's response model from observed actions while learning her own value functions from her rewards and observed transitions.

Let $\mathcal{F}_0^1$ be trivial, let $\mathcal{F}_0^t$ denote the history available before the policy pair in episode t is announced, and set $\mathcal{F}_0^{t+1} := \mathcal{F}_H^t$ after episode t. The pair (π^t, y^t) is $\mathcal{F}_0^t$-measurable, and the initial state satisfies $\mathbb{P}(s_1^t = s \mid \mathcal{F}_0^t) = \rho_0(s)$ for every $s \in \mathcal{S}$. For every $h \in [H]$, define the pre-action, post-action/pre-transition, and post-transition sigma-fields by

$$\mathcal{G}_h^t := \sigma(\mathcal{F}_{h-1}^t, s_h^t), \qquad \mathcal{H}_h^t := \sigma\big(\mathcal{G}_h^t, a_h^t, b_h^t, u_h(s_h^t, a_h^t, b_h^t)\big), \qquad \mathcal{F}_h^t := \sigma(\mathcal{H}_h^t, s_{h+1}^t). \tag{28}$$

Conditional on $\mathcal{G}_h^t$, the joint action distribution factorizes as

$$\mathbb{P}(a_h^t = a, b_h^t = b \mid \mathcal{G}_h^t) = \pi_h^t(a \mid s_h^t)\nu_h^t(b \mid s_h^t).$$

For every bounded function $f : \mathcal{S} \to \mathbb{R}$, the Markov transition kernel satisfies

$$\mathbb{E}[f(s_{h+1}^t) \mid \mathcal{H}_h^t] = (P_h f)(s_h^t, a_h^t, b_h^t). \tag{29}$$

The episode-start conditional expectation and executed state occupancy are defined in (11).

D.2 Measurable likelihood and planning selections

Likelihood minimization. The finite-history space before episode t is standard Borel; for every fixed $\theta \in \Theta_h$, $\ell_h^t(\theta)$ is $\mathcal{F}_0^t$-measurable, and, for every realized pre-episode history, $\theta \mapsto \ell_h^t(\theta)$ is continuous. Because Θ_h is compact metric, the measurable maximum theorem yields an $\mathcal{F}_0^t$-measurable minimizer, and we fix one such exact-MLE selector whenever the argmin is not unique.

Statewise planning. For every fixed (t, h, s), the map $p \mapsto \mathcal{Q}_h^t(s, p)$ is continuous on the compact simplex $\Delta(\mathcal{A})$, so its argmax is nonempty and compact. Starting from the deterministic terminal value $\hat{W}_{H+1}^t \equiv 0$ and proceeding backward, the objective at every step is jointly $\mathcal{F}_0^t \otimes \mathcal{B}(\Delta(\mathcal{A}))$-measurable in the episode-start history and p: the ridge update, finite response sums, covariance widths, bonuses, and truncation preserve measurability, while the fixed MLE and later-step planning selectors supply the induction hypothesis. The measurable maximum theorem therefore provides an $\mathcal{F}_0^t$-measurable maximizer, and we fix one such Borel selection rule once and for all. Since $\mathcal{S}$ is countable and discrete, these coordinatewise selectors assemble into an $\mathcal{F}_0^t$-measurable Markov policy. When $\mathcal{S}$ is infinite, this complete rule is fixed before the episode and evaluated on demand at the states required by regression or execution; repeated queries to the same state return the same output, so no within-episode replanning occurs.

D.3 Response confidence and auxiliary certificates

For $x > 0$, define the Hellinger lower-bound constant and the TV sensitivity constant by

$$c_{\mathrm{H}}(x) := \frac{(1 - e^{-x})^2}{8}, \qquad L_x := \frac{x e^x}{2}.$$

Define the likelihood confidence threshold and the response confidence radius by

$$\beta_\Theta(\delta) := 4\left[d \log(1 + 16 B_\Theta \varepsilon T) + \log \frac{16HT}{\delta} + 2\right], \quad R_\Theta := \sqrt{B_\Theta^2 + \frac{e^{2\varepsilon}}{c_{\mathrm{H}}(\varepsilon)} \beta_\Theta(\delta)}. \tag{30}$$

The response width is $\rho_h^t(s, p) := \sqrt{\mathrm{Tr}((V_h^t)^{-1} \Gamma_h^t(s, p))}$. Define the follower-response bonus and the QR bonus by

$$b_h^{t,\mathrm{F}}(s, p) := \min\{1, L_\varepsilon R_\Theta \rho_h^t(s, p)\}, \qquad b_h^{t,\mathrm{Q}}(s, p) := \min\{1, L_\eta R_\Theta \rho_h^t(s, p)\}. \tag{31}$$

On the response-confidence event, the first upper-bounds the total-variation distance between the fitted recommendation and the true ER, whereas the second upper-bounds the distance between that recommendation and the true QR. We also define the derived preservation certificate

$$b_h^{t,\mathrm{P}}(s, p) := \min\{1, \kappa L_\eta R_\Theta \rho_h^t(s, p)\}. \tag{32}$$

The preservation certificate bounds preservation violation and is not added to the planning objective.

The follower-response bonus b^{F} controls statewise RRD, while the QR bonus b^{Q} accounts for QR estimation error in planning.

Model-based audit certificates. The true recommendation–response distance and preservation violation are latent because the follower parameter is hidden, but the announced policies, feature specification, action log, fixed estimator, declared confidence inputs, and the quantities b^{F} and b^{P} are reproducible from the algorithm. On the simultaneous confidence event in Proposition M.4, b^{F} and b^{P} bound statewise RRD and preservation violation, respectively, at every candidate context. These certificates hold with high probability under the stated realizability and calibration assumptions.

D.4 Ridge-regression representation

At step h, write $\phi_h^i := \phi_h(s_h^i, a_h^i, b_h^i)$ for the observed feature vector and define the Gram matrix $\Lambda_h^t := I_d + \sum_{i=1}^{t-1} \phi_h^i (\phi_h^i)^\top$. Let the ridge-regression estimator $\hat{\omega}_h^t$ solve the ridge regression problem

$$\hat{\omega}_h^t \in \operatorname*{argmin}_{\omega \in \mathbb{R}^d} \left\{ \sum_{i=1}^{t-1} \left(u_h(s_h^i, a_h^i, b_h^i) + \hat{W}_{h+1}^t(s_{h+1}^i) - (\phi_h^i)^\top \omega \right)^2 + \|\omega\|_2^2 \right\}.$$

The objective is strictly convex, so its unique minimizer is the estimator in (18).

D.5 Backward calibration of Bellman confidence

The Bellman confidence radius β_h^{U} accounts for transition noise and the data-dependent continuation value. Equations (33)–(34) specify its deterministic backward calibration together with a cover of the generated value class, so the radius selection is noncircular. At the terminal step, deterministic rewards and the absence of a continuation value give $\beta_H^{\mathrm{U}} = B_\vartheta$.

The generated value class admits a finite-dimensional parameterization even when the state space is countably infinite. Its cover and the Bellman radii are defined backward, which removes any circular constant selection. Let $m_d := d(d+1)/2$ be the dimension of the symmetric-matrix space, initialize the logarithmic covering bound $\mathfrak{L}_{H+1} := 0$, and define the uniform net-extension term by

$$\chi_{T,h} := \mathbf{1}\{h < H\}\left(T^{-1/2} + \frac{\sqrt{T+1}}{T} B_P\right).$$

The quantity $\chi_{T,h}$ is a uniform explicit upper bound on the extension cost of a T^{-1}-net in the Bellman regression; it is zero at the terminal step. Because the leader's reward is deterministic, the only centered regression noise is the continuation-value transition noise, whose range is $H-h$ rather than $M_h = H-h+1$. For $h = H, H-1, \dots, 1$, first define the Bellman confidence radius by

$$\beta_h^{\mathrm{U}} := \frac{H-h}{2}\sqrt{2\left[d\log(1+T/d) + \mathfrak{L}_{h+1} + \log(16H/\delta)\right]} + B_\vartheta + (H-h)B_P + \chi_{T,h}. \tag{33}$$

In particular, $\beta_H^{\mathrm{U}} = B_\vartheta$: at the terminal step there is neither continuation-value noise nor a future-value covering cost. Then define the covering constant, net scale, and logarithmic covering bound by

$$\begin{aligned} K_h^{\mathrm{cov}} &:= 1 + \sqrt{T}M_h + B_\Theta + 2M_h + \beta_h^{\mathrm{U}} + M_h\eta + M_h L_\eta R_\Theta\sqrt{1+3\eta}, \\ q_h &:= (8TK_h^{\mathrm{cov}})^{-1}, \\ \mathfrak{L}_h &:= d\log\left(1 + \frac{2\sqrt{T}M_h}{q_h}\right) + d\log\left(1 + \frac{2B_\Theta}{q_h^2}\right) + 2m_d\log\left(1 + \frac{2\sqrt{d}}{q_h^2}\right). \end{aligned} \tag{34}$$

Proposition N.1 proves directly that $\mathfrak{L}_h$ bounds the logarithm of a T^{-1}-cover of the step-h generated value class.

E Detailed Guarantees and Statistical Consequences

This appendix gives the complete bounds underlying Theorem 6.1, their component guarantees, and the numerical-optimization and statistical extensions. The exact MLE and planner are those of Section 5, with the Bellman calibration in Section D.5.

E.1 The complete joint guarantee

Theorem E.1 (Joint learning and follower-side protection)**.** *Let $T \geq 1$ and $\delta \in (0,1)$. Define the bound on cumulative widths by*

$$\mathfrak{C}_{T,d,H}(\delta) := \sqrt{2Td\log(1+T/d)} + \sqrt{2T\log(16H/\delta)}, \tag{35}$$

and the bound on cumulative squared response widths by

$$\mathfrak{S}_{T,d,H}(\delta) := 4d\log(1+T/d) + 2\log(4H/\delta). \tag{36}$$

Under Definition 2.1, run PROM (Algorithm 1) with the response confidence radius R_Θ defined in (30) and the Bellman confidence radii β_h^{U} defined by (33)–(34). With probability at least $1-\delta$, simultaneously,

$$\mathrm{Reg}^{\mathrm{abs}}(T) \leq 2\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}\beta_h^{\mathrm{U}} + R_\Theta\left[L_\varepsilon + (1+2\kappa)L_\eta\right]\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}M_h, \tag{37}$$

$$\mathrm{RG}(T) \leq \mathrm{CPV}(T) \leq H\kappa L_\eta R_\Theta \mathfrak{C}_{T,d,H}(\delta), \tag{38}$$

$$\mathrm{CG}(T) \leq \frac{\eta\kappa^2 e^{2\eta}}{2} R_\Theta^2 H \mathfrak{S}_{T,d,H}(\delta). \tag{39}$$

For fixed norm bounds $B_\vartheta, B_P, B_\Theta$, behavioral coefficients η, τ, and action-set cardinalities $|\mathcal{A}|, |\mathcal{B}|$, as in Corollary E.6, these bounds give

$$\mathrm{Reg}^{\mathrm{abs}}(T) = \widetilde{O}(H^2 d^{3/2}\sqrt{T}), \quad \mathrm{CPV}(T) = \widetilde{O}(Hd\sqrt{T}), \quad \mathrm{RG}(T) = \widetilde{O}(Hd\sqrt{T}),$$

where the $\widetilde{O}$ notation suppresses logarithmic factors in T, d, H, and δ^{-1}. For CG, the exact bound in (39) displays the dependence on the episode count through both R_Θ^2 and $\mathfrak{S}_{T,d,H}(\delta)$.

The corresponding criterion-gap summary is in Theorem 6.1.

The same event also controls standard leader regret and the auxiliary RRD quantities in Theorem E.2.

E.2 Leader Regret, RRD, and Preservation Guarantees

Throughout this subsection, the environment is the linear LFG in Definition 2.1, and the supplied bounds $B_\Theta, B_\vartheta, B_P$ are valid. The behavioral coefficients (η, τ) supplied to PROM equal the follower's true coefficients; Section R treats calibrated coefficient misspecification separately.

Theorem E.2 (Leader regret and recommendation–response distance)**.** *Let $T \geq 1$ and $\delta \in (0,1)$. Run PROM (Algorithm 1) with the fixed measurable exact negative-log-likelihood minimizers in (15), the fixed measurable exact statewise planning rule in (20), and the confidence quantities in (30) and (33). Then, with probability at least $1-\delta$,*

$$\mathrm{Reg}(T) \leq 2\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} \beta_h^{\mathrm{U}} + \mathfrak{C}_{T,d,H}(\delta) R_\Theta (L_\eta + L_\varepsilon) \sum_{h=1}^{H} M_h, \tag{40}$$

$$\mathrm{RRD}_{\mathrm{cum}}(T) \leq H L_\varepsilon R_\Theta \mathfrak{C}_{T,d,H}(\delta), \tag{41}$$

$$\mathrm{RRD}_{\mathrm{w,cum}}(T) \leq L_\varepsilon R_\Theta \mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h. \tag{42}$$

The quantities $\mathrm{CPV}(T)$ and $\mathrm{CPV}_{\mathrm{w}}(T)$ are defined in Sections C.1 and 4, respectively.

Corollary E.3 (Direct preservation guarantee)**.** *Under the conditions of Theorem E.2, on the same joint event of probability at least $1-\delta$,*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}^t \,\mathrm{TV}\Big(y_h^t(\cdot \mid s_h^t), \mu_{\theta_h^*,h}^{\pi^t}(\cdot \mid s_h^t)\Big) \leq H L_\eta R_\Theta \mathfrak{C}_{T,d,H}(\delta), \tag{43}$$

$$\mathrm{CPV}(T) \leq H\kappa L_\eta R_\Theta \mathfrak{C}_{T,d,H}(\delta), \tag{44}$$

$$\mathrm{CPV}_{\mathrm{w}}(T) \leq \kappa L_\eta R_\Theta \mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h. \tag{45}$$

Together with (41), these inequalities control all three distributions: the fitted recommendation, the true executed response, and the true QR.

The cumulative criterion gap $\mathrm{CG}(T)$ is defined in (14).

Corollary E.4 (Quadratic cumulative criterion-gap guarantee)**.** *Under the conditions of Theorem E.2, on the same joint event of probability at least $1-\delta$,*

$$\mathrm{CG}(T) \leq \frac{\eta\kappa^2 e^{2\eta}}{2} R_\Theta^2 H \mathfrak{S}_{T,d,H}(\delta). \tag{46}$$

For fixed behavioral parameters and norm bounds, the right-hand side is $O(Hd^2 \,\mathrm{polylog}(T))$, with complete logarithmic dependencies retained in the displayed bound.

The criterion-gap guarantee shows that the cumulative recommendation-induced loss in the follower's entropy-regularized criterion grows only polylogarithmically in T. This sharper dependence on T arises because the criterion gap is controlled by squared confidence widths, whereas the TV deviation is controlled at first order; Lemma P.5 provides the corresponding squared-width summation bound.

The standard regret in Theorem E.2 may be negative because the selected pair need not satisfy the true preservation constraint. The RRD, preservation-violation, and criterion-gap quantities are

nonnegative. The factors L_ε and L_η have different roles: L_ε controls the discrepancy between the recommendation and the executed response, whereas L_η controls optimism relative to the quantal-response benchmark. The primary algorithm-specific absolute-regret result for PROM (Algorithm 1) is stated in Theorem E.7. The model-level inverse-map bridge for more general recommendation sequences is stated in Theorem E.9. Approximate negative-log-likelihood minimization is covered by Corollary E.5.

Numerical likelihood optimization. Corollary E.5 extends the guarantees to approximate MLEs with a certified optimization gap by adjusting the response confidence radius and Bellman calibration.

Theorem E.2 uses exact negative-log-likelihood minimizers to keep the principal statement concise. The next corollary isolates the effect of replacing them with outputs having a known deterministic negative-log-likelihood gap.

For the approximate-MLE variant, replace the exact negative-log-likelihood minimizer in (15) by an output $\tilde{\theta}_h^t \in \Theta_h$ of a solver fixed before interaction. Require the output to be $\mathcal{F}_0^t$-measurable and to satisfy

$$\ell_h^t(\tilde{\theta}_h^t) \le \inf_{\theta\in\Theta_h} \ell_h^t(\theta) + \gamma_h^{t,\mathrm{MLE}}, \qquad 0 \le \gamma_h^{t,\mathrm{MLE}} \le \gamma_{\mathrm{MLE}}. \tag{47}$$

We use a superscript γ to denote quantities generated by this approximate-MLE variant. Construct every recommendation and fitted covariance using these approximate minimizers; in particular, use the fitted historical covariance matrix

$$V_h^{t,\gamma} := I_d + \sum_{i=1}^{t-1} \Sigma_h^{\pi^i, y^i, \tilde{\theta}_h^i}(s_h^i)$$

and, for every candidate p, use the estimated QR recommendation, its feature covariance matrix, and the response width given by

$$\begin{aligned} y_h^{t,p,\gamma}(\cdot \mid s) &:= \mu^p_{\tilde{\theta}_h^t,h}(\cdot \mid s), \\ \Gamma_h^{t,\gamma}(s,p) &:= \mathrm{Cov}_{b\sim y_h^{t,p,\gamma}(\cdot|s)}[\phi_h^p(s,b)], \\ \rho_h^{t,\gamma}(s,p) &:= \sqrt{\mathrm{Tr}((V_h^{t,\gamma})^{-1}\Gamma_h^{t,\gamma}(s,p))}. \end{aligned}$$

The approximate variant uses these three objects wherever PROM (Algorithm 1) uses $y_h^{t,p}, \Gamma_h^t, \rho_h^t$. Define the response confidence radius for the approximate MLE by

$$R_{\Theta,\gamma} := \sqrt{B_\Theta^2 + \frac{e^{2\varepsilon}}{c_{\mathrm{H}}(\varepsilon)}\,[\beta_\Theta(\delta) + 2\gamma_{\mathrm{MLE}}]}.$$

Recompute the complete well-specified envelope: start from $\mathfrak{L}_{H+1,\gamma} = 0$, replace R_Θ by $R_{\Theta,\gamma}$ in (33)–(34), and denote the recursively obtained quantities by $\beta^{\mathrm{U}}_{h,\gamma}, K^{\mathrm{cov}}_{h,\gamma}, q_{h,\gamma}, \mathfrak{L}_{h,\gamma}$. Let $\mathcal{W}_{h,\gamma}$ denote the class in (99)–(101) after these substitutions.

Corollary E.5 (Optimization-tolerant follower MLE)**.** *For the approximate-MLE variant just defined, with probability at least $1-\delta/4$, simultaneously for every $t \in [T]$ and $h \in [H]$,*

$$\left\|\tilde{\theta}_h^t - \theta_h^*\right\|_{V_h^{t,\gamma}} \le R_{\Theta,\gamma}. \tag{48}$$

On this event, all conclusions of Proposition M.4 hold for the approximate fitted objects after replacing V_h^t, Γ_h^t, ρ_h^t, and R_Θ by their γ-superscripted counterparts and $R_{\Theta,\gamma}$. After intersecting this event with the Bellman-confidence, first-order potential, and squared-potential events used in Theorem E.2, the conclusions of Theorems E.2 and E.7 and corollaries E.3 and E.4 remain valid with probability at least $1-\delta$ after replacing R_Θ and β_h^{U} by $R_{\Theta,\gamma}$ and $\beta^{\mathrm{U}}_{h,\gamma}$, respectively.

The conditional identities in Corollary E.13 are algebraic facts for every generated policy sequence; substituting the preceding high-probability regret bounds gives the corresponding optimization-tolerant output guarantees. Thus any $\gamma_{\mathrm{MLE}} = \tilde{\mathcal{O}}(d)$ preserves the displayed $\tilde{\mathcal{O}}(\sqrt{T})$ rates for fixed problem parameters.

Dimension-explicit interpretation. The next corollary simplifies the fully explicit bounds while retaining the nonlogarithmic behavioral constants that govern the two response errors. Define the behavioral factors for the follower-response and QR bonuses by

$$\mathfrak{B}_{\mathrm{F}}(\eta,\tau) := L_\varepsilon \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}} = \frac{\sqrt{2}\,\varepsilon e^{2\varepsilon}}{1-e^{-\varepsilon}}, \qquad \mathfrak{B}_{\mathrm{Q}}(\eta,\tau) := L_\eta \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}} = \frac{\sqrt{2}\,\eta e^{\eta+\varepsilon}}{1-e^{-\varepsilon}}. \tag{49}$$

Corollary E.6 (Dimension-explicit rate)**.** *Under the conditions of Theorem E.2, with probability at least $1-\delta$, PROM (Algorithm 1) under exact statewise planning satisfies the following bounds when $B_\vartheta, B_P, B_\Theta, \eta, \tau, |\mathcal{A}|$, and $|\mathcal{B}|$ are treated as fixed problem parameters, while the dependence on H, d, T is displayed explicitly and logarithmic factors are suppressed except in CG:*

$$\begin{aligned} \mathrm{Reg}(T) &\le \tilde{\mathcal{O}}(H^2 d^{3/2}\sqrt{T}), & \mathrm{RRD}_{\mathrm{cum}}(T) &\le \tilde{\mathcal{O}}(Hd\sqrt{T}),\\ \mathrm{RRD}_{\mathrm{w,cum}}(T) &\le \tilde{\mathcal{O}}(H^2 d\sqrt{T}), & \mathrm{CPV}(T) &\le \tilde{\mathcal{O}}(Hd\sqrt{T}),\\ \mathrm{CG}(T) &\le O\big(Hd^2 \log^2(eTH/\delta)\big)\,. \end{aligned}$$

If $\eta \le c_0 \log(e\,|\mathcal{B}|)$ for a fixed constant $c_0 > 0$, and $\tau \le \tau_{\max}$, the constants hidden by $\tilde{\mathcal{O}}$ are polynomial in $|\mathcal{B}|$, with explicit dependence on $c_0, \tau_{\max}$, and the norm bounds.

The preceding convention describes a fixed-action linear-feature regime in which the action cardinalities and norm bounds are held fixed while d varies. It is distinct from the canonical tabular one-step regime below, where $d = |\mathcal{A}|\,|\mathcal{B}|$ and the admissible Euclidean reward-parameter bounds can scale as $\sqrt{d}$.

Additional dependence through β_Θ and the explicit entropy recursion is logarithmic and remains displayed in the preceding formulas. In particular, the response-bonus contribution to leader regret is

$$\tilde{\mathcal{O}}\Big(H^2\left[(L_\eta + L_\varepsilon)B_\Theta\sqrt{d} + (\mathfrak{B}_{\mathrm{F}} + \mathfrak{B}_{\mathrm{Q}})d\right]\sqrt{T}\Big)\,,$$

and the cumulative RRD contribution is

$$\tilde{\mathcal{O}}\Big(H\left[L_\varepsilon B_\Theta\sqrt{d} + \mathfrak{B}_{\mathrm{F}} d\right]\sqrt{T}\Big)\,.$$

Finally, backward induction in (33)–(34) gives $\mathfrak{L}_h = \tilde{\mathcal{O}}(d^2)$ and $\beta_h^{\mathrm{U}} = \tilde{\mathcal{O}}((H-h)d + B_\vartheta + 1)$: the already defined future entropy enters β_h^{U} under a square root, while β_h^{U} itself enters $\mathfrak{L}_h$ only through logarithms. Consequently, under the fixed-parameter convention of Corollary E.6, the generic multi-step Bellman contribution has order $\tilde{\mathcal{O}}(H^2 d\sqrt{Td}) = \tilde{\mathcal{O}}(H^2 d^{3/2}\sqrt{T})$ because the confidence radius of order d is multiplied by the scalar elliptical-potential factor of order $\sqrt{Td}$. When $H = 1$, however, $\hat{W}_2^t \equiv 0$, the continuation-value class is the singleton $\{0\}$, and the terminal confidence radius is $\beta_1^{\mathrm{U}} = B_\vartheta$; Corollary E.10 records the resulting one-step specialization. These observations verify the displayed dimension dependence under the backward, noncircular calibration.

Canonical finite-state tabular scaling. The fixed-parameter convention in Corollary E.6 is a linear-feature regime. For the canonical finite-state representation $\phi_h(s,a,b) = e_{(s,a,b)}$, one has $d = |\mathcal{S}||\mathcal{A}||\mathcal{B}|$. For this representation, one may take $B_P = \sqrt{|\mathcal{S}|d}$ by Cauchy–Schwarz. Consequently, a tabular specialization must substitute the scaling of d, B_P, B_ϑ, and B_Θ before simplifying the rate. The absence of an additional union-bound factor over states does not remove this structural state dependence.

E.3 Absolute Regret

We first state the algorithm-specific absolute-regret guarantee for PROM (Algorithm 1).

Theorem E.7 (Algorithm-specific absolute regret)**.** *Under the conditions of Theorem E.2, on the same event of probability at least $1-\delta$,*

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2\sum_{t=1}^{T}\sum_{h=1}^{H} M_h \mathbb{E}^t\left[b_h^{t,\mathrm{P}}(s_h^t, \pi_h^t)\right] \tag{50}$$

$$\le 2\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}\beta_h^{\mathrm{U}} + R_\Theta[L_\varepsilon + (1+2\kappa)L_\eta]\,\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H} M_h. \tag{51}$$

The connection between preservation and leader performance is captured by the model-level inequality

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2\,\mathrm{CPV}_{\mathrm{w}}(T).$$

By Lemma Q.1, preservation violation bounds the negative part of the signed suboptimality. PROM (Algorithm 1) then controls this deviation directly through b^{P} on its confidence event, giving Theorem E.7 without the additional inverse-map conditioning used by the later RRD-only route.

Corollary E.8 (Dimension-explicit absolute regret)**.** *Under the fixed-parameter convention of Corollary E.6 and the conditions of Theorem E.2, with probability at least $1-\delta$, PROM (Algorithm 1) under exact statewise planning satisfies*

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \tilde{\mathcal{O}}(H^2 d^{3/2}\sqrt{T}).$$

We next state a structural bridge for recommendation sequences satisfying a common positive coordinate lower bound, beyond the particular recommendation sequence generated by PROM (Algorithm 1). The bridge converts remaining-horizon-weighted cumulative RRD into absolute regret through the bound in Proposition C.3.

Theorem E.9 (Model-level absolute-regret bridge)**.** *Let $\{(\pi^t, y^t)\}_{t=1}^T$ be any sequence of policy pairs that may depend on past data. Suppose $y_h^t(b \mid s) \ge \underline{y} > 0$ for every $t \in [T]$, $h \in [H]$, $s \in \mathcal{S}$, and $b \in \mathcal{B}$. Define the uniform lower bound on ER probabilities and the comparison constant by*

$$\underline{n} := |\mathcal{B}|^{-1}\,\underline{y}^{\kappa} e^{-\varepsilon}, \qquad L_{\mathrm{dev}} := 1 + \frac{1+\eta\tau}{2\min\{\underline{y}, \underline{n}\}}.$$

Then

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2L_{\mathrm{dev}}\,\mathrm{RRD}_{\mathrm{w,cum}}(T) \le \mathrm{Reg}(T) + 2HL_{\mathrm{dev}}\,\mathrm{RRD}_{\mathrm{cum}}(T). \tag{52}$$

Recall that $B := |\mathcal{B}|$. For the recommendations generated by PROM (Algorithm 1), Lemma C.2 allows the model-level bridge to take

$$\underline{y} = B^{-1}e^{-\eta}, \qquad \underline{n} = B^{-1}\underline{y}^{\kappa}e^{-\varepsilon} = B^{-(1+\kappa)}e^{-\eta},$$

where the last equality uses $\eta\kappa + \varepsilon = \eta$. Since $\underline{n} \le \underline{y}$, the corresponding inverse-map factor is

$$L_{\mathrm{dev}}^{\mathrm{PROM}} = 1 + \frac{1+\eta\tau}{2} B^{1+\kappa} e^{\eta}.$$

Accordingly, the action-cardinality factor in $L_{\mathrm{dev}}^{\mathrm{PROM}}$ belongs to the model-level bridge in Theorem E.9, while the direct bound for PROM (Algorithm 1) in Theorem E.7 is expressed through the response-confidence and direct-preservation certificates in (51).

The following specialization uses the deterministic leader-reward model and the exact maximum-likelihood and exact-planning interfaces of Theorem E.2.

Corollary E.10 (Dimension-explicit one-step specialization)**.** *Suppose $H = 1$ and $\mathcal{S} = \{s_0\}$. Under the conditions of Theorem E.2, PROM (Algorithm 1) satisfies, with probability at least $1-\delta$,*

$$\mathrm{Reg}(T) \le \mathfrak{C}_{T,d,1}(\delta)\,[2B_\vartheta + R_\Theta(L_\eta + L_\varepsilon)]\,, \tag{53}$$

$$\mathrm{RRD}_{\mathrm{cum}}(T) = \mathrm{RRD}_{\mathrm{w,cum}}(T) \le L_\varepsilon R_\Theta \mathfrak{C}_{T,d,1}(\delta). \tag{54}$$

Moreover,

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathfrak{C}_{T,d,1}(\delta)\,[2B_\vartheta + 3R_\Theta(L_\eta + L_\varepsilon)]\,. \tag{55}$$

Dimension bookkeeping in the one-step case. The model in Corollary E.10 is a policy-recommendation matrix game, while PROM (Algorithm 1) specializes to its exact-MLE, exact-planning one-step implementation. For fixed η and τ, (30) and (35) yield

$$\mathrm{Reg}(T) \le \tilde{\mathcal{O}}\left(\left\{B_\vartheta + (L_\eta + L_\varepsilon)\left[B_\Theta + \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}}\sqrt{d}\right]\right\}\sqrt{Td}\right),$$

$$\mathrm{RRD}_{\mathrm{cum}}(T) = \mathrm{RRD}_{\mathrm{w,cum}}(T) \le \tilde{\mathcal{O}}\left(L_\varepsilon\left[B_\Theta + \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}}\sqrt{d}\right]\sqrt{Td}\right),$$

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \tilde{\mathcal{O}}\left(\left\{B_\vartheta + 3(L_\eta + L_\varepsilon)\left[B_\Theta + \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}}\sqrt{d}\right]\right\}\sqrt{Td}\right).$$

For the canonical one-state tabular specialization, let $\phi(a,b) = e_{(a,b)}$, set $d = |\mathcal{A}|\,|\mathcal{B}|$, take $\vartheta^* \in [0,1]^d$, and let $\Theta \subseteq [0,1]^d$. Then one may take $B_\vartheta \le \sqrt{d}$ and $B_\Theta \le \sqrt{d}$. For fixed η and τ, the bounds on standard regret, absolute regret, and cumulative RRD in (53)–(55) therefore reduce to $\tilde{\mathcal{O}}(d\sqrt{T})$; no additional factor depending on $|\mathcal{B}|$ is hidden beyond its contribution to d and the displayed behavioral constants. This is a sharper upper bound for the deterministic-reward, exact-MLE, exact-planning one-step specialization under the displayed norm and behavioral assumptions. If the behavioral parameters or norm bounds vary with d, their displayed nonlogarithmic dependence must be retained rather than absorbed into $\tilde{\mathcal{O}}$.

Why the one-step upper bound is smaller than a generic Markov bound. The one-step rate follows from the terminal confidence calculation rather than from substituting $H = 1$ into a generic multi-step summary. Chen et al. (2023) establish a generic multi-step regret upper bound for likelihood-based optimistic learning in episodic Markov games. In our one-step specialization, the continuation-value class is $\{0\}$ and the transition martingale vanishes, so $\beta_1^{\mathrm{U}} = B_\vartheta$. Together with the linear covariance-width response bound, these facts yield the stated one-step upper bound under the present assumptions.

The fixed-point and global-deviation results likewise specialize statewise to their matrix-game counterparts.

E.4 Secondary statistical consequences

The remaining corollaries record standard conversions derived from the preceding regret guarantees. They express the same guarantees in total-step, realized-trajectory, single-output, sample-complexity, anytime, and random-horizon forms.

Corollary E.11 (Equivalent bounds in total interaction steps)**.** *Let $N := TH$ be the total number of step interactions. Under the fixed-parameter convention of Corollary E.6 and the conditions of Theorem E.2, with probability at least $1-\delta$, PROM (Algorithm 1) under exact statewise planning satisfies*

$$\mathrm{Reg}(T), \mathrm{Reg}^{\mathrm{abs}}(T) \le \tilde{\mathcal{O}}\left(H^{3/2} d^{3/2} \sqrt{N}\right),$$
$$\mathrm{RRD}_{\mathrm{cum}}(T) \le \tilde{\mathcal{O}}\left(H^{1/2} d \sqrt{N}\right), \quad \mathrm{RRD}_{\mathrm{w,cum}}(T) \le \tilde{\mathcal{O}}\left(H^{3/2} d \sqrt{N}\right).$$

These are algebraic restatements of the episode-indexed bounds and facilitate comparison with work that measures sample size in step interactions.

The main regret criteria are episode-start conditional expectations. They also imply high-probability bounds for the realized leader return and for RRD criteria evaluated along the realized state trajectory. Only the leader return is directly observable. The two RRD criteria below are latent evaluation quantities because they depend on the follower's unknown ER. Let $\mathsf{R}_h^t := u_h(s_h^t, a_h^t, b_h^t)$ denote the leader's realized reward under the paper's deterministic reward model. For compactness, write the statewise recommendation–response distance as $D_h^t := \mathrm{TV}\left(y_h^t(\cdot \mid s_h^t), \nu_{\theta_h^*,h}^{\pi^t, y^t}(\cdot \mid s_h^t)\right)$ and define the realized episodic leader return, RRD, and remaining-horizon-weighted RRD by

$$G^t := \sum_{h=1}^{H} \mathsf{R}_h^t, \qquad C^t := \sum_{h=1}^{H} D_h^t, \qquad C_{\mathrm{w}}^t := \sum_{h=1}^{H} M_h D_h^t.$$

The realized cumulative leader regret, RRD, and remaining-horizon-weighted RRD are

$$\mathrm{Reg}_{\mathrm{traj}}(T) := \sum_{t=1}^{T} (J^* - G^t), \quad \mathrm{RRD}_{\mathrm{traj}}(T) := \sum_{t=1}^{T} C^t, \quad \mathrm{RRD}_{\mathrm{w,traj}}(T) := \sum_{t=1}^{T} C_{\mathrm{w}}^t.$$

Corollary E.12 (Realized trajectory criteria). *For every $\delta_{\mathrm{traj}} \in (0,1)$, with probability at least $1-\delta_{\mathrm{traj}}$, simultaneously,*

$$\mathrm{Reg}_{\mathrm{traj}}(T) \le \mathrm{Reg}(T) + \frac{H}{2}\sqrt{2T\log\frac{3}{\delta_{\mathrm{traj}}}}, \tag{56}$$

$$\mathrm{RRD}_{\mathrm{traj}}(T) \le \mathrm{RRD}_{\mathrm{cum}}(T) + \frac{H}{2}\sqrt{2T\log\frac{3}{\delta_{\mathrm{traj}}}}, \tag{57}$$

$$\mathrm{RRD}_{\mathrm{w,traj}}(T) \le \mathrm{RRD}_{\mathrm{w,cum}}(T) + \frac{S_H}{2}\sqrt{2T\log\frac{3}{\delta_{\mathrm{traj}}}}. \tag{58}$$

Intersecting this event with any of the training events above gives the corresponding explicit realized bounds with total failure probability obtained by a union bound. No sublinear guarantee is asserted for $\sum_{t=1}^{T}|J^* - G^t|$, because irreducible trajectory randomness can make that quantity linear even under an optimal policy.

Corollary E.13 (Online-to-batch guarantee). *Let $\mathcal{F}_0^{T+1}$ be the sigma-field generated by the complete interaction through episode T. After the interaction, draw I uniformly from $[T]$, independently of $\mathcal{F}_0^{T+1}$, and output the complete historical pair $(\hat\pi,\hat y) := (\pi^I, y^I)$. For every integrable $\sigma(\mathcal{F}_0^{T+1}, I)$-measurable random variable Z, define the conditional expectation over the sampled episode index by $\mathbb{E}_I[Z] := \mathbb{E}[Z \mid \mathcal{F}_0^{T+1}]$. Conditionally on $\mathcal{F}_0^{T+1}$, the following identities hold almost surely:*

$$\mathbb{E}_I\,|J^* - J(\hat\pi,\hat y)| = \frac{\mathrm{Reg}^{\mathrm{abs}}(T)}{T},$$

$$\mathbb{E}_I\,\mathrm{RRD}(\hat\pi,\hat y;\theta^*) = \frac{\mathrm{RRD}_{\mathrm{cum}}(T)}{T}, \qquad \mathbb{E}_I\,\mathrm{RRD}_{\mathrm{w}}(\hat\pi,\hat y;\theta^*) = \frac{\mathrm{RRD}_{\mathrm{w,cum}}(T)}{T}.$$

On the high-probability event in Theorem E.2, Theorem E.7 and the two cumulative-RRD bounds control these right-hand sides.

The cumulative guarantees for PROM yield a bound on the expected error of the randomly selected output policy pair. Define the output-error bound by

$$\mathfrak{R}_T^{\mathrm{out}}(\delta) := 2\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}\beta_h^{\mathrm{U}} + 3R_\Theta(L_\eta + L_\varepsilon)\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}M_h + HL_\varepsilon R_\Theta\mathfrak{C}_{T,d,H}(\delta). \tag{59}$$

Corollary E.14 (Expected-error sample complexity). *Under the conditions of Theorem E.2, let $(\hat\pi,\hat y)$ be the randomly selected output in Corollary E.13. On the common training event used in the proof of Theorem E.2, which has probability at least $1-\delta$,*

$$\mathbb{E}_I[|J^* - J(\hat\pi,\hat y)| + \mathrm{RRD}(\hat\pi,\hat y;\theta^*)] = \frac{\mathrm{Reg}^{\mathrm{abs}}(T) + \mathrm{RRD}_{\mathrm{cum}}(T)}{T} \le \frac{\mathfrak{R}_T^{\mathrm{out}}(\delta)}{T}. \tag{60}$$

Under the fixed-parameter convention of Corollary E.6, the last quantity in (60) is

$$\frac{\mathfrak{R}_T^{\mathrm{out}}(\delta)}{T} = \tilde{\mathcal{O}}\left(\frac{H^2 d^{3/2}}{\sqrt{T}}\right).$$

Consequently, it suffices to take $T = \tilde{\mathcal{O}}(H^4 d^3/\epsilon^2)$ to make the conditional expected sum of the two errors at most ϵ, with the constants retaining their explicit dependence on η, τ, and the norm bounds.

Tail conversion. For every $\delta_{\mathrm{out}} \in (0,1)$, Markov's inequality conditional on $\mathcal{F}_0^{T+1}$, followed by the law of total probability over the training event, gives

$$\mathbb{P}\left(|J^* - J(\hat\pi,\hat y)| + \mathrm{RRD}(\hat\pi,\hat y;\theta^*) > \frac{\mathfrak{R}_T^{\mathrm{out}}(\delta)}{T\delta_{\mathrm{out}}}\right) \le \delta + \delta_{\mathrm{out}}, \tag{61}$$

where the probability is taken jointly over the interaction data and the independent uniform index I. Up to logarithmic factors, a single randomly selected historical pair has error at most ϵ with

probability at least $1-\delta-\delta_{\mathrm{out}}$ when $T=\tilde{\mathcal{O}}(H^4d^3/(\epsilon^2\delta_{\mathrm{out}}^2))$. The additional $\delta_{\mathrm{out}}^{-2}$ dependence is not needed for the conditional expected-error guarantee.

Under the conditions of Theorem E.2 and the same output rule as in Corollary E.14, define the value-sensitive output-error bound by

$$\mathfrak{R}^{\mathrm{out}}_{T,\mathrm{w}}(\delta) := 2\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}\beta_h^{\mathrm{U}} + [3(L_\eta+L_\varepsilon)+L_\varepsilon]\,R_\Theta\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}M_h.$$

Corollary E.15 (Value-sensitive output guarantee)**.** *Conditionally on the interaction data, the equality below holds deterministically. On the training event used in Theorem E.2, its right-hand side is bounded as follows:*

$$\mathbb{E}_I[|J^*-J(\hat{\pi},\hat{y})|+\mathrm{RRD}_{\mathrm{w}}(\hat{\pi},\hat{y};\theta^*)] = \frac{\mathrm{Reg}^{\mathrm{abs}}(T)+\mathrm{RRD}_{\mathrm{w,cum}}(T)}{T} \le \frac{\mathfrak{R}^{\mathrm{out}}_{T,\mathrm{w}}(\delta)}{T}.$$

Consequently, under the fixed-parameter convention, the conditional expected value-sensitive error is $\tilde{\mathcal{O}}(H^2d^{3/2}/\sqrt{T})$.

For every $\delta_{\mathrm{out}}\in(0,1)$, Markov's inequality conditional on $\mathcal{F}_0^{T+1}$ and the law of total probability over the training event also give

$$\mathbb{P}\left(|J^*-J(\hat{\pi},\hat{y})|+\mathrm{RRD}_{\mathrm{w}}(\hat{\pi},\hat{y};\theta^*) > \frac{\mathfrak{R}^{\mathrm{out}}_{T,\mathrm{w}}(\delta)}{T\delta_{\mathrm{out}}}\right) \le \delta+\delta_{\mathrm{out}}.$$

This corollary is the output-level counterpart of the value-sensitive RRD quantity, combined with the absolute-regret control for PROM (Algorithm 1) in Theorem E.7.

For an anytime implementation, suppose that the environment is the linear LFG in Definition 2.1, that the leader receives the feedback described in Section 4, that the supplied behavioral parameters (η,τ) are correct, and that the supplied bounds $B_\Theta, B_\vartheta, B_P$ are valid. Fix $\delta\in(0,1)$. The number of episodes need not be known in advance. Phase $j=0,1,2,\dots$ has scheduled length $T_j:=2^j$, covers global episode indices $2^j,\dots,2^{j+1}-1$, and uses confidence level $\delta_j:=6\delta/[\pi^2(j+1)^2]$. At the beginning of phase j, run a fresh copy of PROM (Algorithm 1) and recompute every confidence and covering constant with declared budget T_j and confidence input δ_j. Define the phase-start sigma-field by $\mathcal{F}_{j,0}:=\mathcal{F}_0^{2^j}$, and write $t_j(q):=2^j+q-1$ for the global index of local episode q. Let $\gamma_j^{\mathrm{MLE}}\ge 0$ be a deterministic gap fixed before phase j. Within a phase, use either exact negative-log-likelihood minimization or a solver fixed before that phase whose outputs satisfy the following certificate. Let $\ell^r_{j,h}$ be the step-h negative log-likelihood formed only from local episodes $1,\dots,r-1$ of phase j, namely

$$\ell^r_{j,h}(\theta) := -\sum_{q=1}^{r-1}\log \nu_{\theta,h}^{\pi^{t_j(q)},y^{t_j(q)}}(b_h^{t_j(q)} \mid s_h^{t_j(q)}).$$

The solver rule is selected using $\mathcal{F}_{j,0}$ and then held fixed for the phase. For every local episode $r\in[T_j]$ and step $h\in[H]$, it returns $\tilde{\theta}^r_{j,h}\in\Theta_h$ that is $\mathcal{F}_0^{t_j(r)}$-measurable and satisfies

$$\ell^r_{j,h}(\tilde{\theta}^r_{j,h}) \le \inf_{\theta\in\Theta_h}\ell^r_{j,h}(\theta)+\gamma_j^{\mathrm{MLE}}.$$

Every fitted recommendation, fitted covariance, response bonus, and value recursion in phase j is constructed from these local outputs. In the approximate case, define the phasewise likelihood confidence threshold and response confidence radius by

$$\beta_{\Theta,j} := 4\left[d\log(1+16B_\Theta\varepsilon T_j)+\log\frac{16HT_j}{\delta_j}+2\right],$$

$$R^{(j)}_{\Theta,\gamma} := \sqrt{B_\Theta^2+\frac{e^{2\varepsilon}}{c_{\mathrm{H}}(\varepsilon)}[\beta_{\Theta,j}+2\gamma_j^{\mathrm{MLE}}]}.$$

Starting from $\mathfrak{L}_{H+1}=0$, recompute the complete phasewise value class envelope by replacing T, δ, and R_Θ with T_j, δ_j, and $R^{(j)}_{\Theta,\gamma}$ throughout the backward recursion in (33)–(34). Assume that, for a fixed constant c_γ,

$$\gamma_j^{\mathrm{MLE}} \le c_\gamma\left[d\log(eT_j)+\log\frac{e(j+1)}{\delta}\right] \qquad \text{for every phase } j; \tag{62}$$

exact minimization is the special case $\gamma_j^{\mathrm{MLE}} = 0$. Use the exact statewise planning oracle in every phase. Let $\mathrm{Reg}(n), \mathrm{Reg}^{\mathrm{abs}}(n), \mathrm{RRD}_{\mathrm{cum}}(n)$, and $\mathrm{RRD}_{\mathrm{w,cum}}(n)$ denote the four criteria in (23) and (24) with the cumulative sums stopped after episode n.

Corollary E.16 (Anytime geometric-restart wrapper)**.** *Then, with probability at least $1-\delta$, simultaneously for every number $n \geq 1$ of completed episodes,*

$$\mathrm{Reg}(n), \mathrm{Reg}^{\mathrm{abs}}(n) \leq \tilde{\mathcal{O}}\Big(H^2 d^{3/2}\sqrt{n}\Big), \tag{63}$$

$$\mathrm{RRD}_{\mathrm{cum}}(n) \leq \tilde{\mathcal{O}}\big(Hd\sqrt{n}\big), \tag{64}$$

$$\mathrm{RRD}_{\mathrm{w,cum}}(n) \leq \tilde{\mathcal{O}}\big(H^2 d\sqrt{n}\big). \tag{65}$$

The additional logarithmic factors include at most a $\log\log(en)$ term from the confidence allocation. In the approximate case, the constants hidden by $\tilde{\mathcal{O}}$ may depend on the fixed constant c_γ through $R^{(j)}_{\Theta,\gamma}$. The restart scheme provides statistical guarantees without knowing the episode budget.

Corollary E.17 (Pathwise validity at random evaluation horizons)**.** *Let N be any almost surely finite, positive-integer-valued random variable on the interaction probability space, and consider the geometric-restart wrapper of Corollary E.16. There is an event of probability at least $1-\delta$ on which*

$$\mathrm{Reg}(\mathsf{N}), \mathrm{Reg}^{\mathrm{abs}}(\mathsf{N}) \leq \tilde{\mathcal{O}}\Big(H^2 d^{3/2}\sqrt{\mathsf{N}}\Big), \tag{66}$$

$$\mathrm{RRD}_{\mathrm{cum}}(\mathsf{N}) \leq \tilde{\mathcal{O}}\Big(Hd\sqrt{\mathsf{N}}\Big), \tag{67}$$

$$\mathrm{RRD}_{\mathrm{w,cum}}(\mathsf{N}) \leq \tilde{\mathcal{O}}\Big(H^2 d\sqrt{\mathsf{N}}\Big). \tag{68}$$

The hidden logarithms are the same deterministic functions of the realized value N and δ as in Corollary E.16. No optional-stopping theorem or additional confidence allocation is required: the restart event controls every deterministic episode prefix on each sample path, so one may evaluate the bounds at the realized random horizon pathwise. Almost-sure finiteness is the only probabilistic property needed for this substitution. For a future-dependent N, this is a retrospective evaluation statement rather than an implementable online stopping rule. If the episode-end filtration is $\mathcal{H}_n := \mathcal{F}_0^{n+1}$ and N is a stopping time with respect to $(\mathcal{H}_n)_{n\geq 0}$, then the random horizon is additionally implementable using episode-end information.

E.5 The loss construction under relaxed preservation

Theorem E.18 (Long-term follower-side loss under relaxed preservation)**.** *For the fixed behavioral coefficients η and τ, there is a one-state, one-step linear game with two actions for each player and feature dimension four with the following property. Define the QR probability of follower action one and the map from its recommendation probability to its ER probability by*

$$m := \frac{1}{1+e^{\eta}}, \qquad f(z) := \frac{m^{1-\kappa} z^{\kappa}}{m^{1-\kappa} z^{\kappa} + (1-m)^{1-\kappa}(1-z)^{\kappa}}, \qquad 0 \leq z \leq 1,$$

where $f(0) = 0$ and $f(1) = 1$. For every fixed $\alpha \in (0, 1]$, define the relaxed benchmark value and its gap from the QSE benchmark by

$$v_\alpha := f(\min\{m+\alpha, 1\}), \qquad c_\alpha := v_\alpha - m > 0.$$

The game has $J^*_\alpha = v_\alpha$. Every sequence of policy pairs in $\mathcal{Z}_{\theta^*}(\alpha)$ satisfies, pathwise,

$$\mathrm{CPV}(T) = \mathrm{RG}(T) \geq \big[c_\alpha T - \mathrm{Reg}_\alpha(T)\big]_+, \qquad \mathrm{CG}(T) \geq \frac{2}{\eta T}\big[c_\alpha T - \mathrm{Reg}_\alpha(T)\big]_+^2. \tag{69}$$

Moreover, the cumulative signed difference between the QR reward and the ER reward is at least $c_\alpha T - \mathrm{Reg}_\alpha(T)$. Consequently, $\mathrm{Reg}_\alpha(T) = o(T)$ implies $\mathrm{CPV}(T), \mathrm{RG}(T), \mathrm{CG}(T) = \Theta(T)$ on this game. When $0 < \alpha < 1 - m$, the supremum is attained by a unique jointly optimal leader policy and recommendation; their repeated use gives $\mathrm{CPV}(T) = \mathrm{RG}(T) = c_\alpha T$.

The example holds the follower's parameter fixed and permits the leader to optimize both components of the policy pair. It shows that a fixed relaxation of the preservation constraint does not uniformly ensure vanishing average preservation violation, even under full information. The equal-follower-reward example in Proposition G.5 is retained as a complementary illustration: behavior and the regularized criterion may change even when the reward gap is zero.

E.6 Summary of the protection guarantees

The preceding results compare follower-side effects under the estimated preservation constraint with the losses that can arise without preservation. The positive result also controls absolute leader regret against the QSE benchmark.

Table 3: Follower-side guarantees with and without preservation. The third column gives the high-probability bounds for PROM (Algorithm 1) under the estimated preservation constraint in Theorem E.1. The fourth column summarizes Corollary 6.3: in the constructed game, every leader with $o(T)$ regret against the unconstrained benchmark incurs linear follower-side losses. The same construction also covers fixed relaxations of the preservation constraint (Theorem E.18). The asymptotic summaries keep behavioral and norm parameters fixed.

Quantity	Effect measured	Under estimated preservation	Without preservation
$\mathrm{CPV}(T)$	ER deviation from QR	$\widetilde{O}(Hd\sqrt{T})$	$\Theta(T)$
$\mathrm{RG}(T)$	Absolute difference in expected reward	$\widetilde{O}(Hd\sqrt{T})$	$\Theta(T)$
$\mathrm{CG}(T)$	Regularized criterion loss	$O(Hd^2\,\mathrm{polylog}(T))$	$\Theta(T)$

Table 3 separates behavioral protection, reward protection, and preservation of the decision criterion. The negative result establishes a decrease in the follower's expected reward in this example. This contrast demonstrates the protective role of the preservation constraint.

E.7 Time-varying estimated preservation

The following extension quantifies how time-varying tolerances in the estimated constraint affect the follower-side guarantees. Consider any predictable sequence of policy pairs under the observation protocol of Section 4. Compute $\hat{\theta}_h^t$ by the exact conditional MLE in (15) from the actions generated by these pairs, using the same fixed measurable selection rule. For a sequence $\alpha_t \in [0,1]$ known before episode t, the estimated preservation constraint with tolerance α_t is given by

$$\mathrm{TV}\Big(y_h^t(\cdot \mid s), \mu^{\pi^t}_{\hat{\theta}_h^t,h}(\cdot \mid s)\Big) \le \alpha_t \quad \text{for every } h, s. \tag{70}$$

The policies and estimates remain fixed within each episode. The sequence may depend on the observed history; in that case α_t is $\mathcal{F}_0^t$-measurable.

Theorem E.19 (Preservation with time-varying tolerance)**.** *Under Definition 2.1, consider a predictable policy-pair sequence satisfying (70), with estimates computed by the conditional MLE in (15). For $T \ge 1$ and $\delta \in (0,1)$, the following bounds hold with probability at least $1-\delta$:*

$$\mathrm{RG}(T) \le \mathrm{CPV}(T) \le \sqrt{3} L_\eta R_\Theta H \mathfrak{C}_{T,d,H}(\delta) + \frac{2H}{m_0} \sum_{t=1}^{T} \alpha_t, \tag{71}$$

$$\mathrm{CG}(T) \le \frac{12 L_\eta^2}{\eta m_0} R_\Theta^2 H \mathfrak{S}_{T,d,H}(\delta) + \frac{16H}{\eta m_0^3} \sum_{t=1}^{T} \alpha_t^2, \tag{72}$$

where the uniform lower bound on QR probabilities is $m_0 := B^{-1} e^{-\eta}$, the number of follower actions is $B = |\mathcal{B}|$, and R_Θ is defined in (30) for the stated T and δ.

For PROM (Algorithm 1) with $\alpha_t = 0$, Theorem E.1 gives sharper follower-side bounds and also controls leader regret.

Consider a single learning process over episodes $t = 1, 2, \ldots$. Applying Theorem E.19 to the first T episodes with failure probability $6\delta/(\pi^2 T^2)$ and taking a union bound gives simultaneous bounds for all $T \ge 1$ with probability at least $1-\delta$. On this event, with problem parameters fixed, the condition of vanishing average tolerance

$$\frac{1}{T} \sum_{t=1}^{T} \alpha_t \longrightarrow 0 \tag{73}$$

ensures that $\mathrm{CPV}(T)/T$, $\mathrm{RG}(T)/T$, and $\mathrm{CG}(T)/T$ all tend to zero, as $T \to \infty$. This condition is weaker than requiring $\alpha_t \to 0$: a tolerance equal to one at powers of two and zero otherwise has cumulative tolerance $O(\log T)$. A schedule $\alpha_t \le c/\sqrt{t}$ preserves square-root CPV and RG bounds up to logarithmic factors and polylogarithmic CG, because $\sum_t \alpha_t = O(\sqrt{T})$ whereas $\sum_t \alpha_t^2 = O(\log T)$.

E.8 Proof roadmap

The proof combines likelihood-based optimistic learning with fitted-response transfer bounds that connect leader regret, recommendation–response distance, and preservation violation. The dependency chain is as follows.

1. A square-root likelihood-ratio martingale gives adaptive predictive Hellinger concentration; an exponential-family curvature inequality converts this into the covariance confidence set in Lemma M.3. Corollary E.5 shows how a certified numerical negative-log-likelihood error inflates the same radius.
2. Trace self-normalization yields the follower-response bonus and QR bonus in Proposition M.4: the former controls the statewise recommendation–response distance, while the latter bounds the distance between the fitted recommendation and the true QR for optimistic planning. Under estimated preservation, (98) expresses the ER as the QR at a convex combination of the true and estimated parameters and yields the preservation certificate b^{P}. RG follows from the bounded-reward characterization of TV in Proposition F.2. A predictable squared matrix-potential bound converts the same interpolation into the quadratic cumulative criterion-gap guarantee. The difference between the square-root and polylogarithmic bounds has a direct statistical interpretation. TV and reward differences are first-order in the response estimation error, whereas the criterion gap is a KL divergence and is controlled by squared confidence widths. The determinant-potential bounds accumulate these squared widths only logarithmically; the confidence radius supplies the remaining logarithmic factors.
3. A deterministic cover of every value function generated by planning, together with a self-normalized martingale inequality, proves the uniform Bellman event in Lemma O.1.
4. Lemma P.1 and backward optimism reduce regret to leader and response widths. Scalar and matrix determinant potentials sum these widths.
5. A follower-policy performance-difference identity yields two absolute-regret routes: the global inverse map in Proposition C.3 gives a positivity-dependent model-level bridge for recommendation sequences satisfying a common positive coordinate lower bound, whereas the interpolation identity and b^{P} give the sharper algorithm-specific bridge for PROM (Algorithm 1) in Theorem E.7.
6. The negative construction links the leader's gain from steering to the follower's reward loss, yielding the bounds in Theorem E.18.
7. For time-varying estimated preservation, likelihood concentration and the global response bound in Proposition 3.5 control the contributions of α_t and α_t^2 in Theorem E.19.

The behavioral-misspecification extension is given separately in Section R.

F Proofs of the Structural and Main Results

F.1 Structural results and benchmark characterization

Proof of Proposition 3.2. Fix h, s, π, θ_h, and abbreviate the expected reward, recommendation probability, and QR probability as $r_b := r^{\pi}_{\theta_h,h}(s,b)$, $y_b := y_h(b \mid s)$, and $\mu_b := \mu^{\pi}_{\theta_h,h}(b \mid s)$. Since $\eta(1-\kappa) = \varepsilon$, the unnormalized weights in (4) satisfy

$$y_b^{\kappa} e^{\varepsilon r_b} \propto y_b^{\kappa} \mu_b^{1-\kappa},$$

which proves (5) after normalization. Expanding the objective in (6) and discarding terms independent of q gives

$$\sum_{b \in \mathcal{B}} q(b) \log q(b) - \sum_{b \in \mathcal{B}} q(b) \log\left(\mu_b^{1-\kappa} y_b^{\kappa}\right).$$

Its unique minimizer is the normalized geometric mean in (5). Taking pairwise log odds in that display proves (74). For every $b, b' \in \mathcal{B}$, the response log odds satisfy

$$\log \frac{\nu^{\pi,y}_{\theta_h,h}(b \mid s)}{\nu^{\pi,y}_{\theta_h,h}(b' \mid s)} = (1-\kappa) \log \frac{\mu^{\pi}_{\theta_h,h}(b \mid s)}{\mu^{\pi}_{\theta_h,h}(b' \mid s)} + \kappa \log \frac{y_h(b \mid s)}{y_h(b' \mid s)}. \tag{74}$$

□

Proof of Proposition 3.6. Fix h, s, π, θ_h, abbreviate the expected reward as $r_b := r^{\pi}_{\theta_h,h}(s,b)$, and let the QR normalizing constant be $Z := \sum_{b \in \mathcal{B}} e^{\eta r_b}$. For every $q \in \Delta(\mathcal{B})$,

$$F^{\pi}_{\theta_h,h}(s,q) = \sum_b q(b) r_b - \eta^{-1} \sum_b q(b) \log q(b) = \eta^{-1} \log Z - \eta^{-1} \mathrm{KL}\big(q \| \mu^{\pi}_{\theta_h,h}(\cdot \mid s)\big).$$

Substituting first $q = \mu$ and then $q = \nu$ proves the identity in (7). Pinsker's inequality gives its total-variation lower bound. The equality condition for KL divergence proves the final claim. □

Proof of Proposition C.1. Fix h, s, π, y, and θ_h, and abbreviate the combined entropy coefficient as $c := \eta^{-1} + \tau$. The closed-form response implies that, for a normalizing constant $Z_y > 0$,

$$\log \nu^{\pi,y}_{\theta_h,h}(b \mid s) = c^{-1} r^{\pi}_{\theta_h,h}(s,b) + \frac{\tau}{c} \log y_h(b \mid s) - \log Z_y.$$

Substitution into the definition of KL divergence gives, for every $q \in \Delta(\mathcal{B})$,

$$G^{\pi,y}_{\theta_h,h}(s,q) = c \log Z_y - c\mathrm{KL}\Big(q \| \nu^{\pi,y}_{\theta_h,h}(\cdot \mid s)\Big).$$

The unique maximizer is therefore $\nu^{\pi,y}_{\theta_h,h}(\cdot \mid s)$. Evaluating this identity at the maximizer and at $q = y_h(\cdot \mid s)$ proves (25). □

Proof of Theorem 3.3. Fix h, s, π, θ_h and abbreviate the three policies by y, μ, ν. The geometric formula gives $y = \mu \Longrightarrow \nu = \mu = y$. Conversely, if $y = \nu$, then for every pair of actions b, c its log-odds form implies

$$\log \frac{y(b)}{y(c)} = (1-\kappa) \log \frac{\mu(b)}{\mu(c)} + \kappa \log \frac{y(b)}{y(c)}.$$

Since $1 - \kappa > 0$, the ratios of y and μ agree, and normalization gives $y = \mu$. If instead $\nu = \mu$, the same formula yields

$$\log \frac{\mu(b)}{\mu(c)} = (1-\kappa) \log \frac{\mu(b)}{\mu(c)} + \kappa \log \frac{y(b)}{y(c)}.$$

Since $\kappa > 0$ under the standing model, normalization again gives $y = \mu$. All ratios are well defined on the full-support domain. This proves both equivalences and uniqueness of the fixed point. □

Remark F.1 (Why preservation must be learned). Fix a context (h, s, p). If two admissible parameters θ^0_h and θ^1_h induce distinct QRs, then no single recommendation preserves the QR under both parameters. Indeed, Theorem 3.3 would otherwise require the same recommendation to equal both $\mu^{p}_{\theta^0_h,h}(\cdot \mid s)$ and $\mu^{p}_{\theta^1_h,h}(\cdot \mid s)$. Thus an exact preservation rule cannot in general be selected without learning the response model. This motivates the estimated preservation constraint and cumulative guarantees evaluated under the true parameter.

Proof of Lemma C.2. Fix h, s, π, θ_h, and write the expected reward as $q_b := r^{\pi}_{\theta_h,h}(s,b) \in [0,1]$. For the quantal response,

$$\mu^{\pi}_{\theta_h,h}(b \mid s) = \frac{e^{\eta q_b}}{\sum_{c \in \mathcal{B}} e^{\eta q_c}}.$$

The numerator is at least one, while the denominator is at most Be^{η}. Therefore $\mu^{\pi}_{\theta_h,h}(b \mid s) \geq B^{-1} e^{-\eta} = m_0$. If $y_c \geq \underline{y}$ for every $c \in \mathcal{B}$, then

$$\nu^{\pi,y}_{\theta_h,h}(b \mid s) = \frac{y_b^{\kappa} e^{\varepsilon q_b}}{\sum_{c \in \mathcal{B}} y_c^{\kappa} e^{\varepsilon q_c}}.$$

The numerator is at least $\underline{y}^{\kappa}$ because $q_b \geq 0$, while the denominator is at most Be^{ε} because $y_c^{\kappa} \leq 1$ and $q_c \leq 1$. Consequently, $\nu^{\pi,y}_{\theta_h,h}(b \mid s) \geq B^{-1} \underline{y}^{\kappa} e^{-\varepsilon} = \underline{n}(\underline{y})$. □

Proof of Proposition C.3. The odds ratios in (4) and (2) give the inverse map Ψ_y from the ER to the QR:

$$\mu_b = \Psi_y(\nu)_b := \frac{\nu_b^{\lambda_{\text{inv}}} y_b^{1-\lambda_{\text{inv}}}}{\sum_{c\in\mathcal{B}} \nu_c^{\lambda_{\text{inv}}} y_c^{1-\lambda_{\text{inv}}}}.$$

Also $\Psi_y(y) = y$. Differentiation gives

$$\frac{\partial \Psi_y(z)_b}{\partial z_c} = \frac{\lambda_{\text{inv}} \Psi_y(z)_b}{z_c} \big(\mathbf{1}\{b = c\} - \Psi_y(z)_c\big).$$

Every point on the segment between y and ν has coordinates at least $\underline{p}$. The induced ℓ_1 operator norm satisfies

$$\|D\Psi_y(z)\|_{1\to 1} = \max_{c\in\mathcal{B}} \frac{2\lambda_{\text{inv}} \Psi_y(z)_c(1 - \Psi_y(z)_c)}{z_c} \le \frac{\lambda_{\text{inv}}}{2\underline{p}}.$$

The mean-value theorem yields $\|\mu - y\|_1 \le \lambda_{\text{inv}}(2\underline{p})^{-1} \|\nu - y\|_1$, proving (26) after dividing both sides by two and using $\text{TV}(p, q) = \|p - q\|_1 /2$. Finally, the triangle inequality gives

$$\text{TV}(\mu, \nu) \le \text{TV}(\mu, y) + \text{TV}(y, \nu) \le \left(1 + \frac{\lambda_{\text{inv}}}{2\underline{p}}\right) \text{TV}(y, \nu),$$

which is (27). □

Leader values and the benchmark. For a leader policy π and a follower policy ν, define the leader's action-value function U_h and value function W_h for $h \in [H]$ as

$$U_h^{\pi,\nu}(s, a, b) := u_h(s, a, b) + \sum_{s'\in\mathcal{S}} P_h(s' \mid s, a, b) W_{h+1}^{\pi,\nu}(s'),$$
$$W_h^{\pi,\nu}(s) := \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \pi_h(a \mid s)\nu_h(b \mid s) U_h^{\pi,\nu}(s, a, b), \qquad W_{H+1}^{\pi,\nu} \equiv 0.$$

$U_h^{\pi,\nu}(s, a, b)$ adds the current reward to the expected continuation value after the joint action (a, b). $W_h^{\pi,\nu}(s)$ is the corresponding value function obtained by averaging $U_h^{\pi,\nu}$ under the leader and follower policies at state s. The terminal condition $W_{H+1}^{\pi,\nu} = 0$ closes the finite-horizon Bellman recursion.

For bounded f, the transition-operator series converges absolutely because $\sum_{s\in\mathcal{S}} P_h(s \mid s_0, a, b)|f(s)| \le \|f\|_\infty$.

For concise Bellman expressions, for any function $f : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, write the expectation operator under the joint policy (π_h, ν_h) as

$$(\pi_h \otimes \nu_h) \circ f(s) := \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \pi_h(a \mid s)\nu_h(b \mid s) f(s, a, b).$$

The leader's full-information problem is to maximize her expected return under the preservation constraint imposed by an external regulator:

$$(\pi^*, y^*) \in \operatorname*{argmax}_{(\pi,y)\in\mathcal{Z}_{\theta^*}} J(\pi, y).$$

Proof of Lemma 3.8. We prove the claim by backward induction. The preservation constraint implies that, after the leader chooses $p \in \Delta(\mathcal{A})$ at (h, s), the executed follower distribution is $\mu^p_{\theta^*_h,h}(\cdot \mid s)$. The preservation constraint requires this same recommendation.

At step H, there is no continuation reward. Hence the best preservation value from state s is

$$\max_{p\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} p(a)\mu^p_{\theta^*_H,H}(b \mid s) u_H(s, a, b),$$

which agrees with (9)–(10) because $W^*_{H+1} \equiv 0$.

Now suppose that $W^*_{h+1}(s')$ is the optimal preservation continuation value from every next state s'. If the leader chooses p at (h, s) and then follows an optimal Markov continuation policy, the expected current-plus-future return is

$$\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a)\mu^{p}_{\theta^*_h,h}(b \mid s)\left[u_h(s,a,b) + (P_h W^*_{h+1})(s,a,b)\right].$$

The follower is myopic, so her step-h response depends only on the current state and the current leader mixture; it does not couple the present optimization to future recommendations. Maximizing the preceding display over p therefore gives exactly (9)–(10). For every state, the objective is continuous in p and $\Delta(\mathcal{A})$ is compact, so the maximum is attained. A fixed tie-breaking rule selects one maximizer at each state, and the resulting selector is measurable because $\mathcal{S}$ is countable and carries its discrete sigma-field. Backward induction therefore constructs a Markov policy pair that satisfies the preservation constraint and attains W^*_h at every step. Averaging the initial value under $s_1 \sim \rho_0$ proves $J^* = \mathbb{E}_{s_1\sim\rho_0}[W^*_1(s_1)]$. □

The upper bound connects the recommendation that the leader can choose to the preservation violation that the regulator evaluates. It holds even when the recommendation approaches the boundary of the simplex. The reverse bound is uniform because the QR, rather than the recommendation, has a positive lower bound. As Proposition G.4 shows, the recommendation and ER can be arbitrarily close while remaining separated from the QR.

Proof of Proposition 3.5. Let the recommendation-to-QR probability ratio be $x_b := y(b)/\mu(b) > 0$, and define a path of probability vectors and its normalizing constant by

$$p_v(b) := \frac{\mu(b)x_b^v}{Z(v)}, \qquad Z(v) := \sum_{c\in\mathcal{B}} \mu(c)x_c^v, \qquad 0 \le v \le 1.$$

The geometric identity in Proposition 3.2 gives $p_0 = \mu$, $p_\kappa = \nu$, and $p_1 = y$. Order the actions by nondecreasing x_b and let A be an upper tail in this order, including the empty and full tails. Differentiating the finite sum yields

$$\frac{\mathrm{d}}{\mathrm{d}v}p_v(A) = \mathrm{Cov}_{p_v}(\mathbf{1}_A, \log x) = \sum_{b\in A}\sum_{c\notin A} p_v(b)p_v(c)(\log x_b - \log x_c) \ge 0.$$

Thus the mass of every upper tail is nondecreasing along the path. For $v > 0$, the ratio $p_v(b)/\mu(b) = x_b^v/Z(v)$ is nondecreasing in the same order, so the set on which $p_v(b) > \mu(b)$ is an upper tail. The finite-space characterization of total variation therefore gives

$$\mathrm{TV}(p_v, \mu) = \max_{A \text{ an upper tail}} \{p_v(A) - \mu(A)\}.$$

Every expression in this maximum is nondecreasing in v, including at $v = 0$ by continuity. Consequently, $\mathrm{TV}(\nu, \mu) \le \mathrm{TV}(y, \mu)$.

For the reverse bound, define the inverse map Ψ_μ from the ER to the recommendation using the geometric identity:

$$y(b) = \Psi_\mu(\nu)(b), \qquad \Psi_\mu(z)(b) := \frac{z(b)^{1/\kappa}\mu(b)^{-(1-\kappa)/\kappa}}{\sum_c z(c)^{1/\kappa}\mu(c)^{-(1-\kappa)/\kappa}}.$$

In particular, $\Psi_\mu(\mu) = \mu$. For a normalized power map $G_i(z) = c_i z_i^a / \sum_j c_j z_j^a$, with $c_i > 0$ and $a > 0$, direct differentiation gives

$$\frac{\partial G_i}{\partial z_j} = aG_i\left(\frac{\mathbf{1}_{\{i=j\}}}{z_i} - \frac{G_j}{z_j}\right), \qquad \sum_i \left|\frac{\partial G_i}{\partial z_j}\right| = \frac{2aG_j(1-G_j)}{z_j} \le \frac{a}{2z_j}.$$

Suppose first that $\mathrm{TV}(\nu, \mu) \le m/2$. For probability vectors, each coordinate difference is at most their total variation, so every coordinate on the segment between ν and μ is at least $m/2$. The induced ℓ_1 operator norm of the Jacobian of Ψ_μ is therefore at most $1/(\kappa m)$ on this segment. Integrating the Jacobian along the segment and dividing the resulting ℓ_1 inequality by two gives

$$\mathrm{TV}(y, \mu) = \mathrm{TV}(\Psi_\mu(\nu), \Psi_\mu(\mu)) \le \frac{1}{\kappa m}\,\mathrm{TV}(\nu, \mu).$$

If instead $\mathrm{TV}(\nu, \mu) > m/2$, then

$$\mathrm{TV}(y, \mu) \le 1 < \frac{2}{m}\,\mathrm{TV}(\nu, \mu).$$

Combining the two cases proves the reverse bound with C_m. Finally, the closed-form QR and rewards in $[0, 1]$ give $\mu(b) \ge B^{-1}e^{-\eta}$ for every action, which proves the asserted uniformity. □

Proposition F.2 (Reward protection from preservation)**.** *For every admissible policy-pair sequence,*

$$0 \le \mathrm{RG}(T) \le \mathrm{CPV}(T). \tag{75}$$

Proof of Proposition F.2. Fix a state, step, and leader commitment, and write the expected follower reward as $g(b) := r^{\pi}_{\theta,h}(s, b) \in [0, 1]$. For probability vectors p, q, the total mass of each of the positive and negative parts of $p - q$ equals $\mathrm{TV}(p, q)$. Therefore

$$-\mathrm{TV}(p, q) \le \sum_b (p(b) - q(b))g(b) \le \mathrm{TV}(p, q).$$

Apply this inequality with p equal to the QR and q equal to the ER, take absolute values, then take the same episode-start conditional expectations used in (13). Summing over episodes and steps proves (75). Since the positive part of a signed reward loss is bounded by its absolute value, the stated one-sided interpretation also follows. □

F.2 Learning and protection guarantees

Proof of Theorem E.1. Use the intersection of the MLE, Bellman-confidence, matrix-potential, and squared-matrix-potential events constructed in the proof of Theorem E.2. Its probability is at least $1 - \delta$. On this same event, Theorem E.7 gives (37), and Corollary E.3 gives the CPV inequality in (38). The pathwise bound in Proposition F.2 supplies RG without another probability event. Finally, Corollary E.4 gives (39) on the same intersection. For $\mathrm{CG}(T)$ in Theorem 6.1, the definitions in (30) imply $R^2_\Theta \le C_1 d \log(eTH/\delta)$ for fixed behavioral coefficients and norm bounds, while (36) implies $\mathfrak{S}_{T,d,H}(\delta) \le C_2 d \log(eTH/\delta)$. Substitution in (39) proves the stated squared-logarithmic bound. The leader and CPV dimension summaries follow from the same estimates and the Bellman-radius calculation in Corollary E.6. □

Proof of Theorem 6.1. Apply the dimension estimates in Corollary E.6 to the bounds in Theorem E.1. The estimates for R^2_Θ and $\mathfrak{S}_{T,d,H}(\delta)$ in the proof of Theorem E.1 give the stated criterion-gap bound on the same event. □

Proof of Theorem E.18. Take $H = 1$, $\mathcal{S} = \{s\}$, and $\mathcal{A} = \mathcal{B} = \{1, 2\}$. Define the follower and leader rewards by

$$r(s, a, 1) = 0, \quad r(s, a, 2) = 1 \quad (a = 1, 2),$$
$$u(s, 1, 1) = 1, \quad u(s, 2, 1) = \tfrac{1}{2}, \quad u(s, 1, 2) = u(s, 2, 2) = 0.$$

Set the terminal transition probability $P_1(s \mid s, a, b) = 1$; this terminal transition does not affect either return. To verify membership in Definition 2.1, order the four coordinates as $(1, 1), (1, 2), (2, 1), (2, 2)$ and specify the feature map, follower-reward vector, leader-reward vector, and transition measure by

$$\phi(s, a, b) = e_{(a,b)}, \quad \theta^* = (0, 1, 0, 1)^\top, \quad \vartheta^* = (1, 0, \tfrac{1}{2}, 0)^\top, \quad \mathbf{m}^*_1(\{s\}) = (1, 1, 1, 1)^\top.$$

The feature norms are one, all rewards lie in $[0, 1]$, and one may take $\Theta_1 = [0, 1]^4$, $B_\Theta = 2$, $B_\vartheta = 2$, and $B_P = 2$. Thus the construction uses fixed finite model constants and the same feature representation for the two rewards and the transition.

The follower reward is independent of the leader action, so the QR is $\mu = (m, 1 - m)$ under every leader policy. Write the recommendation as $y = (z, 1 - z)$ with $0 < z < 1$. By Proposition 3.2, the ER assigns probability $q = f(z)$ to follower action one. Its log odds satisfy

$$\log \frac{f(z)}{1 - f(z)} = (1 - \kappa) \log \frac{m}{1 - m} + \kappa \log \frac{z}{1 - z}.$$

Differentiating gives

$$f'(z) = \frac{\kappa f(z)(1-f(z))}{z(1-z)} > 0 \qquad (0 < z < 1),$$

so f is strictly increasing and $f(m) = m$. If the leader assigns probability p to action two, her return is

$$J(\pi, y) = (1 - p/2)f(z) \le f(z).$$

Since $f(z) > 0$, the maximizer over $p \in [0, 1]$ is uniquely $p = 0$. The tolerance in (22) is exactly $|z - m| \le \alpha$. Strict monotonicity of f therefore gives $J^*_\alpha = v_\alpha$ and $c_\alpha > 0$. If $\alpha < 1 - m$, the maximum is attained uniquely at $p = 0$ and $z = m + \alpha$. If $\alpha \ge 1 - m$, admissible z may approach one, giving supremum one, but full support precludes $z = 1$. This proves all optimization claims.

For any feasible, possibly history-dependent sequence, let the ER probability of action one be $q_t := \nu^{\pi^t, y^t}_{\theta^*}(1)$ and the leader return be $J_t := J(\pi^t, y^t)$. The inequality $J_t \le q_t$ yields

$$\sum_{t=1}^{T}(q_t - m) \ge \sum_{t=1}^{T}(J_t - m) = (v_\alpha - m)T - \sum_{t=1}^{T}(v_\alpha - J_t) = c_\alpha T - \mathrm{Reg}_\alpha(T).$$

The QR and ER rewards are $1 - m$ and $1 - q_t$, respectively, so $q_t - m$ is exactly the signed follower reward loss caused by the recommendation in episode t. Since there is one state and one step, the definitions of CPV and RG both reduce to $\sum_t |q_t - m|$. The triangle inequality and their nonnegativity prove the first inequality in (69). The same signed-loss bound also lower-bounds the cumulative positive part of the follower's reward loss.

By Proposition 3.6 and Pinsker's inequality,

$$\mathrm{CG}(T) = \frac{1}{\eta}\sum_{t=1}^{T}\mathrm{KL}\big((q_t, 1-q_t)\|(m, 1-m)\big) \ge \frac{2}{\eta}\sum_{t=1}^{T}(q_t - m)^2.$$

Cauchy–Schwarz implies

$$\sum_{t=1}^{T}(q_t - m)^2 \ge \frac{1}{T}\left(\sum_{t=1}^{T}(q_t - m)\right)^2.$$

If $c_\alpha T - \mathrm{Reg}_\alpha(T) > 0$, substitute the preceding lower bound on the signed sum. If it is nonpositive, nonnegativity of CG suffices. These two cases give the second inequality in (69). When $\mathrm{Reg}_\alpha(T) = o(T)$, the lower bounds are linear because $c_\alpha > 0$ is fixed. The matching elementary upper bounds are $\mathrm{CPV}(T) = \mathrm{RG}(T) \le T$ and $\mathrm{CG}(T) \le T\eta^{-1}\log(1/m)$, since $m < 1/2$ and $\mathrm{KL}(q\|\mu) \le \log(1/m)$ for any policy q. They prove the asserted $\Theta(T)$ growth. For repeated exact optimizers with $\alpha < 1-m$, $q_t = v_\alpha$ in every episode, which proves the stated equalities. □

Proof of Theorem 6.2. Use the game constructed in Theorem E.18, which has $J^*_1 = 1$. The displayed inequalities and signed-loss bound are its conclusions. For fixed $\alpha > 0$, sublinear $\mathrm{Reg}_\alpha(T)$ gives linear lower bounds; bounded rewards and the positive QR probabilities give linear upper bounds. □

Proof of Corollary 6.3. Every two probability vectors have total variation at most one, so the tolerance-one condition imposes no additional restriction on an admissible recommendation. The preceding construction has $v_1 = f(1) = 1$ and $c_1 = 1 - m > 0$. Apply Theorem E.18 with $\alpha = 1$. The recommendations with first coordinate tending to one show directly that the supremum is approachable within the full-support class. □

Proof of Theorem E.19. For the actual recommendation in episode t, define the fitted ER and its feature covariance by

$$\hat{\nu}^t_h(\cdot \mid s) := \nu^{\pi^t, y^t}_{\hat{\theta}^t_h, h}(\cdot \mid s), \qquad \widehat{\Sigma}^t_h(s) := \mathrm{Cov}_{b \sim \hat{\nu}^t_h(\cdot|s)}[\phi^{\pi^t}_h(s, b)].$$

In this proof, use the historical covariance matrix formed from the frozen fitted covariances and the current width

$$V^t_h := I_d + \sum_{i<t}\widehat{\Sigma}^i_h(s^i_h), \qquad w^t_h(s) := \sqrt{\mathrm{Tr}((V^t_h)^{-1}\widehat{\Sigma}^t_h(s))}.$$

The covariance here is taken under the fitted ER for the actual recommendation, not under the fitted QR. This distinction is necessary when the tolerance is positive. Each summand is stored using the estimate from its own episode, so

$$V_h^{t+1} = V_h^t + \widehat{\Sigma}_h^t(s_h^t), \qquad \widehat{\Sigma}_h^t(s_h^t) \succeq 0, \qquad \mathrm{Tr}(\widehat{\Sigma}_h^t(s_h^t)) \le 1.$$

The last inequality follows from $\|\phi_h^{\pi^t}(s,b)\|_2 \le 1$. In particular, $0 \le w_h^t(s_h^t) \le 1$.

We first check the statistical interfaces for this interaction law. In the proof of Lemma M.2, the recommendation enters only as a predictable, positive offset in the conditional likelihood. Neither its likelihood-ratio martingale nor its parameter-net extension uses equality of the recommendation and the fitted QR. The proof of Lemma M.3 compares reward parameters at the same recommendation, so the bounded density ratios and Hellinger-to-covariance bound also remain valid. Applying those proofs to the present histories gives, with failure probability at most $\delta/4$,

$$\|\hat{\theta}_h^t - \theta_h^*\|_{V_h^t} \le R_\Theta \quad \text{for every } t \le T,\ h \le H.$$

This establishes the confidence event for the present interaction histories.

For completeness, the corresponding potential bounds follow from the displayed additive update. With the normalized covariance matrix $A_h^t := (V_h^t)^{-1/2}\widehat{\Sigma}_h^t(s_h^t)(V_h^t)^{-1/2}$, all eigenvalues lie in $[0,1]$, and hence

$$\sum_{t=1}^T w_h^t(s_h^t)^2 = \sum_{t=1}^T \mathrm{Tr}(A_h^t) \le 2\sum_{t=1}^T \log\det(I_d + A_h^t) = 2\log\det V_h^{T+1} \le 2d\log(1+T/d).$$

The last inequality follows from $\mathrm{Tr}(V_h^{T+1}) \le d+T$ and the arithmetic-geometric mean inequality for its eigenvalues. Cauchy–Schwarz and Lemma P.3, with failure probability $\delta/(16H)$ for each h, give

$$\sum_{t=1}^T \mathbb{E}^t w_h^t(s_h^t) \le \mathfrak{C}_{T,d,H}(\delta).$$

To control squared widths without adding a square-root term, set the squared response width $X_t = w_h^t(s_h^t)^2$ and its conditional mean $m_t = \mathbb{E}^t X_t$. The chord bound $e^{-x} \le 1-(1-e^{-1})x$ for $x \in [0,1]$ implies that

$$\exp\left((1-e^{-1})\sum_{i\le t} m_i - \sum_{i\le t} X_i\right)$$

is a nonnegative supermartingale. Markov's inequality, followed by $(1-e^{-1})^{-1} \le 2$, gives

$$\sum_{t=1}^T \mathbb{E}^t X_t \le 2\sum_{t=1}^T X_t + 2\log(4H/\delta) \le \mathfrak{S}_{T,d,H}(\delta)$$

with failure probability at most $\delta/(4H)$. A union bound over h and the likelihood event has failure probability at most $\delta/4 + \delta/16 + \delta/4 < \delta$. Work on their intersection for the remainder of the proof.

Fix t, h, s and abbreviate the estimated QR, true QR, recommendation, and ER as $\hat{\mu} = \mu^{\pi^t}_{\hat{\theta}_h^t,h}(\cdot \mid s)$, $\mu = \mu^{\pi^t}_{\theta_h^*,h}(\cdot \mid s)$, $y = y_h^t(\cdot \mid s)$, and $\nu = \nu^{\pi^t,y^t}_{\theta_h^*,h}(\cdot \mid s)$. Every coordinate of $\hat{\mu}$ and μ is at least m_0. Suppose first that $\alpha_t \le m_0/2$. Since a coordinate difference is bounded by total variation, (70) gives

$$\frac{1}{2} \le \frac{y(b)}{\hat{\mu}(b)} \le \frac{3}{2} \qquad \text{for every } b.$$

Using the geometric response identity with parameter $\hat{\theta}_h^t$ yields

$$\frac{\hat{\nu}_h^t(b \mid s)}{\hat{\mu}(b)} = \frac{(y(b)/\hat{\mu}(b))^\kappa}{\sum_c \hat{\mu}(c)(y(c)/\hat{\mu}(c))^\kappa} \ge 3^{-\kappa} \ge \frac{1}{3}.$$

For every real function g on $\mathcal{B}$, the variational characterization of variance therefore gives

$$\mathrm{Var}_{\hat{\mu}}(g) = \min_a \mathbb{E}_{\hat{\mu}}(g-a)^2 \le \mathbb{E}_{\hat{\mu}}(g - \mathbb{E}_{\hat{\nu}} g)^2 \le 3\,\mathrm{Var}_{\hat{\nu}}(g).$$

Consequently, the covariance under $\hat{\mu}$ is bounded in Loewner order by $3\widehat{\Sigma}_h^t(s)$. Apply Lemma L.5 to the two QRs, followed by trace self-normalization and the confidence event, to obtain

$$\mathrm{TV}(\hat{\mu}, \mu) \le L_\eta \|\hat{\theta}_h^t - \theta_h^*\|_{\mathrm{Cov}_{\hat{\mu}}\, \phi_h^{\pi^t}} \le \sqrt{3} L_\eta R_\Theta w_h^t(s).$$

By Proposition 3.5 and the triangle inequality,

$$\mathrm{TV}(\nu, \mu) \le \mathrm{TV}(y, \mu) \le \alpha_t + \sqrt{3} L_\eta R_\Theta w_h^t(s).$$

If $\alpha_t > m_0/2$, use instead $\mathrm{TV}(\nu, \mu) \le 1 < 2\alpha_t/m_0$. The two cases imply the uniform bound

$$\mathrm{PV}_h(s, \pi^t, y^t; \theta_h^*) \le \sqrt{3} L_\eta R_\Theta w_h^t(s) + \frac{2}{m_0}\alpha_t. \tag{76}$$

Taking episode-start conditional expectations and summing over t, h proves (71), including RG by Proposition F.2.

For probability vectors q, p with $p(b) \ge m_0$, the inequality $\log x \le x - 1$ gives

$$\mathrm{KL}(q\|p) \le \sum_b \frac{(q(b) - p(b))^2}{p(b)} \le \frac{1}{m_0} \sum_b (q(b) - p(b))^2 \le \frac{2}{m_0} \mathrm{TV}(q, p)^2.$$

For the last step, the positive and negative parts of $q - p$ each have total mass $\mathrm{TV}(q, p)$, and the sum of squares of each part is at most the square of that mass. By Proposition 3.6, (76), and $(a+b)^2 \le 2a^2 + 2b^2$,

$$F^{\pi^t}_{\theta_h^*, h}(s, \mu) - F^{\pi^t}_{\theta_h^*, h}(s, \nu) \le \frac{12 L_\eta^2}{\eta m_0} R_\Theta^2 w_h^t(s)^2 + \frac{16}{\eta m_0^3} \alpha_t^2.$$

Summing the conditional expectations and using the squared-potential bound proves (72). For one infinite predictable sequence, the stated failure probabilities sum to δ over all T, which justifies the simultaneous-horizon conclusion. Finally, $0 \le \alpha_t^2 \le \alpha_t$ implies that (73) suffices for all three average quantities to vanish on this common event. The two example schedules follow from $\sum_{t \le T} t^{-1/2} \le 2\sqrt{T}$ and $\sum_{t \le T} t^{-1} \le 1 + \log T$. □

G Structural Complements and Scope

This section collects consequences that are useful for interpreting the main results and for distinguishing them from QSE learning without policy recommendation.

The MLE problem is convex over Θ_h. The MLE can be nonunique because reward shifts or other centered-feature directions unexcited by the observed contexts can induce the same ER at those contexts. The analysis therefore selects a measurable MLE representative and controls its error only in the response-relevant covariance seminorms needed for prediction and regret. Lemma G.1 and corollary G.2 separately characterize observational equivalence and the exact structural condition for unique parameter identification over Θ_h.

Lemma G.1 (Observational equivalence of response parameters)**.** *Fix $h \in [H]$, $s \in \mathcal{S}$, $p \in \Delta(\mathcal{A})$, and a full-support recommendation y. For any $\theta, \theta' \in \Theta_h$, the ERs $\nu^{p,y}_{\theta,h}(\cdot \mid s)$ and $\nu^{p,y}_{\theta',h}(\cdot \mid s)$ are equal if and only if*

$$\langle \phi_h^p(s, b) - \phi_h^p(s, c), \theta - \theta' \rangle = 0 \qquad \text{for every } b, c \in \mathcal{B}.$$

For a full-support distribution $q \in \mathrm{int}\,\Delta(\mathcal{B})$, define the feature covariance matrix $\Sigma_{h,s,p}(q) := \mathrm{Cov}_{b \sim q}[\phi_h^p(s, b)]$. Equivalently, $\theta - \theta' \in \ker \Sigma_{h,s,p}(q)$ for one, and hence for every, full-support distribution q. The same characterization holds for the quantal responses $\mu^p_{\theta,h}(\cdot \mid s)$ and $\mu^p_{\theta',h}(\cdot \mid s)$. For a finite collection of contexts $\{(s_j, p_j, y_j)\}_{j=1}^m$ at the same step h, with every y_j having full support, observational equivalence at every context is equivalent to

$$\theta - \theta' \in \ker\left(\sum_{j=1}^m \Sigma_{h,s_j,p_j}(q_j)\right) = \bigcap_{j=1}^m \ker \Sigma_{h,s_j,p_j}(q_j),$$

where every q_j may be any full-support distribution on $\mathcal{B}$.

Proof of Lemma G.1. For every $b, c \in \mathcal{B}$, the conditional log-odds identity gives

$$\log \frac{\nu^{p,y}_{\theta,h}(b \mid s)}{\nu^{p,y}_{\theta,h}(c \mid s)} - \kappa \log \frac{y(b)}{y(c)} = \varepsilon \langle \phi^p_h(s,b) - \phi^p_h(s,c), \theta \rangle.$$

Because y and both ERs have full support and $\varepsilon > 0$, equality of the ERs is equivalent to equality of all displayed feature contrasts. For any full-support q and any vector v, the identity $v^\top \Sigma_{h,s,p}(q) v = \mathrm{Var}_{b \sim q}(v^\top \phi^p_h(s,b))$ shows that $v \in \ker \Sigma_{h,s,p}(q)$ if and only if $v^\top \phi^p_h(s,b)$ is constant in b. This proves both the covariance characterization and its independence from the particular full-support distribution q. For the quantal response, (2) gives, for every $b, c \in \mathcal{B}$,

$$\log \frac{\mu^p_{\theta,h}(b \mid s)}{\mu^p_{\theta,h}(c \mid s)} = \eta \langle \phi^p_h(s,b) - \phi^p_h(s,c), \theta \rangle.$$

Since $\eta > 0$ and every quantal response has full support, $\mu^p_{\theta,h}(\cdot \mid s) = \mu^p_{\theta',h}(\cdot \mid s)$ if and only if

$$\langle \phi^p_h(s,b) - \phi^p_h(s,c), \theta - \theta' \rangle = 0 \qquad \text{for every } b, c \in \mathcal{B}.$$

Thus the same covariance-kernel characterization holds for quantal responses. Finally, for positive semidefinite matrices $\{A_j\}_{j=1}^m$, $\ker(\sum_{j=1}^m A_j) = \cap_{j=1}^m \ker(A_j)$, which proves the aggregate statement. In particular, adding a contextwise common scalar to every follower-action reward is one, but not necessarily the only, observationally null perturbation. □

Corollary G.2 (Unique response-parameter identification over Θ_h)**.** *Fix $h \in [H]$ and a finite collection of contexts $\{(s_j, p_j, y_j)\}_{j=1}^m$, where $m \in \mathbb{N}$ and every y_j has full support. These contexts uniquely identify the response parameter over Θ_h if and only if, for one, equivalently every, collection of full-support weighting distributions $\{q_j\}_{j=1}^m$,*

$$\ker\left(\sum_{j=1}^m \Sigma_{h,s_j,p_j}(q_j)\right) \cap (\Theta_h - \Theta_h) = \{0\}. \tag{77}$$

Proof of Corollary G.2. By Lemma G.1, two parameters in Θ_h induce the same ER at all listed contexts if and only if their difference belongs to the kernel in (77). Restricting this difference to $\Theta_h - \Theta_h$ proves the equivalence. The same lemma shows that the kernel is independent of the particular full-support weighting distributions. □

In particular, positive definiteness of the aggregate centered covariance is sufficient for unique identification. This is a structural identifiability statement only: the regret theorems neither assume that the algorithm visits or sufficiently excites these contexts nor require unique Euclidean recovery.

Proposition G.3 (Markov sufficiency for the preservation benchmark)**.** *Enlarge the comparison class by allowing the leader, at each finite history ending in (h, s), to choose a leader-action distribution and a full-support recommendation as measurable functions of that history. Impose preservation at every such history. Under a myopic follower, this enlargement does not improve the optimal value: an optimal pair can be chosen Markov. Its recommendation satisfies*

$$y^*_h(\cdot \mid s) = \mu^{\pi^*}_{\theta^*_h,h}(\cdot \mid s),$$

and its value is the Bellman value in Lemma 3.8.

Proof of Proposition G.3. Let $\mathfrak{h}_h = (s_1, a_1, b_1, \ldots, s_h)$ be an arbitrary finite history ending at step h and state $s_h = s$. A history-dependent leader strategy selects a distribution $p_h(\cdot \mid \mathfrak{h}_h) \in \Delta(\mathcal{A})$ and a full-support recommendation $y_h(\cdot \mid \mathfrak{h}_h) \in \mathrm{int}\, \Delta(\mathcal{B})$. Preservation at that history implies that the executed response is $\mu^{p_h(\cdot \mid \mathfrak{h}_h)}_{\theta^*_h,h}(\cdot \mid s)$. We prove by backward induction that the conditional leader value of every preservation-constrained history-dependent strategy is at most $W^*_h(s)$. At step $H+1$, both values are zero, so the claim holds. Assume that the claim holds at step $h+1$ for every state and every history ending at that state. Fix a history $\mathfrak{h}_h$ ending at s and abbreviate $p := p_h(\cdot \mid \mathfrak{h}_h)$. Conditionally on $\mathfrak{h}_h$, the myopic follower uses $\mu^p_{\theta^*_h,h}(\cdot \mid s)$. For each current joint action (a, b) and each realized next state s', the induction hypothesis bounds the continuation value of the original

history-dependent strategy by $W^*_{h+1}(s')$. Therefore, the conditional value of the original strategy at $\mathfrak{h}_h$ is at most

$$\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a)\mu^p_{\theta^*_h,h}(b \mid s)\left[u_h(s,a,b) + (P_h W^*_{h+1})(s,a,b)\right].$$

By (9), the last display equals

$$\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a)\mu^p_{\theta^*_h,h}(b \mid s) U^*_h(s,a,b).$$

Maximizing over $p \in \Delta(\mathcal{A})$ gives $W^*_h(s)$ by (10). This proves the induction upper bound for every history-dependent strategy satisfying preservation. Conversely, the objective in (10) is continuous in p and $\Delta(\mathcal{A})$ is compact, so a maximizer $p^*_h(s)$ exists for every (h,s). Because $\mathcal{S}$ is countable and carries its discrete sigma-field, a fixed tie-breaking rule produces a measurable Markov selector $s \mapsto p^*_h(s)$. Define the leader policy and recommendation policy by

$$\pi^*_h(\cdot \mid s) := p^*_h(s), \qquad y^*_h(\cdot \mid s) := \mu^{\pi^*}_{\theta^*_h,h}(\cdot \mid s).$$

The pair (π^*, y^*) satisfies the preservation constraint and induces the QR by Theorem 3.3. Backward induction in (9)–(10) shows that it attains $W^*_h(s)$ at every step and state. Hence enlarging the comparison class to history-dependent preserving pairs does not improve the optimal value. □

Proposition G.4 (Boundary noncoercivity of the recommendation–response distance)**.** *Suppose $0 < \kappa < 1$ and consider an equal-reward two-action context, so $\mu = (1/2, 1/2)$. For $\epsilon \in (0, 1/2)$, let the recommendation be $y_\epsilon = (1-\epsilon, \epsilon)$ and let ν^{y_ϵ} be the induced response. Then*

$$\mathrm{TV}(y_\epsilon, \nu^{y_\epsilon}) \longrightarrow 0, \qquad \mathrm{TV}(\nu^{y_\epsilon}, \mu) \longrightarrow \frac{1}{2} \qquad \text{as } \epsilon \downarrow 0.$$

Consequently, on the entire open simplex there is no uniform modulus g with $g(r) \to 0$ as $r \downarrow 0$ such that

$$\mathrm{TV}(\nu^y, \mu) \le g(\mathrm{TV}(y, \nu^y))$$

for every recommendation $y \in \operatorname{int}\Delta(\mathcal{B})$.

Proof of Proposition G.4. The equal-reward response has second coordinate

$$\nu_2^{y_\epsilon} = \frac{\epsilon^\kappa}{(1-\epsilon)^\kappa + \epsilon^\kappa}.$$

For sufficiently small ϵ, $\nu_2^{y_\epsilon} > \epsilon$, and hence

$$\mathrm{TV}(y_\epsilon, \nu^{y_\epsilon}) = \nu_2^{y_\epsilon} - \epsilon = \Theta(\epsilon^\kappa) - \epsilon \longrightarrow 0.$$

At the same time, $\nu_2^{y_\epsilon} \to 0$, so $\mathrm{TV}(\nu^{y_\epsilon}, \mu) = 1/2 - \nu_2^{y_\epsilon} \to 1/2$. The final claim follows by contradiction along this sequence. This boundary obstruction explains the coordinate-floor assumption in the generic bound on preservation violation; it does not affect the direct certificate in (97) for PROM (Algorithm 1). □

Proposition G.5 (Relaxed preservation can induce linear absolute regret)**.** *Suppose $\tau > 0$ and fix any radius $\alpha \in (0, 1/2)$. There exists a one-state, one-step, two-follower-action instance in which maximizing the leader's value over the relaxation*

$$\mathrm{TV}(y, \mu) \le \alpha$$

selects a recommendation at distance α from the QR and has suboptimality $-c_\kappa(\alpha)$, where the gain over the QSE benchmark is

$$c_\kappa(\alpha) := \frac{(1/2+\alpha)^\kappa}{(1/2+\alpha)^\kappa + (1/2-\alpha)^\kappa} - \frac{1}{2} > 0.$$

Repeating this relaxed-optimal pair for T episodes gives

$$\mathrm{Reg}(T) = -c_\kappa(\alpha)T, \qquad \mathrm{Reg}^{\mathrm{abs}}(T) = c_\kappa(\alpha)T, \qquad \mathrm{RRD}_{\mathrm{cum}}(T) = [\alpha - c_\kappa(\alpha)]T.$$

Moreover,

$$c_\kappa(\alpha) = \kappa\alpha + \frac{4}{3}\kappa(1-\kappa^2)\alpha^3 + \mathcal{O}(\alpha^5). \tag{78}$$

Proof of Proposition G.5. Let the leader have one action and let the follower have two actions. Give both follower actions reward zero, so $\mu = (1/2, 1/2)$, and give the leader rewards $u(1) = 0$ and $u(2) = 1$. Hence the preservation benchmark is $J^* = 1/2$. The leader's return is increasing in the second coordinate of y, so the relaxed optimum is $y_\alpha = (1/2 - \alpha, 1/2 + \alpha)$. Its executed probability of action two is

$$\nu_2 = \frac{(1/2+\alpha)^\kappa}{(1/2+\alpha)^\kappa + (1/2-\alpha)^\kappa} = \frac{1}{2} + c_\kappa(\alpha).$$

Because $0 < \kappa < 1$, the map $x \mapsto x^\kappa$ pulls the odds ratio toward one, so $0 < c_\kappa(\alpha) < \alpha$. It follows that the one-step suboptimality is $-c_\kappa(\alpha)$, its absolute value is $c_\kappa(\alpha)$, and the RRD is $\mathrm{TV}(y_\alpha, \nu) = \alpha - c_\kappa(\alpha)$. Repetition proves the three cumulative identities. Finally, expand $\log[(1+2\alpha)/(1-2\alpha)]$ and the logistic map at zero to obtain (78). □

Proposition G.6 (Conservative dominance of the QR bonus)**.** *For every $t \in [T]$, $h \in [H]$, $s \in \mathcal{S}$, and $p \in \Delta(\mathcal{A})$,*

$$b_h^{t,\mathrm{F}}(s,p) \le b_h^{t,\mathrm{Q}}(s,p), \qquad b_h^{t,\mathrm{P}}(s,p) \le b_h^{t,\mathrm{Q}}(s,p).$$

Consequently, the QR bonus can conservatively control both the execution edge and the true QR-deviation edge. Retaining the separate follower-response bonus gives the sharper RRD constant, while the derived preservation certificate gives the sharper preservation and absolute-regret constants.

Proof of Proposition G.6. The bonuses have the same confidence radius and covariance width. Moreover, $0 < \varepsilon \le \eta$, and $x \mapsto xe^x/2$ is increasing on $[0, \infty)$. Hence $L_\varepsilon \le L_\eta$, and truncation at one preserves the ordering. The second inequality follows from $0 \le \kappa < 1$ and the common radius and covariance width. □

State and action spaces. The main results allow a finite or countably infinite discrete state space. Conditional on any episode-start history, every state-indexed rule on such a space is measurable. The bounded-variation condition in Definition 2.1 makes the vector series defining each linear Bellman parameter absolutely convergent. Neither the likelihood argument nor the deterministic sup-norm cover takes a union bound over states, so the concentration proof introduces no additional state-cardinality factor.

The finite leader action set makes $\Delta(\mathcal{A})$ compact and ensures attainment of the continuous statewise planning objective. The finite follower action set makes the softmax normalizer and conditional likelihood finite sums and supports the finite-action covariance and total-variation analysis. Extending either action set to a countably infinite set would require additional attainment, base-measure, and positivity conditions and is not covered here.

An extension to an uncountable standard-Borel state space would require measurable features, a bounded-variation vector transition measure, measurable statewise maximizers or a measurable planning oracle, and the same deterministic sup-norm entropy bound for the generated value class.

G.1 Additional implementation properties

Remark G.7 (A uniform recommendation is not absence of recommendation). Suppose $\tau > 0$ and $y_h(\cdot \mid s)$ is uniform on $\mathcal{B}$. Then

$$\nu_{\theta,h}^{\pi,y}(b \mid s) = \frac{\exp\{\varepsilon r_{\theta,h}^\pi(s,b)\}}{\sum_{c\in\mathcal{B}} \exp\{\varepsilon r_{\theta,h}^\pi(s,c)\}},$$

where $\varepsilon < \eta$. If the action-reward vector is nonconstant, its log odds differ from those of $\mu_{\theta,h}^\pi$, so the uniform recommendation does not preserve the QR. The recommendation-free counterfactual removes the KL penalty; a uniform recommendation at positive sensitivity retains it.

Remark G.8 (The full-support qualification is essential). If the protocol were extended to boundary recommendations, then for any nonempty proper support $S \subsetneq \mathcal{B}$ and $\tau > 0$, the distribution obtained by conditioning μ on S would be a fixed point of the recommendation-to-response map on that support. It would not satisfy preservation because the QR has full support. Conversely, every boundary recommendation equal to its induced response under positive sensitivity is obtained by conditioning μ on the support of the recommendation. Thus the uniqueness and the equivalence

between preservation and $y = \nu$ in Theorem 3.3 are statements about the full-support protocol class specified in the model, not about arbitrary boundary references.

Remark G.9 (Why the recommendation-free entropy term is not redundant). If the entropy term were removed from (3) while $\tau > 0$, the response would satisfy $\nu_b^y \propto y_b e^{r_b/\tau}$. An interior solution of $y = \nu^y$ would then require $e^{r_b/\tau}$ to be constant over all follower actions, which fails in a generic reward context. The entropy term therefore does more than duplicate the KL regularizer: it defines the QR benchmark and permits a unique recommendation $y \in \operatorname{int}\Delta(\mathcal{B})$ satisfying $y = \nu^y$ under heterogeneous rewards.

Proposition G.10 (Information attenuation under an estimated preservation constraint). *Fix (h, s, p), a fitted parameter $\widehat{\theta} \in \Theta_h$, and hold the recommendation $y = \mu^p_{\widehat{\theta},h}(\cdot \mid s)$ fixed while varying a candidate parameter $\theta \in \Theta_h$. Define the single-observation conditional Fisher information matrices*

$$\mathcal{I}_{\mathrm{rec}}(\theta; y) := \mathbb{E}_{b\sim\nu^{p,y}_{\theta,h}}\left[\left(\nabla_\theta \log \nu^{p,y}_{\theta,h}(b \mid s)\right)\left(\nabla_\theta \log \nu^{p,y}_{\theta,h}(b \mid s)\right)^\top\right],$$

$$\mathcal{I}_{\mathrm{QR}}(\vartheta) := \mathbb{E}_{b\sim\mu^{p}_{\vartheta,h}}\left[\left(\nabla_\vartheta \log \mu^{p}_{\vartheta,h}(b \mid s)\right)\left(\nabla_\vartheta \log \mu^{p}_{\vartheta,h}(b \mid s)\right)^\top\right].$$

With the interpolated parameter $\bar{\theta}(\theta) := (1-\kappa)\theta + \kappa\widehat{\theta}$, the interpolation identity gives

$$\mathcal{I}_{\mathrm{rec}}(\theta; y) = (1-\kappa)^2 \mathcal{I}_{\mathrm{QR}}\big(\bar{\theta}(\theta)\big). \tag{79}$$

In particular, at the fitted parameter,

$$\mathcal{I}_{\mathrm{rec}}(\widehat{\theta}; y) = (1-\kappa)^2 \mathcal{I}_{\mathrm{QR}}(\widehat{\theta}). \tag{80}$$

Proof of Proposition G.10. Holding y fixed, the recommendation-conditioned score is

$$\nabla_\theta \log \nu^{p,y}_{\theta,h}(b \mid s) = \varepsilon\left(\phi^p_h(s,b) - \mathbb{E}_{c\sim\nu^{p,y}_{\theta,h}}[\phi^p_h(s,c)]\right).$$

The interpolation identity gives $\nu^{p,y}_{\theta,h} = \mu^p_{\bar{\theta}(\theta),h}$, while the QR score at ϑ is η times the centered feature under $\mu^p_{\vartheta,h}$. Taking second moments and using $\varepsilon/\eta = 1-\kappa$ proves (79); setting $\theta = \widehat{\theta}$ gives $\bar{\theta}(\widehat{\theta}) = \widehat{\theta}$ and proves (80). □

Proposition G.10 is a fixed-context comparison with the fitted recommendation held fixed.

G.2 The inactive-channel boundary

The following supplementary specialization allows $\tau = 0$, outside the positive-sensitivity standing condition of the principal model. This case connects the model to QSE learning with an inactive recommendation channel. At zero sensitivity, the ER equals the QR for every recommendation; the rule $y = \mu$ remains a canonical announcement, but is no longer necessary for preservation. For PROM (Algorithm 1), $\varepsilon = \eta$ and the two numerical response bonuses coincide: $b_h^{t,\mathrm{F}} = b_h^{t,\mathrm{Q}}$. RRD need not vanish during learning, because the announced fitted QR can still differ from the true QR, even though true preservation holds automatically when $\tau = 0$.

Corollary G.11 (Sharp zero-recommendation-weight baseline). *If $\tau = 0$, then for every step-parameter tuple $\theta = \{\theta_h\}_{h\in[H]}$, leader policy π, and full-support recommendation y,*

$$\nu^{\pi,y}_\theta = \mu^\pi_\theta.$$

Consequently, every evaluated pair satisfies $J(\pi, y) = J^{\mathrm{pol}}(\pi, \mu^\pi_{\theta^*}) \le J^*$, and hence

$$\mathrm{SubOpt}(\pi, y) \ge 0, \qquad \mathrm{Reg}^{\mathrm{abs}}(T) = \mathrm{Reg}(T). \tag{81}$$

H Further Discussion of Scope

Behavioral and protocol scope. The ER is a tractable reduced-form representation of recommendation sensitivity. At each visited state, the follower conditions on the publicly specified distribution $y_h(\cdot \mid s)$ rather than only on one action sampled from that distribution. The main theorem treats η and τ as known common coefficients.

Information and deployment scope. The leader estimates the follower's reward parameter from observed actions, without access to her reward function or realized rewards.

The recommendation is a behavioral reference and carries no payoff-relevant private signal beyond the state and commitment already represented in the model. If it conveyed new information about payoffs, states, or types, the appropriate no-channel benchmark would need to incorporate that information and preservation relative to the uninformed QR could be too conservative.

Constraint and enforcement scope. The exact implementation of the preservation constraint in Definition 3.4 requires the follower's true QR under the chosen commitment. However, under the no-disclosure requirement, θ^* cannot be disclosed to either the leader or the regulator, so the preservation constraint should be implemented using estimates of θ^*. The leader determines an estimate $\hat{\theta}_h^t$ from the interaction history available before episode t, and then selects the policy pair (π^t, y^t) subject to the *estimated preservation constraint* based on $\hat{\theta}_h^t$.

This yields (12), the estimated version of Definition 3.4.

The regulator requires the recommendation to satisfy (12) using the same reward-parameter estimate as the leader's planner. The leader retains the choice of commitment and estimation method. Our algorithm uses conditional MLE, as specified in Section 5.

The preservation constraint formalizes an external regulatory requirement and defines the full-information benchmark. PROM uses the estimated preservation constraint and provides statistical certificates for statewise RRD and preservation violation. The regulator is not modeled as a strategic player, and the analysis does not model enforcement of the constraint.

Normative scope of preservation. The preservation constraint regulates the additional effect of the recommendation while retaining the leader's choice of commitment. Preservation is a conservative, model-relative channel-design target under hidden follower rewards and potentially misaligned leader objectives. Under full information, the constrained game reproduces the QSE outcome set and value under the same commitment class, while allowing arbitrary recommendations may increase the leader's return through steering. The construction in Theorem E.18 shows how pursuing this additional return can produce long-term follower-side losses. The identity in (7) measures loss in the entropy-regularized criterion and is not automatically an unregularized expected-reward or welfare loss.

Scope of the response-control guarantees. During learning, PROM (Algorithm 1) provides cumulative guarantees under the true response model, combining vanishing average absolute leader regret with vanishing average follower-side effects. CPV, RG, CG, and auxiliary cumulative RRD use episode-start conditional expectations and the actual occupancies induced by the announced policy pairs. The bounds imply vanishing averages for fixed problem parameters, rather than true preservation in every episode or uniformly over all states. RG compares expected rewards at the same visited state and commitment, while CG compares the corresponding entropy-regularized decision criteria; neither is a difference in total returns under distinct trajectory laws or QSE commitments. During learning, the estimated preservation constraint can coexist with nonzero true RRD and nonzero preservation violation.

Myopic follower. Only the current reward enters the follower's response. A farsighted follower would solve a recommendation-dependent entropy-regularized MDP, coupling recommendations across future states. The present myopic conditional likelihood would have to be replaced by one derived from the corresponding dynamic response model; its soft Bellman recursion would introduce additional dependencies.

Unknown episode budget. PROM (Algorithm 1) uses a declared episode budget to tune the confidence radii and the value-cover resolution. Corollary E.16 removes this input by geometric restarts and preserves the same rates up to logarithmic factors. The wrapper permits either exact MLEs or certified phasewise negative-log-likelihood gaps satisfying (62). A nonrestarting confidence-sequence implementation could reuse all observations. Because the restart event is simultaneous over all episode prefixes, Corollary E.17 also permits any almost surely finite episode-end stopping rule without a new probability budget.

Computational scope. The algorithm is statistically efficient under the estimation and planning assumptions in Section 5. By fixing the MLE before planning, PROM optimizes only over the leader's action distribution at each state.

The theorem uses an exact statewise optimization oracle over the continuum $\Delta(\mathcal{A})$. For a finite state set, explicit enumeration adds a linear factor in the number of states. For a countably infinite state set, feature evaluation and statewise planning are on-demand oracle operations. The follower negative log-likelihood is convex and can be minimized to prescribed numerical accuracy when each Θ_h has a suitable projection or separation oracle. The main theorem analyzes exact negative-log-likelihood minimizers, while Corollary E.5 propagates a certified additive negative-log-likelihood-optimization error. Fixing the MLE before planning removes a search over the reward confidence set from the planning step, while the statewise maximization remains an optimization interface. A direct replay implementation of (18) revisits $O(HT^2)$ historical transition records over T episodes because each regression uses the current continuation value. This count excludes the costs of likelihood optimization, matrix operations, and statewise planning; for finite state spaces, the regression sums can instead be aggregated by successor state. A fully end-to-end polynomial-time theorem would additionally need to specify the oracle complexity of the parameter sets, the state representation, and the arithmetic model. A planner polynomial in a growing action dimension, or optimization over an arbitrary convex parameter set without an oracle, requires additional structure.

State-space dependence. The generic linear-feature regret bound has no separate union-bound factor in $|\mathcal{S}|$. The state space nevertheless enters through the feature dimension, the linear transition representation, the bound B_P, and the cost of statewise feature and planning queries. In a canonical finite-state tabular representation, these quantities generally scale with $|\mathcal{S}|$.

Behavioral misspecification. The well-specified proof cannot be reused unchanged when the fitted behavioral parameters are wrong, because the data then lie outside the fitted likelihood family. The robustness result begins with an approximate-likelihood concentration theorem and explicitly separates statistical uncertainty from model bias.

Equilibrium interpretation. Under the main model, the preservation constraint makes the recommendation and ER equal to the QR, so each constrained optimum projects to an ordinary myopic QSE policy pair. The inactive-channel specialization is described in Section G.2. The recommendation channel changes the likelihood when $y \neq \nu$ and the online learning dynamics but does not create a new payoff-relevant full-information equilibrium in this model. This distinction is important when comparing the result with Chen et al. (2023).

Rate comparison. In the generic multi-step standard-regret bound, a Bellman confidence radius of order d is multiplied by an elliptical-potential factor of order $\sqrt{Td}$ and thereby yields the displayed $d^{3/2}$ dependence. In the one-step deterministic-reward specialization, the leader-regret term is smaller because the continuation-value class is the singleton $\{0\}$ and the terminal transition martingale vanishes. The one-step cumulative RRD rate follows separately by setting $H = 1$ in the response-confidence bound; it does not rely on the continuation-value cover. For PROM (Algorithm 1), absolute regret is controlled directly by the analysis-only preservation certificate and therefore has the same displayed order without an inverse-map factor.

Future research. Richer recommendation interfaces, farsighted follower responses, offline learning under the preservation constraint, and runtime guarantees for statewise planning are natural directions for future research.

I Modeling Scope, Comparisons, and Illustrative Examples

I.1 Recommendation sensitivity and regulatory motivation

Recommendations are an established part of decision support, and their effects can extend beyond communicating available options. In an online-shopping experiment, Häubl & Trifts (2000) study recommendation agents that help consumers screen products and change their consideration sets and decision quality. Advice-taking research documents how decision makers adjust their judgments

after receiving recommendations (Bailey et al., 2023), while three controlled laboratory experiments show that displayed recommendation ratings anchor consumers' subsequent preference ratings (Adomavicius et al., 2013). These findings motivate treating a recommendation as behaviorally active. We model its influence through a Kullback–Leibler (KL) penalty relative to the recommended policy, using the recommendation as a reference for the follower's executed response. Relative-entropy costs and reference-sensitive stochastic choice have foundations in bounded rationality and rational inattention (Ortega & Braun, 2013; Matějka & McKay, 2015); the particular response model adopted here is a reduced-form modeling choice.

However, a self-interested leader may use this influence of recommendations to pursue her own interests at the follower's expense. Aridor & Gonçalves (2022) model a platform that biases recommendations toward its own products and identify circumstances in which consumers lose from its dual role as producer and recommender. Chen & Tsai (2024) provide evidence that Amazon's strategic incentives affect algorithmic recommendations in ways not explained by consumer preferences. External regulation of such interfaces also has concrete precedents: the European Union's Digital Services Act specifies recommender-system transparency and, for very large online platforms and search engines, an option not based on profiling, while the Digital Markets Act restricts preferential ranking of a gatekeeper's own products and services (European Parliament and Council of the European Union, 2022b;a). These examples motivate studying externally imposed constraints on recommendations, and we propose the *preservation constraint* as a theoretical design for this purpose. The leader remains self-interested and optimizes her return over the feasible set induced by this constraint.

I.2 Comparison with mechanism-design IC

Requiring the recommendation to equal its induced response superficially resembles an obedience requirement in mechanism and information design (Myerson, 1979; 1983; Kamenica & Gentzkow, 2011; Bergemann & Morris, 2019). The distinction lies in the behavioral and information structure: our follower follows the specified ER, and the announced distribution is required to equal that response. It is therefore a model-specific distributional condition, not Bayesian incentive compatibility or a family of obedience inequalities under a signal and type model. For comparison, Bayesian incentive-compatible bandit exploration uses information disclosure to make recommendations compatible with agents' Bayesian incentives (Mansour et al., 2015). In our setting, the leader receives no follower-reward feedback and fits the conditional response model from endogenous follower actions; unique reconstruction of the reward vector is neither assumed nor required.

The absence of transfers separates our model from principal-agent bandits with explicit payments (Scheid et al., 2024). Here the leader influences behavior only through a commitment distribution and a recommendation distribution. The statistical problem is therefore to learn an unknown response map from observed endogenous follower actions while propagating its uncertainty through dynamic planning.

I.3 Comparison with steering and incentive payments

Recent work on steering learning agents also seeks to move strategic behavior toward a desired outcome, but it uses a different mechanism and information structure (Zhang et al., 2024). In payment-based steering, a mediator changes the players' effective payoffs by offering transfers and typically computes those transfers from a known or directly observed game model. Our leader offers no transfer and does not alter the follower's reward. She announces only a commitment policy and a recommendation policy, and she must fit the follower's conditional response model from observed actions. The follower in our model responds to the current policy pair through a specified bounded-rational rule, rather than running an arbitrary no-regret learning algorithm across episodes. These differences explain why payment design and equilibrium-selection guarantees cannot replace the conditional-likelihood, RRD, and Bellman-confidence analyses developed here.

I.4 Comparison with other Stackelberg-learning models

The closest statistical model is the myopic-follower setting of Chen et al. (2023), where the follower's action distribution depends on the leader policy alone. Here it depends on both the leader policy

and the announced recommendation, so the distribution executed during learning can differ from the quantal response that defines the preservation benchmark. This leads to the occupancy-weighted RRD and preservation-violation criteria, the model-level absolute-regret bridge in Theorem E.9, and the sharper algorithm-specific bridge in Theorem E.7. As in the online linear-myopic analysis of Chen et al. (2023), fitted response uncertainty enters optimistic planning and makes the generated continuation values data dependent. Here the fitted ER arises from a recommendation-conditioned likelihood and the recommendation is fixed by the estimated preservation constraint. We therefore rederive a deterministic noncircular cover for this recursion. The follower-response bonus b^{F} and the direct preservation certificate b^{P} are used only in the performance analysis and do not enlarge the generated value class. These technical links support joint learning and protection guarantees under the estimated constraint, while the construction in Theorem E.18 and corollary 6.3 identifies follower losses that low leader regret alone does not rule out without preservation.

Table 4 summarizes the research objectives and statistical connections between the online linear-myopic setting of Chen et al. (2023) and the present work.

Table 4: Research objectives and statistical connections to the online linear-myopic QSE-learning framework of Chen et al. (2023).

Comparison	Online linear-myopic QSE learning	This paper
Full-information target	Optimal commitment against true QR	QSE benchmark preserved by the constraint
Observed response	Commitment-conditioned true QR	True ER, which may differ from the true QR
Learning task and evaluation	Learn the commitment from QR feedback	Commitment learning under estimated preservation, with joint protection guarantees and a loss construction without preservation
Statistical backbone and added link	Likelihood-based response learning and optimistic value learning	Builds on that backbone; fitted-response interpolation links execution, preservation, and the signed suboptimality

Calibrated Stackelberg learning studies a follower who reacts to forecasts of the leader's play (Haghtalab et al., 2023). By contrast, the follower here observes the leader's current commitment and recommendation directly, while the leader fits a reward-parameterized conditional response model through strategic feedback. The preservation constraint restricts the recommendation to the QR under the specified model; its estimated implementation is distinct from calibration of a forecast sequence.

I.5 Executed Response Policy: Interpretation and Scope

For fixed (h, s, π, y), the variational problem in (3) is an as-if representation of reference-dependent stochastic choice. The Shannon-entropy term defines the QR benchmark, while the KL term introduces incremental dependence on the public recommendation. The first-order condition yields the closed form (4), which makes the response model amenable to likelihood-based estimation. Free-energy and rational-inattention models provide theoretical precedents for KL-to-reference objectives (Ortega & Braun, 2013; Matějka & McKay, 2015). For any fixed finite collection of contexts, the represented reward parameter is uniquely identified over Θ_h exactly under the kernel condition in (77); identification can fail along reward-shift or other unexcited centered-feature directions. The regret proof does not assume this condition and instead controls response-prediction error in covariance seminorms. The parameter τ changes both the effective inverse temperature ε and the recommendation exponent κ, and therefore also changes estimation curvature and robustness constants.

Rationality limits. As $\eta \downarrow 0$, the quantal response converges to the uniform distribution on $\mathcal{B}$. As $\eta \to \infty$, its probability mass concentrates on the follower actions that maximize the expected reward.

Follower's value functions. As the follower is myopic, her step-h action-value function is simply the expected reward $Q^{\pi}_{\theta,h}(s,b) := r^{\pi}_{\theta,h}(s,b) = \langle \phi^{\pi}_h(s,b), \theta \rangle$. For the ER, the optimal value of the recommendation-dependent objective in (3) is $V^{\pi,y}_{\theta,h}(s) := \varepsilon^{-1} \log\Big(\sum_{b\in\mathcal{B}} y_h(b \mid s)^{\kappa} \exp\{\varepsilon Q^{\pi}_{\theta,h}(s,b)\}\Big)$. Define the corresponding advantage function by

$A^{\pi,y}_{\theta,h}(s,b) := Q^{\pi}_{\theta,h}(s,b) - V^{\pi,y}_{\theta,h}(s)$. The response policy then admits the equivalent representation $\nu^{\pi,y}_{\theta,h}(b \mid s) = y_h(b \mid s)^{\kappa} \exp\{\varepsilon A^{\pi,y}_{\theta,h}(s,b)\}$.

In particular, when $\tau = 0$ in (3), the follower's response reduces to the QR. In this case, the follower's value function is $V^{\pi}_{\theta,h}(s) := \eta^{-1} \log \sum_{b \in \mathcal{B}} \exp\{\eta Q^{\pi}_{\theta,h}(s,b)\}$, and the corresponding advantage function is $A^{\pi}_{\theta,h}(s,b) := Q^{\pi}_{\theta,h}(s,b) - V^{\pi}_{\theta,h}(s)$. Then, (2) is equivalently $\mu^{\pi}_{\theta,h}(b \mid s) = \exp\{\eta A^{\pi}_{\theta,h}(s,b)\}$.

Under the preservation constraint in Definition 3.4, the follower's recommendation-dependent and recommendation-free optimal values also coincide: $V^{\pi,y}_{\theta_h,h}(s) = V^{\pi}_{\theta_h,h}(s)$. Indeed, the ER equals the QR by Theorem 3.3, and the KL penalty in (3) vanishes.

I.6 Stylized dynamic examples

The following examples illustrate how current joint actions affect future states.

1. **Inventory coordination.** A manufacturer commits to shipment probabilities and recommends an ordering distribution to a recommendation-sensitive retailer. Inventory and backlog form the state, joint actions determine the next inventory, and the retailer's holding and shortage costs are hidden from the manufacturer.
2. **Security allocation.** A central authority commits to a patrol mixture and recommends a local deployment mixture. Threat level and recent incidents form the state. The local unit has private operational costs and responds myopically to the current recommendation, while deployments affect future threat states.
3. **Demand response.** A system operator commits to a supply policy and recommends a consumption profile. Load, storage, and congestion form the state. Consumers have hidden comfort costs; current joint decisions change the next load state.

In these dynamic settings, state visitation determines which centered-feature directions of the follower's response model and which transition-feature directions are statistically informed, while backward value propagation determines how errors at later steps affect earlier decisions.

J Key Technical Challenges

Adaptive follower likelihood. The contexts in the MLE are generated by earlier estimates and by state visitation. Pointwise iid likelihood arguments are therefore insufficient. After normalization by the conditional Hellinger affinities, the square-root likelihood-ratio process in Lemma M.2 is a nonnegative martingale under predictable adaptive contexts. The resulting concentration bound is uniformized over the compact parameter class.

Two response errors. At the fitted parameter, the recommendation equals the fitted QR and is a fixed point of the recommendation-to-response map. At the true parameter, however, the executed response and the benchmark QR are different distributions. The two TV errors in Proposition M.4 consequently enter different lines of the performance-difference proof. The regret analysis requires controlling the distance between the true and estimated quantal responses, whereas observed actions follow the executed response. Likelihood guarantees for the executed response therefore do not directly yield the quantal-response error bounds needed for sublinear regret. We exploit invariance to constant reward shifts to center reward errors and use the equality of the fitted executed response and quantal response to control both response errors through covariance bounds. This transfers statistical guarantees from executed-response data to the quantal-response quantities governing regret and preservation.

Adaptive Bellman targets. The target in (18) contains a future value function produced by the same data. A self-normalized inequality for a single fixed target does not apply. Proposition N.1 covers the entire generated class, and Lemma O.1 pays the corresponding logarithmic complexity. The extension from a value-function net to transition targets contributes the factor $1 + B_P$.

Negative suboptimality. Deviations from preservation under the estimated constraint can produce negative values of suboptimality that offset positive values, so an upper bound on standard regret alone does not control absolute regret. We use fitted-response interpolation to control preservation violations, which also bound the negative part of the leader's suboptimality.

A pair that violates the preservation constraint may outperform the constrained benchmark in an episode, so standard regret can be negative. For recommendation sequences satisfying a common positive coordinate lower bound, remaining-horizon-weighted RRD and the global inverse map provide a model-level absolute-regret bridge. For PROM (Algorithm 1), the estimated-preservation identity, interpolation identity, and direct preservation certificate b^{P} control the same negative part and remove the inverse-map factor from the algorithm-specific rate.

Planning. The statewise objective is nonlinear in the leader mixture because the fitted QR and its covariance both depend on that mixture. The theorem uses an exact statewise planning oracle.

K Constants and Failure-Probability Ledger

This section records the numerical interfaces needed to keep the concentration, covering, and regret arguments mutually consistent.

Hellinger curvature. Centering the logit increment by its log-normalizer secant produces $g_b \in [-2, 2]$ and $p_{\theta'}(b)/p_\theta(b) = e^{x g_b}$. The pointwise inequality

$$|1 - e^z| \geq \frac{1 - e^{-x}}{x} |z|, \qquad z \in [-x, x],$$

then gives the following explicit curvature bound:

$$D^2_{\mathrm{H}}(p_\theta, p_{\theta'}) \geq \frac{(1 - e^{-x})^2}{8} \|\theta' - \theta\|^2_{\Sigma_\theta}.$$

This proof is independent of the conditional base measure y^κ, which cancels in the likelihood ratio. The elementary inequality $1 - e^{-x} \geq x/(1 + x)$ also gives $c_{\mathrm{H}}(x) \geq x^2/[8(1 + x)^2]$.

Deterministic-reward Bellman scale. For a fixed continuation value $W \in [0, H - h]$, the centered transition noise

$$W(s_{h+1}) - (P_h W)(s_h, a_h, b_h)$$

has conditional support in an interval of length $H - h$. Hoeffding's lemma therefore gives the valid conditional sub-Gaussian proxy $(H - h)/2$ under the convention $\mathbb{E}[\exp\{\lambda Z\} \mid \mathcal{F}] \leq \exp\{\lambda^2 R^2/2\}$. This is the coefficient used in (33) and (124). At $h = H$, it vanishes exactly.

Value cover. The explicit nets in (34) and (125) give the cover directly and avoid introducing a generic constant $C_{\mathcal{W}}$. The one-half Hölder dependence is paid for by using squared mesh for A, C, θ.

Probability budgets. The following table is the complete event ledger for $\delta \in (0, 1)$. Planning and all covering statements are deterministic.

Each certified approximate-MLE variant uses the predictive Hellinger event of its exact counterpart and analogous Bellman and potential events constructed from its recomputed deterministic cover and fitted covariance sequence. The well-specified transfer in Corollary E.5 also uses the squared-potential event because it transfers the criterion-gap guarantee; its four events receive budgets $\delta/4$ each. The misspecified transfer in Corollary R.5 does not assert a quadratic criterion-gap result and uses only its MLE, Bellman, and first-order-potential events, for total failure at most $3\delta/4$. Optimization-gap certificates and radius substitutions introduce no additional random event.

The geometric-restart wrapper does not assert a squared criterion-gap guarantee, so phase j repeats the three events needed for prefix regret and RRD with δ replaced by δ_j. The potential allocation is a maximal, prefix-uniform version of the same event, rather than a fourth event. Thus the conditional phase failure is at most $3\delta_j/4$, and summing over phases costs at most $3\delta/4$ because $\sum_{j \geq 0} \delta_j = \delta$.

Event	Content	Failure budget
$\mathcal{E}_{\mathrm{MLE}}$	All steps and episodes in Lemma M.3	$\delta/4$
$\mathcal{E}_{\mathrm{Bell}}$	Uniform Bellman event in Lemma O.1	$\delta/4$
$\mathcal{E}_{\mathrm{pot}}$	Both scalar and matrix expected potentials in Lemma P.4	$\delta/4$
$\mathcal{E}_{\mathrm{sq}}$	Squared expected matrix potential used for the criterion-gap result in Lemma P.5	$\delta/4$
Well-specified intersection	Union bound over the preceding events	$4\delta/4 = \delta$
$\bar{\mathcal{E}}_{\mathrm{MLE}}$	Predictive Hellinger and near-model log loss combined in Lemma S.3	$\delta/4$
$\bar{\mathcal{E}}_{\mathrm{Bell}}$	Robust value cover and Bellman event in Proposition R.2	$\delta/4$
$\bar{\mathcal{E}}_{\mathrm{pot}}$	Fitted scalar and matrix potentials in Corollary R.3	$\delta/4$
Misspecified intersection	Union bound over the preceding robust events	$3\delta/4 \le \delta$

Table 5: Failure-probability bookkeeping for the well-specified and behaviorally misspecified guarantees.

Misspecified Bernstein event. The proof of Lemma S.2 states the conditional sub-gamma Bernstein inequality it uses. For martingale differences bounded above by b and with deterministic conditional-variance bound v, it gives $S_n \le \sqrt{2vL} + bL/3$. Thus $b = 3D_{\mathrm{beh}}$ yields the explicit linear correction $D_{\mathrm{beh}}L$. A coarser inversion of the usual Freedman tail would instead yield the also valid correction $2D_{\mathrm{beh}}L$. The behavioral constants suppressed by simplified notation are exposed in (49) and (133).

L Exponential-Family and Matrix Identities

Fix a step, state, leader distribution, and recommendation. Write the feature vectors and reference weights as

$$x_b := \phi_h^\pi(s, b), \qquad w_b := y_h(b \mid s)^\kappa.$$

For $x > 0$ and parameter θ, define the exponential-family distribution and its log-partition function by

$$p_\theta(b) := \frac{w_b \exp\{x x_b^\top \theta\}}{\sum_{c \in \mathcal{B}} w_c \exp\{x x_c^\top \theta\}}, \qquad A_x(\theta) := \log \sum_{c \in \mathcal{B}} w_c \exp\{x x_c^\top \theta\}. \tag{82}$$

Then

$$\nabla A_x(\theta) = x \mathbb{E}_{p_\theta}[x_b], \qquad \nabla^2 A_x(\theta) = x^2 \operatorname{Cov}_{p_\theta}(x_b).$$

Lemma L.1 (Likelihood-ratio and covariance comparison). *Suppose $x_b^\top \theta, x_b^\top \theta' \in [0, 1]$ for every b. Then*

$$e^{-2x} \le \frac{p_{\theta'}(b)}{p_\theta(b)} \le e^{2x} \qquad \textit{for every } b. \tag{83}$$

Consequently, for every $v \in \mathbb{R}^d$,

$$e^{-2x} v^\top \Sigma_\theta v \le v^\top \Sigma_{\theta'} v \le e^{2x} v^\top \Sigma_\theta v, \tag{84}$$

where the feature covariance matrix is $\Sigma_\theta = \operatorname{Cov}_{p_\theta}(x_b)$.

Proof of Lemma L.1. Let the expected rewards under the two parameters be $q_b = x_b^\top \theta$ and $q'_b = x_b^\top \theta'$. The log-normalizer difference lies between the minimum and maximum of $x(q'_b - q_b)$. Since $q'_b - q_b \in [-1, 1]$, the log probability ratio lies in $[-2x, 2x]$, proving (83). For any scalar $g(b)$,

$$\operatorname{Var}_{p_{\theta'}}(g) = \min_{c \in \mathbb{R}} \mathbb{E}_{p_{\theta'}}[(g - c)^2] \le e^{2x} \mathbb{E}_{p_\theta}[(g - \mathbb{E}_{p_\theta} g)^2].$$

Apply this to $g(b) = v^\top x_b$, and reverse the roles of θ and θ'. □

Lemma L.2 (Trace self-normalization). *For $V \succ 0$, $\Sigma \succeq 0$, and $z \in \mathbb{R}^d$,*

$$z^\top \Sigma z \le \operatorname{Tr}(V^{-1}\Sigma) \, \|z\|_V^2 .$$

Proof of Lemma L.2. Let the transformed vector be $q = V^{1/2} z$. Since $qq^\top \preceq \|q\|_2^2 \, I$,

$$z z^\top \preceq \|z\|_V^2 \, V^{-1}.$$

Multiply by $\Sigma \succeq 0$ and take traces. □

Lemma L.3 (Logit perturbation). *Let the softmax distributions be $p_v = \operatorname{softmax}(v)$ and $p_w = \operatorname{softmax}(w)$ on a finite set. If $\|v - w\|_\infty \le D$, then*

$$\max_{b \in \mathcal{B}} |\log[p_v(b)/p_w(b)]| \le 2D, \tag{85}$$

$$\mathrm{KL}(p_v \| p_w) \le \frac{1}{2} D^2, \tag{86}$$

$$\mathrm{TV}(p_v, p_w) \le \frac{1}{2} D. \tag{87}$$

Proof of Lemma L.3. The log-sum-exp normalizer difference lies between the minimum and maximum coordinate differences, proving (85). Let the logit difference be $\Delta = w - v$. The Bregman identity gives

$$\mathrm{KL}(p_v \| p_w) = \int_0^1 (1 - z) \operatorname{Var}_{p_{v + z\Delta}}(\Delta_b) \, dz.$$

Because $\Delta_b \in [-D, D]$, its variance is at most D^2. Integrating proves (86); Pinsker's inequality gives (87). □

Lemma L.4 (Hellinger lower bound by response covariance)**.** *Define the squared Hellinger distance by*

$$D_{\mathrm{H}}^2(P,Q) := \frac{1}{2}\sum_{b\in\mathcal{B}}(\sqrt{P(b)}-\sqrt{Q(b)})^2.$$

Suppose that $x_b^\top\theta, x_b^\top\theta' \in [0,1]$ for every b. For the corresponding two distributions in (82),

$$D_{\mathrm{H}}^2(p_\theta, p_{\theta'}) \geq c_{\mathrm{H}}(x)\,\|\theta'-\theta\|_{\Sigma_\theta}^2 .$$

Proof of Lemma L.4. Let the parameter difference be $\Delta = \theta'-\theta$, the expected-reward difference be $d_b = x_b^\top\Delta$, and the scaled log-partition difference be

$$c = \frac{A_x(\theta')-A_x(\theta)}{x}.$$

Since $d_b \in [-1,1]$, the log-sum-exp variational bound gives $c \in [\min_{b\in\mathcal{B}} d_b, \max_{b\in\mathcal{B}} d_b]$. Thus $g_b = d_b - c \in [-2,2]$, and

$$\frac{p_{\theta'}(b)}{p_\theta(b)} = e^{xg_b}.$$

Consequently,

$$D_{\mathrm{H}}^2(p_\theta, p_{\theta'}) = \frac{1}{2}\mathbb{E}_{p_\theta}\left(1-e^{xg_b/2}\right)^2.$$

For $z \in [-x,x]$, concavity on the negative half-line and $e^z - 1 \geq z$ on the positive half-line imply

$$|1-e^z| \geq \frac{1-e^{-x}}{x}\,|z|\,.$$

Applying this with $z = xg_b/2$ yields

$$D_{\mathrm{H}}^2(p_\theta, p_{\theta'}) \geq \frac{(1-e^{-x})^2}{8}\mathbb{E}_{p_\theta}[g_b^2] \geq \frac{(1-e^{-x})^2}{8}\,\mathrm{Var}_{p_\theta}(d_b) = c_{\mathrm{H}}(x)\,\|\Delta\|_{\Sigma_\theta}^2 .$$

□

Lemma L.5 (Covariance-form TV sensitivity)**.** *Under the same assumptions,*

$$\mathrm{TV}(p_\theta, p_{\theta'}) \leq L_x\,\|\theta'-\theta\|_{\Sigma_\theta}\,, \qquad L_x = xe^x/2.$$

Proof of Lemma L.5. The Bregman identity and the upper half of (84) give

$$\mathrm{KL}(p_\theta\|p_{\theta'}) \leq \frac{x^2e^{2x}}{2}\,\|\theta'-\theta\|_{\Sigma_\theta}^2 .$$

Pinsker's inequality completes the proof. □

Lemma L.6 (Segment-scaled softmax sensitivity)**.** *Let the softmax distribution be $p_\theta(b) \propto \exp\{\eta x_b^\top\theta\}$ on a finite set, and suppose $x_b^\top\theta, x_b^\top\theta' \in [0,1]$ for every b. For $z \in [0,1]$, let the interpolated parameter be $\theta_z := (1-z)\theta + z\theta'$, let the corresponding distribution be $p_z := p_{\theta_z}$, and let the feature covariance matrix at the initial parameter be $\Sigma_0 := \mathrm{Cov}_{p_0}(x_b)$. Then, for every $0 \leq u \leq v \leq 1$,*

$$\mathrm{KL}(p_u\|p_v) \leq \frac{\eta^2(v-u)^2e^{2\eta v}}{2}\,\|\theta'-\theta\|_{\Sigma_0}^2\,, \tag{88}$$

$$\mathrm{TV}(p_u, p_v) \leq \frac{\eta(v-u)e^{\eta v}}{2}\,\|\theta'-\theta\|_{\Sigma_0}\,. \tag{89}$$

Proof of Lemma L.6. Let the log-partition function be $A(\vartheta) := \log\sum_b \exp\{\eta x_b^\top\vartheta\}$ and the parameter difference be $\Delta := \theta'-\theta$. The exponential-family Bregman identity gives

$$\mathrm{KL}(p_u\|p_v) = \eta^2(v-u)^2\int_0^1 (1-z)\,\|\Delta\|_{\Sigma_{u+z(v-u)}}^2\,\mathrm{d}z,$$

where the feature covariance matrix is $\Sigma_w := \mathrm{Cov}_{p_w}(x_b)$. For every $w \in [0, v]$, convexity of the reward interval gives $x_b^\top \theta_w \in [0, 1]$. Set the expected-reward difference $d_b := x_b^\top(\theta' - \theta) \in [-1, 1]$. Then

$$\log \frac{p_w(b)}{p_0(b)} = \eta w d_b - \log \mathbb{E}_{p_0} e^{\eta w d} \in [-2\eta w, 2\eta w].$$

Applying the same likelihood-ratio variance comparison as in (84), now with range $2\eta w$, yields $\Sigma_w \preceq e^{2\eta w}\Sigma_0 \preceq e^{2\eta v}\Sigma_0$. Substitution and $\int_0^1 (1 - z)\,\mathrm{d}z = 1/2$ prove (88). Pinsker's inequality proves (89). □

M Adaptive MLE Concentration and Response Bonuses

Proposition M.1 (Convexity of the follower likelihood)**.** *For fixed historical contexts, $\ell_h^t(\theta)$ is convex and continuously differentiable on Θ_h, with Hessian*

$$\nabla^2 \ell_h^t(\theta) = \varepsilon^2 \sum_{i=1}^{t-1} \Sigma_h^{\pi^i, y^i, \theta}(s_h^i) \succeq 0. \tag{90}$$

Proof of Proposition M.1. Fix h and abbreviate the observed feature vectors and recommendation probabilities as

$$x_i(b) := \phi_h^{\pi^i}(s_h^i, b), \qquad y_i(b) := y_h^i(b \mid s_h^i).$$

By (4), the contribution of observation i to the negative log likelihood is

$$\ell_{h,i}(\theta) := -\log \nu_{\theta,h}^{\pi^i, y^i}(b_h^i \mid s_h^i) = -\kappa \log y_i(b_h^i) - \varepsilon x_i(b_h^i)^\top \theta + \log \sum_{c \in \mathcal{B}} y_i(c)^\kappa \exp\{\varepsilon x_i(c)^\top \theta\}.$$

The first term is independent of θ, the second term is linear in θ, and the final term is a log-sum-exp of finitely many affine functions of θ. Hence $\ell_{h,i}$ is convex and infinitely differentiable on $\mathbb{R}^d$. Differentiating the displayed formula gives

$$\nabla \ell_{h,i}(\theta) = \varepsilon \left[\mathbb{E}_{b \sim \nu_{\theta,h}^{\pi^i, y^i}(\cdot \mid s_h^i)} x_i(b) - x_i(b_h^i) \right].$$

Differentiating once more yields

$$\nabla^2 \ell_{h,i}(\theta) = \varepsilon^2 \,\mathrm{Cov}_{b \sim \nu_{\theta,h}^{\pi^i, y^i}(\cdot \mid s_h^i)}[x_i(b)] = \varepsilon^2 \Sigma_h^{\pi^i, y^i, \theta}(s_h^i).$$

Every covariance matrix is positive semidefinite. Summing over $i = 1, \ldots, t-1$ proves

$$\nabla^2 \ell_h^t(\theta) = \varepsilon^2 \sum_{i=1}^{t-1} \Sigma_h^{\pi^i, y^i, \theta}(s_h^i) \succeq 0,$$

which is (90) and proves the proposition. □

For fixed h, let $p_h^{i,*}$ be the true conditional distribution of b_h^i at the observed context, and let $p_h^{i,\theta}$ be the conditional response under candidate $\theta \in \Theta_h$, with the same announced context.

Lemma M.2 (Adaptive predictive Hellinger concentration)**.** *With probability at least $1 - \delta/8$, simultaneously for all $h \in [H]$, $t \le T$, and $\theta \in \Theta_h$,*

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, p_h^{i,\theta}) \le \frac{1}{2}[\ell_h^t(\theta) - \ell_h^t(\theta_h^*)] + \frac{1}{4}\beta_\Theta(\delta). \tag{91}$$

Proof of Lemma M.2. Fix h and θ. Using the within-episode filtration in (28), define the fixed-step inter-episode filtration $\mathcal{J}_i^{(h)} := \mathcal{G}_h^{i+1}$ for $i \ge 0$. The inclusions $\mathcal{G}_h^i \subseteq \mathcal{H}_h^i \subseteq \mathcal{F}_H^i = \mathcal{F}_0^{i+1} \subseteq \mathcal{G}_h^{i+1}$ show that $\{\mathcal{J}_i^{(h)}\}_{i \ge 0}$ is a filtration. The announced context and the conditional distributions $p_h^{i,*}$ and

$p_h^{i,\theta}$ are $\mathcal{J}_{i-1}^{(h)}$-measurable, whereas b_h^i and the likelihood-ratio factor based on it are $\mathcal{J}_i^{(h)}$-measurable. Define the Hellinger affinity by

$$A_h^i(\theta) := \sum_{b\in\mathcal{B}} \sqrt{p_h^{i,*}(b)p_h^{i,\theta}(b)} = 1 - D_{\mathrm{H}}^2(p_h^{i,*}, p_h^{i,\theta}).$$

The square-root likelihood-ratio process

$$\mathcal{M}_h^t(\theta) := \prod_{i=1}^{t-1} \frac{\sqrt{p_h^{i,\theta}(b_h^i)/p_h^{i,*}(b_h^i)}}{A_h^i(\theta)}$$

is a nonnegative martingale with initial value one relative to $\{\mathcal{J}_i^{(h)}\}_{i\ge 0}$, because each factor has conditional mean one given $\mathcal{J}_{i-1}^{(h)} = \mathcal{G}_h^i$. Thus adaptive contexts cause no difficulty. This likelihood-ratio construction is a specialized exponential-supermartingale argument; compare the general conditional log-moment-generating-function inequality in Foster et al. (2021, Lemma A.4). Ville's inequality implies that, with probability at least $1-\delta'$, simultaneously in t,

$$\log \mathcal{M}_h^t(\theta) \le \log(1/\delta').$$

Using $-\log(1-z) \ge z$ gives

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, p_h^{i,\theta}) \le \frac{1}{2}[\ell_h^t(\theta) - \ell_h^t(\theta_h^*)] + \log(1/\delta'). \tag{92}$$

Let the net radius be $r = (8\varepsilon T)^{-1}$ and take an r-net $\mathcal{N}_h$ of Θ_h. Its cardinality is at most

$$N_\Theta \le (1 + 16B_\Theta \varepsilon T)^d.$$

Apply (92) to all net points and steps with $\delta' = \delta/(16HTN_\Theta)$. Ville's inequality is already uniform in t; the extra factor T is retained as slack.

Fix $\theta \in \Theta_h$ and choose $\theta^0 \in \mathcal{N}_h$ with $\|\theta - \theta^0\|_2 \le r$. For every outcome b,

$$\left|\log p_h^{i,\theta}(b) - \log p_h^{i,\theta^0}(b)\right| \le 2\varepsilon r =: D = \frac{1}{4T}.$$

It follows that

$$\left|\ell_h^t(\theta) - \ell_h^t(\theta^0)\right| \le (t-1)D \le \frac{1}{4}.$$

For the Hellinger term, write $p_h^{i,\theta}(b) = p_h^{i,\theta^0}(b)e^{d_b}$, where $|d_b| \le D$. Then

$$\left|D_{\mathrm{H}}^2(p_h^{i,*}, p_h^{i,\theta}) - D_{\mathrm{H}}^2(p_h^{i,*}, p_h^{i,\theta^0})\right| \le \sum_{b\in\mathcal{B}} \sqrt{p_h^{i,*}(b)p_h^{i,\theta^0}(b)}\left|e^{d_b/2} - 1\right| \le e^{D/2}\frac{D}{2}.$$

Therefore the cumulative Hellinger change is at most $Te^{D/2}D/2 \le e^{1/8}/8$, and the change in one half of the likelihood difference is at most $1/8$. Applying the net-point inequality and using

$$\log(1/\delta') \le d\log(1 + 16B_\Theta\varepsilon T) + \log(16HT/\delta)$$

shows that the total extension slack is less than one. The $+2$ term in (30) absorbs it, proving (91). □

Lemma M.3 (MLE concentration)**.** *With probability at least $1-\delta/4$, simultaneously for every $t \in [T]$ and $h \in [H]$,*

$$\sum_{i=1}^{t-1} \left\|\hat{\theta}_h^t - \theta_h^*\right\|^2_{\Sigma_h^{\pi^i, y^i, \hat{\theta}_h^t}(s_h^i)} \le \frac{\beta_\Theta(\delta)}{c_{\mathrm{H}}(\varepsilon)}. \tag{93}$$

Consequently,

$$\left\|\hat{\theta}_h^t - \theta_h^*\right\|_{V_h^t} \le R_\Theta. \tag{94}$$

Proof of Lemma M.3. Fix $t \in [T]$ and $h \in [H]$, and substitute the random MLE $\hat{\theta}_h^t$ into the event of Lemma M.2. This substitution is valid because that event is simultaneous over every $\theta \in \Theta_h$. Since $\hat{\theta}_h^t$ minimizes ℓ_h^t over Θ_h and $\theta_h^* \in \Theta_h$, we have $\ell_h^t(\hat{\theta}_h^t) - \ell_h^t(\theta_h^*) \le 0$. Therefore,

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, p_h^{i,\hat{\theta}_h^t}) \le \frac{1}{4}\beta_\Theta(\delta).$$

For every historical context i, apply Lemma L.4 with inverse temperature ε, covariance reference $\hat{\theta}_h^t$, and second parameter θ_h^*. This gives

$$D_{\mathrm{H}}^2(p_h^{i,\hat{\theta}_h^t}, p_h^{i,*}) \ge c_{\mathrm{H}}(\varepsilon) \left\| \hat{\theta}_h^t - \theta_h^* \right\|^2_{\Sigma_h^{\pi^i, y^i, \hat{\theta}_h^t}(s_h^i)}.$$

After summing over i, we obtain the stronger inequality

$$\sum_{i=1}^{t-1} \left\| \hat{\theta}_h^t - \theta_h^* \right\|^2_{\Sigma_h^{\pi^i, y^i, \hat{\theta}_h^t}(s_h^i)} \le \frac{\beta_\Theta(\delta)}{4c_{\mathrm{H}}(\varepsilon)},$$

and hence (93).

It remains to relate the preceding covariances to the historical matrix V_h^t. At every historical context, all admissible follower rewards lie in $[0, 1]$, so Lemma L.1 gives

$$\Sigma_h^{\pi^i, y^i, \hat{\theta}_h^i}(s_h^i) \preceq e^{2\varepsilon} \Sigma_h^{\pi^i, y^i, \hat{\theta}_h^t}(s_h^i).$$

Writing the parameter estimation error as $\Delta_h^t := \hat{\theta}_h^t - \theta_h^*$ and using $\|\Delta_h^t\|_2 \le B_\Theta$, we obtain

$$\left\|\Delta_h^t\right\|^2_{V_h^t} = \left\|\Delta_h^t\right\|_2^2 + \sum_{i=1}^{t-1} \left\|\Delta_h^t\right\|^2_{\Sigma_h^{\pi^i, y^i, \hat{\theta}_h^i}(s_h^i)} \le B_\Theta^2 + \frac{e^{2\varepsilon}}{c_{\mathrm{H}}(\varepsilon)}\beta_\Theta(\delta) = R_\Theta^2.$$

Taking square roots proves (94). The argument is simultaneous in (t, h) because the predictive Hellinger event is simultaneous in (t, h, θ). □

Proof of Corollary E.5. Fix $t \in [T]$ and $h \in [H]$. Apply Lemma M.2 with $\theta = \tilde{\theta}_h^t$. The certificate (47) and the inequality $\inf_{\theta \in \Theta_h} \ell_h^t(\theta) \le \ell_h^t(\theta_h^*)$ give $\ell_h^t(\tilde{\theta}_h^t) - \ell_h^t(\theta_h^*) \le \gamma_{\mathrm{MLE}}$. Substituting this inequality into (91) yields

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, p_h^{i,\tilde{\theta}_h^t}) \le \frac{1}{2}\gamma_{\mathrm{MLE}} + \frac{1}{4}\beta_\Theta(\delta).$$

Apply Lemma L.4 at each historical context, with $p_h^{i,\tilde{\theta}_h^t}$ as the covariance reference. After summing over i, one obtains

$$c_{\mathrm{H}}(\varepsilon) \sum_{i=1}^{t-1} \left\| \tilde{\theta}_h^t - \theta_h^* \right\|^2_{\Sigma_h^{\pi^i, y^i, \tilde{\theta}_h^t}(s_h^i)} \le \frac{1}{2}\gamma_{\mathrm{MLE}} + \frac{1}{4}\beta_\Theta(\delta).$$

In particular,

$$\sum_{i=1}^{t-1} \left\| \tilde{\theta}_h^t - \theta_h^* \right\|^2_{\Sigma_h^{\pi^i, y^i, \tilde{\theta}_h^t}(s_h^i)} \le \frac{\beta_\Theta(\delta) + 2\gamma_{\mathrm{MLE}}}{c_{\mathrm{H}}(\varepsilon)}.$$

For every $i < t$, Lemma L.1 gives

$$\Sigma_h^{\pi^i, y^i, \tilde{\theta}_h^i}(s_h^i) \preceq e^{2\varepsilon} \Sigma_h^{\pi^i, y^i, \tilde{\theta}_h^t}(s_h^i).$$

Writing the parameter estimation error as $\Delta_h^t := \tilde{\theta}_h^t - \theta_h^*$ and using $\|\Delta_h^t\|_2 \le B_\Theta$, we obtain

$$\left\|\Delta_h^t\right\|^2_{V_h^{t,\gamma}} = \left\|\Delta_h^t\right\|_2^2 + \sum_{i=1}^{t-1} \left\|\Delta_h^t\right\|^2_{\Sigma_h^{\pi^i, y^i, \tilde{\theta}_h^i}(s_h^i)} \le B_\Theta^2 + \frac{e^{2\varepsilon}}{c_{\mathrm{H}}(\varepsilon)}[\beta_\Theta(\delta) + 2\gamma_{\mathrm{MLE}}] = R_{\Theta,\gamma}^2.$$

This proves (48) on the predictive Hellinger event, whose failure probability is at most $\delta/8$ and hence at most $\delta/4$.

We next verify the substitutions used in the remaining conclusions. For every candidate (h, s, p), the estimated-preservation identity gives

$$y_h^{t,p,\gamma} = \nu_{\hat{\theta}_h^t,h}^{p,y_h^{t,p,\gamma}}.$$

Applying Lemmas L.2 and L.5 with the ellipsoidal bound just proved yields the two response inequalities with radius $R_{\Theta,\gamma}$ and the γ-superscripted covariance width. The historical approximate fitted covariance matrices satisfy the deterministic update

$$V_h^{t+1,\gamma} = V_h^{t,\gamma} + \Gamma_h^{t,\gamma}(s_h^t, \pi_h^t),$$

so the scalar and matrix determinant-potential identities hold for every realized sequence. The value-class envelope is recomputed with $R_{\Theta,\gamma}$ and $\beta_{h,\gamma}^{\mathrm{U}}$ before interaction. It is therefore deterministic, and the proof of Proposition N.1 gives the stated cover for the approximate fitted recursion. Apply the self-normalized Bellman argument of Lemma O.1 to this recomputed deterministic cover. This gives a Bellman-confidence event of failure probability at most $\delta/4$. Apply Lemma P.3 to the scalar and approximate fitted matrix widths and union bound over both families and all steps. This gives the potential event with failure probability at most $\delta/4$. Apply the argument of Lemma P.5 to the approximate fitted matrix widths. This gives the squared-potential event with failure probability at most $\delta/4$. On the intersection of the MLE, Bellman, and potential events, the proof of Lemma P.2 applies with the approximate fitted recommendation and radius $R_{\Theta,\gamma}$. Substituting the resulting Bellman and response bounds into Lemma P.1 gives the three conclusions of Theorem E.2 with the substitutions stated in the corollary. The direct interpolation and preservation arguments use the same fitted-QR identity and therefore give the corresponding optimization-tolerant conclusions. The total failure probability is at most $\delta/4 + \delta/4 + \delta/4 + \delta/4 = \delta$, which proves the corollary. □

Proposition M.4 (Response bounds and direct preservation certificate)**.** *On the event of Lemma M.3, for every $t \in [T]$, $h \in [H]$, $s \in \mathcal{S}$, and $p \in \Delta(\mathcal{A})$,*

$$\mathrm{TV}\Big(y_h^{t,p}(\cdot \mid s), \nu_{\theta_h^*,h}^{p,y_h^{t,p}}(\cdot \mid s)\Big) \le b_h^{t,\mathrm{F}}(s,p), \tag{95}$$

$$\mathrm{TV}\Big(\mu_{\hat{\theta}_h^t,h}^{p}(\cdot \mid s), \mu_{\theta_h^*,h}^{p}(\cdot \mid s)\Big) \le b_h^{t,\mathrm{Q}}(s,p). \tag{96}$$

$$\mathrm{TV}\Big(\nu_{\theta_h^*,h}^{p,y_h^{t,p}}(\cdot \mid s), \mu_{\theta_h^*,h}^{p}(\cdot \mid s)\Big) \le b_h^{t,\mathrm{P}}(s,p). \tag{97}$$

Moreover, with the interpolated parameter

$$\bar{\theta}_h^t := (1-\kappa)\theta_h^* + \kappa\hat{\theta}_h^t,$$

the true executed response obeys the interpolation identity

$$\nu_{\theta_h^*,h}^{p,y_h^{t,p}}(\cdot \mid s) = \mu_{\bar{\theta}_h^t,h}^{p}(\cdot \mid s). \tag{98}$$

Proof of Proposition M.4. Fix t, h, s, and p, and abbreviate the parameter estimation error as $\Delta_h^t := \hat{\theta}_h^t - \theta_h^*$ and the feature covariance matrix as $\Gamma := \Gamma_h^t(s,p)$. By the estimated-preservation identity (17),

$$y_h^{t,p}(\cdot \mid s) = \nu_{\hat{\theta}_h^t,h}^{p,y_h^{t,p}}(\cdot \mid s).$$

Applying Lemma L.5 within this recommendation-dependent exponential family gives

$$\mathrm{TV}\Big(y_h^{t,p}(\cdot \mid s), \nu_{\theta_h^*,h}^{p,y_h^{t,p}}(\cdot \mid s)\Big) \le L_\varepsilon \left\|\Delta_h^t\right\|_\Gamma.$$

The trace self-normalization inequality and (94) yield

$$\left\|\Delta_h^t\right\|_\Gamma^2 \le \mathrm{Tr}((V_h^t)^{-1}\Gamma)\left\|\Delta_h^t\right\|_{V_h^t}^2 \le R_\Theta^2[\rho_h^t(s,p)]^2.$$

Hence the first total-variation distance is at most $L_\varepsilon R_\Theta \rho_h^t(s,p)$, and taking the minimum with one proves (95). For the QR inequality, the fitted recommendation is itself $\mu^p_{\hat\theta^t_h,h}(\cdot \mid s)$, so its feature covariance is again Γ. Applying the same two lemmas with inverse temperature η gives

$$\mathrm{TV}\Big(\mu^p_{\hat\theta^t_h,h}(\cdot \mid s), \mu^p_{\theta^*_h,h}(\cdot \mid s)\Big) \le L_\eta R_\Theta \rho_h^t(s,p).$$

Truncation at one proves (96). For the fitted recommendation $y_h^{t,p} = \mu^p_{\hat\theta^t_h,h}$, the closed forms in (4) and (2), together with the identity $\varepsilon = \eta(1-\kappa)$, give

$$\nu^{p,y_h^{t,p}}_{\theta^*_h,h}(b \mid s) \propto \Big[\exp\Big\{\eta\phi_h^p(s,b)^\top \hat\theta_h^t\Big\}\Big]^\kappa \exp\big\{\eta(1-\kappa)\phi_h^p(s,b)^\top\theta_h^*\big\} \propto \exp\big\{\eta\phi_h^p(s,b)^\top \bar\theta_h^t\big\}.$$

Normalization proves (98). Convexity of Θ_h ensures $\bar\theta_h^t \in \Theta_h$.

Apply Lemma L.6 with $\theta = \hat\theta_h^t$, $\theta' = \theta_h^*$, $u = 1-\kappa$, and $v = 1$. The corresponding distributions are respectively the true executed response and the true QR, while the covariance Σ_0 is Γ. Equation (89), trace self-normalization, and (94) yield

$$\mathrm{TV}\Big(\nu^{p,y_h^{t,p}}_{\theta^*_h,h}(\cdot \mid s), \mu^p_{\theta^*_h,h}(\cdot \mid s)\Big) \le \kappa L_\eta R_\Theta \rho_h^t(s,p).$$

Truncation at one proves (97). □

N Covering the Generated Value Class

Set the terminal value class $\mathcal{W}_{H+1} := \{0\}$, whose logarithmic covering number is $\mathfrak{L}_{H+1} = 0$.

For fixed h, consider parameters

$$\omega \in \mathbb{R}^d, \quad A, C \in \mathbb{S}_+^d, \quad \theta \in \Theta_h, \quad 0 \preceq A, C \preceq I_d, \quad \|\omega\|_2 \le \sqrt{T} M_h.$$

Define the optimistic action-value function, QR policy, and its feature covariance matrix by

$$U_{\omega,A}(s,a,b) := \mathrm{Trun}_{[0,M_h]}\Big(\phi_h(s,a,b)^\top\omega + \beta_h^{\mathrm{U}}\sqrt{\phi_h(s,a,b)^\top A \phi_h(s,a,b)}\Big), \tag{99}$$

$$\mu^p_{\theta,h}(b \mid s) := \frac{\exp\{\eta\langle\phi_h^p(s,b),\theta\rangle\}}{\sum_{c\in\mathcal{B}}\exp\{\eta\langle\phi_h^p(s,c),\theta\rangle\}},$$

$$\Gamma^p_{\theta,h}(s) := \mathrm{Cov}_{b\sim\mu^p_{\theta,h}(\cdot|s)}[\phi_h^p(s,b)].$$

Define the planning objective and the corresponding value function by

$$G_{\omega,A,\theta,C}(s,p) := \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a)\mu^p_{\theta,h}(b \mid s) U_{\omega,A}(s,a,b) + M_h \min\Big\{1, L_\eta R_\Theta \sqrt{\mathrm{Tr}(C\Gamma^p_{\theta,h}(s))}\Big\}, \tag{100}$$

$$F_{\omega,A,\theta,C}(s) := \mathrm{Trun}_{[0,M_h]}\Big(\sup_{p\in\Delta(\mathcal{A})} G_{\omega,A,\theta,C}(s,p)\Big). \tag{101}$$

Let the generated value class $\mathcal{W}_h$ be the collection of these functions over the displayed parameter ranges.

Why recommendation and policy improvement add no hidden entropy. For every candidate leader distribution p, the fitted recommendation is the fixed-point response $y^p_{\theta,h} = \mu^p_{\theta,h}$, and consequently $\nu^{p,y^p_{\theta,h}}_{\theta,h} = y^p_{\theta,h}$. Thus recommendation is not an independent function-valued parameter: a net for θ controls the fitted recommendation uniformly over s and p through (105). Policy improvement also requires no cover of the possibly discontinuous or nonunique argmax selector. Indeed, for arbitrary objectives f, g on $\Delta(\mathcal{A})$,

$$\Big|\sup_{p\in\Delta(\mathcal{A})} f(p) - \sup_{p\in\Delta(\mathcal{A})} g(p)\Big| \le \sup_{p\in\Delta(\mathcal{A})} |f(p) - g(p)|. \tag{102}$$

The right-hand side is controlled uniformly in p by the finite-dimensional parameter nets below. Finally, only the QR bonus enters (100). The follower-response bonus is used later to control RRD and execution error, so it does not enlarge the class of values generated by the Bellman recursion.

Proposition N.1 (Covering number of the generated value class)**.** *Every $\hat{W}_h^t$ generated by PROM (Algorithm 1) with exact statewise planning belongs to $\mathcal{W}_h$, and, simultaneously for every step,*

$$\log N_\infty(\mathcal{W}_h, T^{-1}) \le \mathfrak{L}_h, \tag{103}$$

where every quantity on the right-hand side is explicit in (33)–(34).

Moreover, β_h^{U} depends only on the already defined future-step entropy $\mathfrak{L}_{h+1}$, so the entropy claim is noncircular.

Proposition N.1 controls the data dependence of the continuation values in Bellman regression. At step h, the regression target contains the random continuation value $\hat{W}_{h+1}^t$, which is computed from the same historical trajectories that form the regression design. A fixed-function self-normalized inequality therefore cannot be applied directly to $\hat{W}_{h+1}^t$. The proposition embeds every continuation value that the algorithm can generate in a deterministic class $\mathcal{W}_{h+1}$, covers that class before the data are realized, and permits a union bound over the resulting net.

Under the estimated preservation constraint, the planner has the fitted-QR form used in QSE learning, while its trajectories are generated by the ER. The fitted recommendation is determined by (θ, p), and (102) controls the maximization over p. These properties allow the same finite-dimensional parameterization to cover the continuation values generated by this planner.

Proof of Proposition N.1. First, $(\Lambda_h^t)^{-1}$ and $(V_h^t)^{-1}$ are positive semidefinite and bounded above by I_d. The ridge objective at its minimizer is no larger than at the zero vector. Since every regression target lies in $[0, M_h]$,

$$\left\|\hat{\omega}_h^t\right\|_2^2 \le \sum_{i=1}^{t-1} \left(u_h(s_h^i, a_h^i, b_h^i) + \hat{W}_{h+1}^t(s_{h+1}^i)\right)^2 \le T M_h^2,$$

and hence $\|\hat{\omega}_h^t\|_2 \le \sqrt{T} M_h$. Thus $\hat{W}_h^t$ has the form (101), with $A = (\Lambda_h^t)^{-1}$, $C = (V_h^t)^{-1}$, and $\theta = \hat{\theta}_h^t$.

We now establish uniform continuity in the finite-dimensional parameters. Truncation is one-Lipschitz, and for $\|x\|_2 \le 1$,

$$\left|\sqrt{x^\top A x} - \sqrt{x^\top A' x}\right| \le \sqrt{\|A - A'\|_{\mathrm{op}}}.$$

Therefore

$$\|U_{\omega,A} - U_{\omega',A'}\|_\infty \le \|\omega - \omega'\|_2 + \beta_h^{\mathrm{U}} \sqrt{\|A - A'\|_{\mathrm{op}}}. \tag{104}$$

For fixed p, s, the follower's induced reward vectors under θ and θ' differ by at most $\|\theta - \theta'\|_2$ in sup norm. Lemma L.3 therefore gives

$$\mathrm{TV}(\mu_{\theta,h}^p, \mu_{\theta',h}^p) \le \frac{\eta}{2} \|\theta - \theta'\|_2 . \tag{105}$$

For a fixed family $(X_b)_{b \in \mathcal{B}}$ with $\|X_b\|_2 \le 1$, changing its sampling distribution by ℓ_1 distance r changes the second moment by at most r in nuclear norm and the outer product of the mean by at most $2r$. Hence

$$\left\|\Gamma_{\theta,h}^p(s) - \Gamma_{\theta',h}^p(s)\right\|_* \le 3 \left\|\mu_{\theta,h}^p - \mu_{\theta',h}^p\right\|_1 \le 3\eta \|\theta - \theta'\|_2 .$$

For $0 \preceq C, C' \preceq I_d$,

$$\begin{aligned}\left|\mathrm{Tr}(C\Gamma_{\theta,h}^p) - \mathrm{Tr}(C'\Gamma_{\theta',h}^p)\right| &\le \|C - C'\|_{\mathrm{op}} \mathrm{Tr}(\Gamma_{\theta,h}^p) + \left\|\Gamma_{\theta,h}^p - \Gamma_{\theta',h}^p\right\|_* \\ &\le \|C - C'\|_{\mathrm{op}} + 3\eta \|\theta - \theta'\|_2 .\end{aligned} \tag{106}$$

Using $\left|\sqrt{x} - \sqrt{y}\right| \le \sqrt{|x - y|}$, the response-width term is therefore one-half Hölder in (C, θ).

Combining (104), (105), and (106), uniformly over s and p, shows that the objective in (100) changes by at most

$$\|\omega - \omega'\|_2 + \beta_h^{\mathrm{U}} \sqrt{\|A - A'\|_{\mathrm{op}}} + \frac{M_h \eta}{2} \|\theta - \theta'\|_2 + M_h L_\eta R_\Theta \sqrt{\|C - C'\|_{\mathrm{op}} + 3\eta \|\theta - \theta'\|_2}.$$

The inequality

$$\left|\sup_{p\in\Delta(\mathcal{A})} f(p) - \sup_{p\in\Delta(\mathcal{A})} g(p)\right| \le \sup_{p\in\Delta(\mathcal{A})} |f(p) - g(p)|$$

justifies taking the supremum without covering p.

Use mesh q_h for ω and mesh q_h^2 for θ, A, C. The continuity bound above and the definition of K_h^{cov} give function error at most

$$q_h + 3K_h^{\mathrm{cov}} q_h \le T^{-1}.$$

A Euclidean ball of radius R in $\mathbb{R}^d$ has an r-net of size at most $(1+2R/r)^d$. The symmetric matrices satisfying $0 \preceq A \preceq I$ lie in a Frobenius ball of radius $\sqrt{d}$ in dimension m_d, and thus have an r-net of size at most $(1+2\sqrt{d}/r)^{m_d}$. Multiplying the nets for ω, θ, A, C gives exactly the three displayed contributions defining $\mathfrak{L}_h$ in (34); the two symmetric-matrix nets are combined in the coefficient $2m_d$. This proves (103). Since $\mathfrak{L}_{H+1} = 0$ and the recursion defines β_h^{U} before $\mathfrak{L}_h$, no quantity depends on itself, even indirectly. □

O Uniform Bellman Confidence

For a bounded $W : \mathcal{S} \to [0, H-h]$, define the Bellman parameter by

$$\omega_h[W] := \vartheta_h^* + \sum_{s'\in\mathcal{S}} \mathbf{m}_h^*(\{s'\})W(s'). \tag{107}$$

The vector series satisfies

$$\sum_{s'\in\mathcal{S}} \|\mathbf{m}_h^*(\{s'\})W(s')\|_2 \le B_P \|W\|_\infty .$$

The displayed bound proves absolute convergence and hence makes $\omega_h[W]$ well defined. Consequently,

$$u_h(s,a,b) + (P_hW)(s,a,b) = \phi_h(s,a,b)^\top \omega_h[W], \qquad \|\omega_h[W]\|_2 \le B_\vartheta + (H-h)B_P.$$

Lemma O.1 (Uniform Bellman confidence)**.** *With probability at least $1-\delta/4$, simultaneously for all $t \in [T]$, $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $b \in \mathcal{B}$,*

$$\left|\phi_h(s,a,b)^\top \hat{\omega}_h^t - [u_h(s,a,b) + (P_h\hat{W}_{h+1}^t)(s,a,b)]\right| \le b_h^{t,\mathrm{U}}(s,a,b). \tag{108}$$

Proof of Lemma O.1. Fix h and a deterministic $W \in \mathcal{W}_{h+1}$, and define the ridge-regression estimator by

$$\hat{\omega}_h^t[W] := (\Lambda_h^t)^{-1} \sum_{i=1}^{t-1} \phi_h^i [u_h(s_h^i, a_h^i, b_h^i) + W(s_{h+1}^i)].$$

The centered transition noise is

$$\zeta_h^i(W) := W(s_{h+1}^i) - (P_hW)(s_h^i, a_h^i, b_h^i).$$

By (29), $\mathbb{E}[\zeta_h^i(W) \mid \mathcal{H}_h^i] = 0$. Since $0 \le W \le M_{h+1} = H-h$, Hoeffding's lemma makes $\zeta_h^i(W)$ conditionally $(H-h)/2$-sub-Gaussian under our stated moment-generating-function convention. We use this valid proxy without further relaxation. In particular, there is no martingale-noise contribution at the terminal step.

To make the ridge contribution explicit, use (107) and the identity

$$\hat{\omega}_h^t[W] - \omega_h[W] = (\Lambda_h^t)^{-1} \left(\sum_{i=1}^{t-1} \phi_h^i \zeta_h^i(W) - \omega_h[W] \right). \tag{109}$$

Indeed, the conditional mean of the regression target is $(\phi_h^i)^\top \omega_h[W]$, while $\Lambda_h^t = I_d + \sum_{i=1}^{t-1} \phi_h^i (\phi_h^i)^\top$. Taking the Λ_h^t-norm in (109) and applying the triangle inequality yields

$$\left\|\hat{\omega}_h^t[W] - \omega_h[W]\right\|_{\Lambda_h^t} \le \left\|\sum_{i=1}^{t-1} \phi_h^i \zeta_h^i(W)\right\|_{(\Lambda_h^t)^{-1}} + \|\omega_h[W]\|_{(\Lambda_h^t)^{-1}} .$$

Because $\Lambda_h^t \succeq I_d$, the second term is at most $\|\omega_h[W]\|_2 \le B_\vartheta + (H-h)B_P$.

For fixed h, define the inter-episode filtration $\mathcal{K}_i^{(h)} := \mathcal{H}_h^{i+1}$ for $i \ge 0$. The inclusions $\mathcal{H}_h^i \subseteq \mathcal{F}_h^i \subseteq \mathcal{F}_H^i = \mathcal{F}_0^{i+1} \subseteq \mathcal{H}_h^{i+1}$ show that $\{\mathcal{K}_i^{(h)}\}_{i\ge 0}$ is a filtration. The design vector $\phi_h^i = \phi_h(s_h^i, a_h^i, b_h^i)$ is $\mathcal{K}_{i-1}^{(h)}$-measurable, whereas $\zeta_h^i(W)$ is $\mathcal{K}_i^{(h)}$-measurable and satisfies $\mathbb{E}[\zeta_h^i(W) \mid \mathcal{K}_{i-1}^{(h)}] = 0$ for every fixed continuation function W. The conditional sub-Gaussian bound established above is therefore also valid relative to $\mathcal{K}_{i-1}^{(h)}$. Hence the pairs $(\phi_h^i, \zeta_h^i(W))$ have exactly the predictable-design martingale structure required by the self-normalized inequality. This predictable-design argument parallels the proof of Jin et al. (2021c, Lemma 5.2), which uses their data-compliance condition. We handle the data dependence of the continuation value through a uniform bound over the generated value class. The self-normalized martingale inequality gives, uniformly in t,

$$\left\| \sum_{i=1}^{t-1} \phi_h^i \zeta_h^i(W) \right\|_{(\Lambda_h^t)^{-1}} \le \frac{H-h}{2} \sqrt{2\left[d\log(1+T/d) + \log(1/\delta')\right]}. \tag{110}$$

For each step, fix a deterministic T^{-1}-net $\mathcal{N}_{h+1}$ of $\mathcal{W}_{h+1}$ with $|\mathcal{N}_{h+1}| \le \exp\{\mathfrak{L}_{h+1}\}$, and set the failure probability for each net point to $\delta_h' := \delta \exp\{-\mathfrak{L}_{h+1}\}/(16H)$. Apply (110) with $\delta' = \delta_h'$ to every net point and union bound over the net and over steps. Since the self-normalized inequality is already uniform in t, no episode-wise union bound is needed. The resulting failure probability is at most $\sum_{h=1}^H |\mathcal{N}_{h+1}|\, \delta_h' \le \delta/16$, which is within the $\delta/4$ budget of the lemma, and $\log(1/\delta_h') = \mathfrak{L}_{h+1} + \log(16H/\delta)$. Let W^0 be a nearest net point to an arbitrary W. For fixed h and t, define the design matrix $\Phi := [\phi_h^1, \dots, \phi_h^{t-1}] \in \mathbb{R}^{d\times(t-1)}$, so that $\Lambda_h^t = I_d + \Phi\Phi^\top$. The inequality $\Phi^\top(\Lambda_h^t)^{-1}\Phi \preceq I_{t-1}$ gives

$$\left\|\hat\omega_h^t[W] - \hat\omega_h^t[W^0]\right\|_{\Lambda_h^t} \le \sqrt{T}\left\|W - W^0\right\|_\infty. \tag{111}$$

Furthermore,

$$\left\|\omega_h[W] - \omega_h[W^0]\right\|_2 \le B_P \left\|W - W^0\right\|_\infty,$$

and $\lambda_{\max}(\Lambda_h^t) \le T+1$, so

$$\left\|\omega_h[W] - \omega_h[W^0]\right\|_{\Lambda_h^t} \le \sqrt{T+1} B_P \left\|W - W^0\right\|_\infty. \tag{112}$$

When $h = H$, $\mathcal{W}_{H+1} = \{0\}$, so no net extension is needed. When $h < H$, a net radius of T^{-1} makes the sum of the estimator and target extensions in (111)– (112) exactly bounded by

$$T^{-1/2} + \frac{\sqrt{T+1}}{T} B_P = \chi_{T,h}.$$

Combining the net-point self-normalized bound, ridge bias, and these extensions yields

$$\left\|\hat\omega_h^t[W] - \omega_h[W]\right\|_{\Lambda_h^t} \le \beta_h^{\mathrm{U}}.$$

Cauchy–Schwarz in the Λ_h^t-norm gives the desired pointwise bound. The step-$(h+1)$ covering logarithm is $\mathfrak{L}_{h+1}$ by (34). Finally substitute the random $W = \hat W_{h+1}^t$, which belongs to the deterministic envelope by Proposition N.1. □

P Performance Difference, Optimism, and Regret

Lemma P.1 (Performance difference with a response estimate)**.** *Fix a committed pair (π, y) and let $\nu_h = \nu_{\theta_h^*,h}^{\pi,y}$. Let $\bar\nu = \{\bar\nu_h : \mathcal{S} \to \Delta(\mathcal{B})\}_{h\in[H]}$ be any follower policy, let $\bar U_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \mathbb{R}$ and $\bar W_h : \mathcal{S} \to \mathbb{R}$ be bounded functions for every $h \in [H]$, and set $\bar W_{H+1} \equiv 0$. Then*

$$\begin{aligned}\mathbb{E}_{s_1\sim\rho_0}[\bar W_1(s_1)] - J(\pi, y) = &\sum_{h=1}^H \mathbb{E}^{\pi,\nu}\left[(\bar U_h - u_h)(s_h, a_h, b_h) - \bar W_{h+1}(s_{h+1})\right] \\ &+ \sum_{h=1}^H \mathbb{E}^{\pi,\nu}\left[\bar W_h(s_h) - (\pi_h \otimes \bar\nu_h)\circ \bar U_h(s_h)\right] + \sum_{h=1}^H \mathbb{E}^{\pi,\nu}\left[\pi_h \otimes (\bar\nu_h - \nu_h)\circ \bar U_h(s_h)\right]. \end{aligned} \tag{113}$$

Proof of Lemma P.1. For every step h, condition first on the current state s_h. Under the executed trajectory law, (a_h, b_h) has conditional distribution $\pi_h \otimes \nu_h$. Therefore the three summands on the right-hand side of (113) at step h add up to

$$\begin{aligned}
&\mathbb{E}^{\pi,\nu}[\bar{U}_h(s_h, a_h, b_h) - u_h(s_h, a_h, b_h) - \bar{W}_{h+1}(s_{h+1})] \\
&\quad + \mathbb{E}^{\pi,\nu}[\bar{W}_h(s_h)] - \mathbb{E}^{\pi,\nu}[(\pi_h \otimes \bar{\nu}_h) \circ \bar{U}_h(s_h)] \\
&\quad + \mathbb{E}^{\pi,\nu}[(\pi_h \otimes \bar{\nu}_h) \circ \bar{U}_h(s_h)] - \mathbb{E}^{\pi,\nu}[(\pi_h \otimes \nu_h) \circ \bar{U}_h(s_h)] \\
&= \mathbb{E}^{\pi,\nu}[\bar{W}_h(s_h) - u_h(s_h, a_h, b_h) - \bar{W}_{h+1}(s_{h+1})].
\end{aligned}$$

The last equality uses the conditional-expectation identity

$$\mathbb{E}^{\pi,\nu}[\bar{U}_h(s_h, a_h, b_h)] = \mathbb{E}^{\pi,\nu}[(\pi_h \otimes \nu_h) \circ \bar{U}_h(s_h)].$$

Summing over $h = 1, \dots, H$ makes the estimated value terms telescope:

$$\sum_{h=1}^{H} \mathbb{E}^{\pi,\nu}[\bar{W}_h(s_h) - \bar{W}_{h+1}(s_{h+1})] = \mathbb{E}_{s_1 \sim \rho_0}[\bar{W}_1(s_1)],$$

because $\bar{W}_{H+1} \equiv 0$. The remaining reward sum is $J(\pi, y) = \mathbb{E}^{\pi,\nu}[\sum_{h=1}^{H} u_h(s_h, a_h, b_h)]$, which proves (113). □

Recall the optimal dynamic-programming values from Lemma 3.8:

$$\begin{aligned}
W^*_{H+1} &\equiv 0, \\
U^*_h(s, a, b) &:= u_h(s, a, b) + (P_h W^*_{h+1})(s, a, b), \\
W^*_h(s) &:= \max_{p \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) \mu^{p}_{\theta^*_h, h}(b \mid s) U^*_h(s, a, b).
\end{aligned}$$

Because the follower is myopic, standard backward induction gives $\mathbb{E}_{\rho_0} W^*_1 = J^*$.

Lemma P.2 (Optimism)**.** *On the intersection of the events in Lemmas O.1 and M.3, PROM (Algorithm 1) satisfies*

$$\hat{W}^t_h(s) \geq W^*_h(s) \qquad \text{for every } t \in [T],\ h \in [H],\ s \in \mathcal{S}.$$

Proof of Lemma P.2. Use backward induction. At step $H+1$, both sides are zero. Suppose $\hat{W}^t_{h+1} \geq W^*_{h+1}$. By Lemma O.1,

$$\phi_h^\top \hat{\omega}^t_h + b^{t,\mathrm{U}}_h \geq u_h + P_h \hat{W}^t_{h+1} \geq u_h + P_h W^*_{h+1}.$$

The true target lies in $[0, M_h]$, so truncation preserves the lower bound and $\hat{U}^t_h \geq U^*_h$.

Fix s and evaluate $\mathcal{Q}^t_h(s, p)$ at any p. By (96), changing the follower distribution from the true QR to the estimated QR changes the expectation of a function in $[0, M_h]$ by at most $M_h b^{t,\mathrm{Q}}_h(s, p)$. Therefore

$$\mathcal{Q}^t_h(s, p) \geq \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) \mu^{p}_{\theta^*_h, h}(b \mid s) U^*_h(s, a, b).$$

Taking the maximum of each side over $p \in \Delta(\mathcal{A})$ and using (20) and (10) gives $\mathcal{Q}^t_h(s, \pi^t_h(\cdot \mid s)) \geq W^*_h(s)$. Monotone truncation therefore gives $\hat{W}^t_h(s) \geq W^*_h(s)$. □

Lemma P.3 (Episode-start expectation conversion)**.** *Let X^t be $\mathcal{F}_0^{t+1}$-measurable and satisfy $0 \leq X^t \leq 1$ almost surely for every $t \in [T]$. Then, for every $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,*

$$\sum_{t=1}^{T} \mathbb{E}^t X^t \leq \sum_{t=1}^{T} X^t + \sqrt{2T \log(1/\alpha)}. \tag{114}$$

Proof of Lemma P.3. Define the martingale differences by

$$M^t := \mathbb{E}^t X^t - X^t.$$

The variable M^t is $\mathcal{F}_0^{t+1}$-measurable and satisfies

$$\mathbb{E}[M^t \mid \mathcal{F}_0^t] = 0.$$

Moreover, $|M^t| \le 1$ because both X^t and $\mathbb{E}^t X^t$ belong to $[0,1]$. Thus $\{M^t\}_{t=1}^T$ is a martingale-difference sequence with increments bounded by one. The one-sided Azuma–Hoeffding inequality gives

$$\mathbb{P}\left(\sum_{t=1}^T M^t > \sqrt{2T\log(1/\alpha)}\right) \le \alpha.$$

Substituting the definition of M^t proves (114). □

Lemma P.4 (Expected scalar and matrix potentials)**.** *With probability at least $1-\delta/4$, for every h,*

$$\sum_{t=1}^T \mathbb{E}^t \sqrt{\phi_h(s_h^t,a_h^t,b_h^t)^\top (\Lambda_h^t)^{-1}\phi_h(s_h^t,a_h^t,b_h^t)} \le \mathfrak{C}_{T,d,H}(\delta), \tag{115}$$

$$\sum_{t=1}^T \mathbb{E}^t \rho_h^t(s_h^t,\pi_h^t) \le \mathfrak{C}_{T,d,H}(\delta). \tag{116}$$

Here $\mathbb{E}^t$ is the episode-start conditional expectation defined in (11).

Proof of Lemma P.4. Fix $h \in [H]$ and define the realized scalar width

$$z_h^t := \sqrt{\phi_h(s_h^t,a_h^t,b_h^t)^\top (\Lambda_h^t)^{-1}\phi_h(s_h^t,a_h^t,b_h^t)}.$$

Because $\Lambda_h^t \succeq I_d$ and $\|\phi_h(s,a,b)\|_2 \le 1$, one has $0 \le z_h^t \le 1$. The determinant lemma and the update

$$\Lambda_h^{t+1} = \Lambda_h^t + \phi_h(s_h^t,a_h^t,b_h^t)\phi_h(s_h^t,a_h^t,b_h^t)^\top$$

show that

$$\sum_{t=1}^T (z_h^t)^2 \le 2\log\frac{\det \Lambda_h^{T+1}}{\det I_d} \le 2d\log(1+T/d).$$

Cauchy–Schwarz gives

$$\sum_{t=1}^T z_h^t \le \sqrt{2Td\log(1+T/d)}.$$

Apply Lemma P.3 to $X^t = z_h^t$ with $\alpha = \delta/(16H)$. With probability at least $1-\delta/(16H)$,

$$\sum_{t=1}^T \mathbb{E}^t z_h^t \le \sqrt{2Td\log(1+T/d)} + \sqrt{2T\log(16H/\delta)} = \mathfrak{C}_{T,d,H}(\delta).$$

For the response widths, define the feature covariance matrix and its normalized version by

$$X_h^t := \Gamma_h^t(s_h^t,\pi_h^t), \qquad Y_h^t := (V_h^t)^{-1/2} X_h^t (V_h^t)^{-1/2}.$$

The estimated-preservation identity implies

$$V_h^{t+1} = V_h^t + X_h^t.$$

Moreover, $X_h^t \succeq 0$, $\mathrm{Tr}(X_h^t) \le 1$, and $V_h^t \succeq I_d$. Hence every eigenvalue of Y_h^t belongs to $[0,1]$. For every $\lambda \in [0,1]$, $\log(1+\lambda) \ge \lambda/2$, so

$$\log\det(I_d + Y_h^t) \ge \frac{1}{2}\mathrm{Tr}(Y_h^t).$$

Using the determinant lemma and telescoping gives

$$\begin{aligned}\sum_{t=1}^T [\rho_h^t(s_h^t,\pi_h^t)]^2 &= \sum_{t=1}^T \mathrm{Tr}((V_h^t)^{-1}X_h^t) \le 2\sum_{t=1}^T \log\frac{\det V_h^{t+1}}{\det V_h^t} \\ &= 2\log\det V_h^{T+1} \le 2d\log(1+T/d).\end{aligned}$$

Cauchy–Schwarz gives

$$\sum_{t=1}^{T} \rho_h^t(s_h^t, \pi_h^t) \le \sqrt{2Td\log(1+T/d)}.$$

The response widths also lie in $[0,1]$, so Lemma P.3 with $\alpha = \delta/(16H)$ yields

$$\sum_{t=1}^{T} \mathbb{E}^t \rho_h^t(s_h^t, \pi_h^t) \le \mathfrak{C}_{T,d,H}(\delta)$$

with failure probability at most $\delta/(16H)$. A union bound over the H steps and the two width families has failure probability at most $\delta/8$, which is smaller than the stated budget $\delta/4$. This proves (115) and (116) simultaneously for every $h \in [H]$. □

Lemma P.5 (Predictable squared matrix potential)**.** *With probability at least $1-\delta/4$, simultaneously for every $h \in [H]$,*

$$\sum_{t=1}^{T} \mathbb{E}^t\left[\left(\rho_h^t(s_h^t, \pi_h^t)\right)^2\right] \le \mathfrak{S}_{T,d,H}(\delta). \tag{117}$$

Proof of Lemma P.5. Fix h, let the squared response width be $X_t := [\rho_h^t(s_h^t, \pi_h^t)]^2 \in [0,1]$, and let its conditional mean be $m_t := \mathbb{E}^t X_t$. Convexity of $x \mapsto e^{-x}$ on $[0,1]$ gives

$$\mathbb{E}^t e^{-X_t} \le 1 - (1-e^{-1})m_t \le \exp\{-(1-e^{-1})m_t\}.$$

Consequently,

$$\exp\left\{(1-e^{-1})\sum_{i=1}^{t} m_i - \sum_{i=1}^{t} X_i\right\}$$

is a nonnegative supermartingale. Markov's inequality applied to the terminal value of this nonnegative supermartingale implies that, with probability at least $1-\alpha$,

$$\sum_{t=1}^{T} m_t \le \frac{1}{1-e^{-1}}\left(\sum_{t=1}^{T} X_t + \log(1/\alpha)\right) \le 2\sum_{t=1}^{T} X_t + 2\log(1/\alpha).$$

The realized determinant-potential calculation in the proof of Lemma P.4 gives $\sum_t X_t \le 2d\log(1+T/d)$ deterministically. Taking $\alpha = \delta/(4H)$ and applying a union bound over $h \in [H]$ proves (117). □

Proof of Theorem E.2. Let $\mathcal{E}_{\mathrm{MLE}}$ and $\mathcal{E}_{\mathrm{Bell}}$ denote the simultaneous events in Lemma M.3 and Lemma O.1, respectively. Let $\mathcal{E}_{\mathrm{pot}}$ and $\mathcal{E}_{\mathrm{sq}}$ denote the simultaneous events in Lemma P.4 and Lemma P.5, respectively. Their failure probabilities are each at most $\delta/4$, so

$$\mathbb{P}(\mathcal{E}_{\mathrm{MLE}} \cap \mathcal{E}_{\mathrm{Bell}} \cap \mathcal{E}_{\mathrm{pot}} \cap \mathcal{E}_{\mathrm{sq}}) \ge 1-\delta.$$

We work on this intersection.

By Lemma P.2, $\hat{W}_1^t(s) \ge W_1^*(s)$ for every episode t and state s. Since $J^* = \mathbb{E}_{\rho_0}[W_1^*(s_1)]$, this gives

$$J^* - J(\pi^t, y^t) \le \mathbb{E}_{\rho_0}[\hat{W}_1^t(s_1)] - J(\pi^t, y^t).$$

Fix episode t and apply Lemma P.1 with $\bar{\nu}_h = y_h^t$, $\bar{U}_h = \hat{U}_h^t$, and $\bar{W}_h = \hat{W}_h^t$, while the executed response is $\nu_h^t := \nu_{\theta_h^*,h}^{\pi^t,y^t}$. We bound the three groups in (113) separately.

For the Bellman group, conditional expectation over the next state gives

$$\mathbb{E}^t[(\hat{U}_h^t - u_h)(s_h^t, a_h^t, b_h^t) - \hat{W}_{h+1}^t(s_{h+1}^t)] = \mathbb{E}^t[\hat{U}_h^t(s_h^t, a_h^t, b_h^t) - g_h^t(s_h^t, a_h^t, b_h^t)],$$

where the Bellman target function is

$$g_h^t(s,a,b) := u_h(s,a,b) + (P_h\hat{W}_{h+1}^t)(s,a,b) \in [0, M_h].$$

On $\mathcal{E}_{\mathrm{Bell}}$, (108) implies $\left|\phi_h^\top \hat{\omega}_h^t - g_h^t\right| \le b_h^{t,\mathrm{U}}$. Since $\hat{U}_h^t = \mathrm{Trun}_{[0,M_h]}\left(\phi_h^\top \hat{\omega}_h^t + b_h^{t,\mathrm{U}}\right)$ and $g_h^t \in [0, M_h]$,

$$\hat{U}_h^t - g_h^t \le \phi_h^\top \hat{\omega}_h^t + b_h^{t,\mathrm{U}} - g_h^t \le 2b_h^{t,\mathrm{U}}.$$

Thus the Bellman group is at most $2\mathbb{E}^t[b_h^{t,\mathrm{U}}(s_h^t, a_h^t, b_h^t)]$.

For the policy-improvement group, the fitted recommendation satisfies $y_h^t(\cdot \mid s) = y_h^{t,\pi_h^t}(\cdot \mid s)$, and hence

$$\mathcal{Q}_h^t(s, \pi_h^t) = (\pi_h^t \otimes y_h^t) \circ \hat{U}_h^t(s) + M_h b_h^{t,\mathrm{Q}}(s, \pi_h^t).$$

The definition of $\hat{W}_h^t$ and the inequality $\mathrm{Trun}_{[0,M]}(x + c) - x \le c$ for $x \in [0, M]$ and $c \ge 0$ give

$$\hat{W}_h^t(s) - (\pi_h^t \otimes y_h^t) \circ \hat{U}_h^t(s) \le M_h b_h^{t,\mathrm{Q}}(s, \pi_h^t).$$

For the response-estimation group, $\hat{U}_h^t(s, a, b) \in [0, M_h]$. The variational characterization of total variation and (95) therefore imply

$$\pi_h^t \otimes (y_h^t - \nu_h^t) \circ \hat{U}_h^t(s) \le M_h \,\mathrm{TV}(y_h^t(\cdot \mid s), \nu_h^t(\cdot \mid s)) \le M_h b_h^{t,\mathrm{F}}(s, \pi_h^t).$$

Combining the preceding bounds and summing over episodes and steps yields

$$\mathrm{Reg}(T) \le 2 \sum_{h=1}^{H} \beta_h^{\mathrm{U}} \sum_{t=1}^{T} \mathbb{E}^t \sqrt{\phi_h(s_h^t, a_h^t, b_h^t)^\top (\Lambda_h^t)^{-1} \phi_h(s_h^t, a_h^t, b_h^t)}$$
$$+ \sum_{h=1}^{H} M_h R_\Theta (L_\varepsilon + L_\eta) \sum_{t=1}^{T} \mathbb{E}^t \rho_h^t(s_h^t, \pi_h^t).$$

Applying (115) and (116) proves (40).

Finally, (95) gives, for each episode,

$$\mathrm{RRD}(\pi^t, y^t; \theta^*) \le L_\varepsilon R_\Theta \sum_{h=1}^{H} \mathbb{E}^t \rho_h^t(s_h^t, \pi_h^t), \quad \mathrm{RRD}_{\mathrm{w}}(\pi^t, y^t; \theta^*) \le L_\varepsilon R_\Theta \sum_{h=1}^{H} M_h \mathbb{E}^t \rho_h^t(s_h^t, \pi_h^t).$$

Summing over t and applying the matrix-potential bound for every h gives

$$\mathrm{RRD}_{\mathrm{cum}}(T) \le H L_\varepsilon R_\Theta \mathfrak{C}_{T,d,H}(\delta), \qquad \mathrm{RRD}_{\mathrm{w,cum}}(T) \le L_\varepsilon R_\Theta \mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h.$$

This proves (41) and (42). □

Proof of Corollary E.3. On the event used in the proof of Theorem E.2, (96) gives

$$\mathrm{TV}\left(y_h^t(\cdot \mid s_h^t), \mu_{\theta_h^*,h}^{\pi^t}(\cdot \mid s_h^t)\right) \le L_\eta R_\Theta \rho_h^t(s_h^t, \pi_h^t).$$

Likewise, the direct preservation certificate in (97) gives

$$\mathrm{TV}\left(\nu_{\theta_h^*,h}^{\pi^t, y^t}(\cdot \mid s_h^t), \mu_{\theta_h^*,h}^{\pi^t}(\cdot \mid s_h^t)\right) \le \kappa L_\eta R_\Theta \rho_h^t(s_h^t, \pi_h^t).$$

Take $\mathbb{E}^t$, sum over episodes and steps, and apply (116) for each h. This proves (43) and (44); inserting M_h proves (45).

□

Proof of Corollary E.4. Fix t, h, condition on $\mathcal{F}_0^t$, and abbreviate the parameter difference as $\Delta_h^t := \theta_h^* - \hat{\theta}_h^t$ and the feature covariance matrix as $\Gamma := \Gamma_h^t(s_h^t, \pi_h^t)$. The interpolation identity (98) and Lemma L.6 with $u = 1 - \kappa$ and $v = 1$ give

$$\mathrm{KL}\left(\nu_{\theta_h^*,h}^{\pi^t, y^t}(\cdot \mid s_h^t) \middle\| \mu_{\theta_h^*,h}^{\pi^t}(\cdot \mid s_h^t)\right) \le \frac{\eta^2 \kappa^2 e^{2\eta}}{2} \left\|\Delta_h^t\right\|_\Gamma^2.$$

On the MLE event, trace self-normalization yields

$$\left\|\Delta_h^t\right\|_\Gamma^2 \le R_\Theta^2 [\rho_h^t(s_h^t, \pi_h^t)]^2.$$

Divide by η and apply the criterion identity (7). After taking $\mathbb{E}^t$ and summing over t, h, (117) proves (46). □

Proof of Corollary E.12. By the episode-start definitions,

$$\mathbb{E}^t[G^t] = J(\pi^t, y^t), \quad \mathbb{E}^t[C^t] = \mathrm{RRD}(\pi^t, y^t; \theta^*), \quad \mathbb{E}^t[C_{\mathrm{w}}^t] = \mathrm{RRD}_{\mathrm{w}}(\pi^t, y^t; \theta^*).$$

Define the martingale differences for the realized leader return, RRD, and weighted RRD by

$$M_{\mathrm{G}}^t := J(\pi^t, y^t) - G^t, \quad M_{\mathrm{C}}^t := C^t - \mathrm{RRD}(\pi^t, y^t; \theta^*), \quad M_{\mathrm{W}}^t := C_{\mathrm{w}}^t - \mathrm{RRD}_{\mathrm{w}}(\pi^t, y^t; \theta^*).$$

Each variable is $\mathcal{F}_0^{t+1}$-measurable and has conditional mean zero given $\mathcal{F}_0^t$. Thus, each of the three sequences is a martingale-difference sequence with respect to $\{\mathcal{F}_0^{t+1}\}_{t\geq 0}$. Since $G^t \in [0, H]$, the conditional support of M_{G}^t is an interval of length at most H. Since $C^t \in [0, H]$, the conditional support of M_{C}^t is an interval of length at most H. Since $C_{\mathrm{w}}^t \in [0, S_H]$, the conditional support of M_{W}^t is an interval of length at most S_H. The conditional Hoeffding lemma therefore gives the sub-Gaussian proxies $H/2$, $H/2$, and $S_H/2$, respectively. Applying the one-sided Azuma–Hoeffding inequality with failure probability $\delta_{\mathrm{traj}}/3$ to the first sequence gives

$$\sum_{t=1}^{T} M_{\mathrm{G}}^t \leq \frac{H}{2}\sqrt{2T \log \frac{3}{\delta_{\mathrm{traj}}}}.$$

Moreover,

$$\mathrm{Reg}_{\mathrm{traj}}(T) = \mathrm{Reg}(T) + \sum_{t=1}^{T} M_{\mathrm{G}}^t,$$

which proves (56). Applying the same argument to $\{M_{\mathrm{C}}^t\}_{t=1}^T$ gives

$$\sum_{t=1}^{T} M_{\mathrm{C}}^t \leq \frac{H}{2}\sqrt{2T \log \frac{3}{\delta_{\mathrm{traj}}}},$$

and the identity

$$\mathrm{RRD}_{\mathrm{traj}}(T) = \mathrm{RRD}_{\mathrm{cum}}(T) + \sum_{t=1}^{T} M_{\mathrm{C}}^t$$

proves (57). Finally,

$$\sum_{t=1}^{T} M_{\mathrm{W}}^t \leq \frac{S_H}{2}\sqrt{2T \log \frac{3}{\delta_{\mathrm{traj}}}}$$

with failure probability at most $\delta_{\mathrm{traj}}/3$, and

$$\mathrm{RRD}_{\mathrm{w,traj}}(T) = \mathrm{RRD}_{\mathrm{w,cum}}(T) + \sum_{t=1}^{T} M_{\mathrm{W}}^t.$$

A union bound over the three martingale events has total failure probability at most δ_{traj} and proves all three inequalities (56)– (58) simultaneously. □

Proof of Corollary E.16. Fix a phase j and condition on its starting sigma-field. Within this phase, write $r \in [T_j]$ for the local episode index, $\mathbb{E}_j^r$ for conditional expectation at the start of local episode r, and $z_{j,h}^r$ and $\rho_{j,h}^r$ for the realized scalar and matrix widths from Lemma P.4. Both widths lie in $[0, 1]$. The determinant arguments in that lemma hold deterministically at every local prefix and give, for every $k \in [T_j]$,

$$\sum_{r=1}^{k} z_{j,h}^r \leq \sqrt{2kd \log(1 + T_j/d)}, \qquad \sum_{r=1}^{k} \rho_{j,h}^r \leq \sqrt{2kd \log(1 + T_j/d)}.$$

For either width sequence $q_{j,h}^r \in \{z_{j,h}^r, \rho_{j,h}^r\}$, $\mathbb{E}_j^r q_{j,h}^r - q_{j,h}^r$ is a martingale difference bounded by one. The one-sided maximal Azuma–Hoeffding inequality therefore implies

$$\mathbb{P}\left(\max_{k\in[T_j]} \sum_{r=1}^{k} (\mathbb{E}_j^r q_{j,h}^r - q_{j,h}^r) > \sqrt{2T_j \log(8H/\delta_j)} \,\middle|\, \mathcal{F}_{j,0} \right) \leq \frac{\delta_j}{8H},$$

where $\mathcal{F}_{j,0}$ is the phase-start sigma-field. A union bound over the H steps and the two width families yields an event $\mathcal{E}_{j,\mathrm{pot}}$ whose conditional failure probability is at most $\delta_j/4$ and on which, simultaneously for every h and $k \in [T_j]$,

$$\sum_{r=1}^{k} \mathbb{E}_j^r z_{j,h}^r \le \sqrt{2kd\log(1+T_j/d)} + \sqrt{2T_j\log(8H/\delta_j)}, \tag{118}$$

$$\sum_{r=1}^{k} \mathbb{E}_j^r \rho_{j,h}^r \le \sqrt{2kd\log(1+T_j/d)} + \sqrt{2T_j\log(8H/\delta_j)}. \tag{119}$$

Let $\mathcal{E}_{j,\mathrm{MLE}}$ and $\mathcal{E}_{j,\mathrm{Bell}}$ be the phasewise MLE and Bellman events. For exact minimization, these are the events used in the proof of Theorem E.2 with T and δ replaced by T_j and δ_j. For a certified negative-log-likelihood gap, these are the recomputed events from Corollary E.5 with T, δ, and γ_{MLE} replaced by T_j, δ_j, and γ_j^{MLE}. The finite phase-start history takes values in a standard Borel space, so a regular conditional law given $\mathcal{F}_{j,0}$ exists. Under each such conditional law, the realized history at the start of the phase is fixed, while the local policies, likelihood contexts, and Bellman design vectors remain predictable with respect to the corresponding within-phase filtrations. The conditional joint-action factorization stated in Section D.1 and the transition conditional mean in (29) remain valid within the phase. Consequently, the square-root likelihood-ratio martingales and the self-normalized Bellman martingales retain their martingale properties under the conditional law. Ville's inequality and the self-normalized inequality are uniform over the local prefix, so the phasewise events satisfy, almost surely,

$$\mathbb{P}(\mathcal{E}_{j,\mathrm{MLE}}^c \mid \mathcal{F}_{j,0}) \le \frac{\delta_j}{4}, \qquad \mathbb{P}(\mathcal{E}_{j,\mathrm{Bell}}^c \mid \mathcal{F}_{j,0}) \le \frac{\delta_j}{4}.$$

Under the fixed-parameter convention of Corollary E.6, (62) keeps the phasewise response confidence radius at order $\tilde{\mathcal{O}}(\sqrt{d})$ and the logarithmic covering bounds at order $\tilde{\mathcal{O}}(d^2)$. Define the phase event $\mathcal{E}_j := \mathcal{E}_{j,\mathrm{MLE}} \cap \mathcal{E}_{j,\mathrm{Bell}} \cap \mathcal{E}_{j,\mathrm{pot}}$. The preceding failure allocations imply

$$\mathbb{P}(\mathcal{E}_j^c \mid \mathcal{F}_{j,0}) \le \frac{3\delta_j}{4} \le \delta_j.$$

Taking expectations removes the conditioning. Since

$$\sum_{j=0}^{\infty} \delta_j = \frac{6\delta}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} = \delta,$$

a countable union bound yields $\mathbb{P}(\cap_{j\ge 0}\mathcal{E}_j) \ge 1 - 3\delta/4 \ge 1 - \delta$.

For a phase j and a local prefix $k \in [T_j]$, define the bound on cumulative widths over the phase prefix by

$$\mathfrak{C}_{j,k} := \sqrt{2kd\log(1+T_j/d)} + \sqrt{2T_j\log(8H/\delta_j)}.$$

Let $R_{\Theta,j}$ and $\beta_{j,h}^{\mathrm{U}}$ denote the response radius and Bellman radius recomputed for phase j. For exact likelihood minimization, $R_{\Theta,j}$ is the exact phasewise radius. For the certified approximate solver, $R_{\Theta,j}$ is the radius obtained by substituting γ_j^{MLE} in Corollary E.5. Define the standard leader regret over the phase prefix by

$$\mathrm{Reg}_j(k) := \sum_{r=1}^{k} \left[J^* - J(\pi^{t_j(r)}, y^{t_j(r)}) \right],$$

and define the cumulative RRD $\mathrm{RRD}_{j,\mathrm{cum}}(k)$, weighted cumulative RRD $\mathrm{RRD}_{j,\mathrm{w,cum}}(k)$, and absolute leader regret $\mathrm{Reg}_j^{\mathrm{abs}}(k)$ by restricting the corresponding cumulative criteria to the first k local episodes of phase j. On $\mathcal{E}_j$, the MLE event gives the two phasewise response-confidence bounds, the Bellman event gives the pointwise Bellman bound for every local episode, and (118)–(119) control the two expected width sums at every prefix. Applying Lemma P.1 and the same

three-group decomposition used in the proof of Theorem E.2 therefore gives

$$\begin{aligned}\mathrm{Reg}_j(k) &\le 2\mathfrak{C}_{j,k}\sum_{h=1}^{H}\beta_{j,h}^{\mathrm{U}} + R_{\Theta,j}(L_\eta+L_\varepsilon)\mathfrak{C}_{j,k}\sum_{h=1}^{H}M_h,\\ \mathrm{RRD}_{j,\mathrm{cum}}(k) &\le HL_\varepsilon R_{\Theta,j}\mathfrak{C}_{j,k},\\ \mathrm{RRD}_{j,\mathrm{w,cum}}(k) &\le L_\varepsilon R_{\Theta,j}\mathfrak{C}_{j,k}\sum_{h=1}^{H}M_h.\end{aligned}$$

For the absolute criterion, the episodewise negative-part argument in the proof of Theorem E.7 applies on the same event and gives

$$\mathrm{Reg}_j^{\mathrm{abs}}(k) \le 2\mathfrak{C}_{j,k}\sum_{h=1}^{H}\beta_{j,h}^{\mathrm{U}} + 3R_{\Theta,j}(L_\eta+L_\varepsilon)\mathfrak{C}_{j,k}\sum_{h=1}^{H}M_h.$$

The growth condition (62) implies $R_{\Theta,j}=\tilde{\mathcal{O}}(\sqrt{d})$, while the phasewise backward entropy recursion gives

$$\beta_{j,h}^{\mathrm{U}} = \tilde{\mathcal{O}}((H-h)d + B_\vartheta + 1).$$

Since $k \le T_j$, the definition of $\mathfrak{C}_{j,k}$ gives $\mathfrak{C}_{j,k}=\tilde{\mathcal{O}}(\sqrt{T_j d})$. Consequently, for every local prefix,

$$\begin{aligned}\mathrm{Reg}_j(k), \mathrm{Reg}_j^{\mathrm{abs}}(k) &\le \tilde{\mathcal{O}}(H^2 d^{3/2}\sqrt{T_j}),\\ \mathrm{RRD}_{j,\mathrm{cum}}(k) \le \tilde{\mathcal{O}}(Hd\sqrt{T_j}), &\qquad \mathrm{RRD}_{j,\mathrm{w,cum}}(k) \le \tilde{\mathcal{O}}(H^2 d\sqrt{T_j}).\end{aligned}$$

Fix a completed-episode count $n \ge 1$, set the active phase index $J := \lfloor \log_2 n \rfloor$, and let $k := n - 2^J + 1 \in [T_J]$ be the number of episodes completed in the active phase. The preceding phases contain exactly $2^J - 1 = T_J - 1$ episodes. The four cumulative criteria add across phases, so we apply the preceding prefix bounds to every completed phase and to the active prefix. The logarithms in phase j contain

$$\log(1/\delta_j) = O(\log(1/\delta) + \log(j+1)).$$

Moreover,

$$\sum_{j=0}^{J}\sqrt{T_j} = \sum_{j=0}^{J}2^{j/2} \le \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{T_J} \le C\sqrt{n+1}$$

for a constant C. Summing the phasewise inequalities proves (63)–(65). Because $J \le \log_2(en)$, the phase-confidence term contributes at most $\log\log(en)$ inside $\tilde{\mathcal{O}}$. □

Proof of Corollary E.17. Let $\mathcal{E}_{\mathrm{any}}$ be the common event constructed in the proof of Corollary E.16. It has probability at least $1-\delta$ and, for every sample path $\omega \in \mathcal{E}_{\mathrm{any}}$, the three prefix bounds in (63)–(65) hold for every deterministic integer $n \ge 1$. Since $\mathbb{P}(\mathsf{N} < \infty) = 1$, the event $\mathcal{E}_{\mathrm{any}} \cap \{\mathsf{N} < \infty\}$ still has probability at least $1-\delta$. On this intersection, substitute $n = \mathsf{N}(\omega)$ in the pathwise inequalities. This proves (66)–(68) on $\mathcal{E}_{\mathrm{any}} \cap \{\mathsf{N} < \infty\}$. The proof uses only the simultaneous prefix event and therefore neither requires adaptedness of N nor invokes an optional-stopping identity. □

Proof of Corollary E.10. When $H = 1$, the deterministic ridge target is the observed value $u_1(s_1, a_1, b_1)$ because $\hat{W}_2^t \equiv 0$. Hence the future-value class is the singleton $\{0\}$, and the union over an adaptive value cover is unnecessary in the self-normalized regression argument. The backward definition sets $\mathfrak{L}_2 = \mathfrak{L}_{H+1} = 0$ and $\chi_{T,1} = 0$. Because the leader's reward is deterministic and there is no continuation value, (33) gives the sharper identity $\beta_1^{\mathrm{U}} = B_\vartheta$. Furthermore, $M_1 = 1$ and $\sum_{h=1}^{H} M_h = 1$. Substitution into (40)–(42) proves (53)–(54). Applying (51) with $M_1 = 1$ and $\beta_1^{\mathrm{U}} = B_\vartheta$ gives (55). Finally, (30) implies, up to logarithmic factors,

$$R_\Theta \le \tilde{\mathcal{O}}\left(B_\Theta + \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}}\sqrt{d}\right), \qquad \mathfrak{C}_{T,d,1}(\delta) = \tilde{\mathcal{O}}(\sqrt{Td}).$$

These relations give the dimension expansions following the corollary. When $H = 1$ and $\mathcal{S} = \{s_0\}$, the payoff-relevant recursion reduces to a single joint-action draw, while the terminal transition is

immaterial to the value recursion. The state-occupancy expectation is therefore trivial, the continuation value and its covering cost vanish, and $\mathrm{RRD} = \mathrm{RRD}_{\mathrm{w}}$ reduces to the single-state total-variation discrepancy. The fixed-point and global-deviation identities reduce statewise to their one-state forms, and the exact negative-log-likelihood minimization and exact-planning interfaces reduce to the corresponding one-step updates. These observations establish the claimed matrix-game specialization. For completeness, the stochastic martingale contribution to the Bellman-confidence radius in (33) is

$$\frac{H-h}{2}\sqrt{2\left[d\log(1+T/d)+\mathfrak{L}_{h+1}+\log(16H/\delta)\right]},$$

where $\mathfrak{L}_{h+1}$ is the deterministic upper bound on $\log N_\infty(\mathcal{W}_{h+1}, T^{-1})$ established in Proposition N.1. When $H=1$, necessarily $h=1$, and $\mathcal{W}_2=\{0\}$, $\mathfrak{L}_2=0$, and $(H-h)/2=0$. The remaining terms in (33) then reduce to $\beta_1^{\mathrm{U}} = B_\vartheta$. □

Q Absolute-Regret Bridge

For example, a sequence with constant suboptimality $-c<0$ has standard regret $-cT$ and absolute regret cT.

The preservation and leader guarantees are also connected beyond this algorithm. For every admissible policy-pair sequence, the follower-policy performance-difference identity gives

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2\,\mathrm{CPV}_{\mathrm{w}}(T) \le \mathrm{Reg}(T) + 2H\,\mathrm{CPV}(T).$$

The first inequality accounts for changes in future state occupancy through the remaining-horizon weights. Its proof is included with Theorem E.7 in Section Q. Thus cumulative preservation controls the extent to which execution can outperform the constrained benchmark by departing from the QR.

Lemma Q.1 (Performance difference between follower policies)**.** *Fix a leader policy π and follower policies ν, μ. Then*

$$J^{\mathrm{pol}}(\pi,\nu) - J^{\mathrm{pol}}(\pi,\mu) = \sum_{h=1}^{H} \mathbb{E}^{\pi,\nu}\left[\sum_{b\in\mathcal{B}} (\nu_h-\mu_h)(b\mid s_h) Q_h^{\pi,\mu}(s_h,b)\right], \tag{120}$$

where the leader's action-value function averaged over her own action is

$$Q_h^{\pi,\mu}(s,b) := \sum_{a\in\mathcal{A}} \pi_h(a\mid s)[u_h(s,a,b) + (P_h W_{h+1}^{\pi,\mu})(s,a,b)] \in [0, M_h].$$

Consequently,

$$J^{\mathrm{pol}}(\pi,\nu) - J^{\mathrm{pol}}(\pi,\mu) \le \sum_{h=1}^{H} M_h \mathbb{E}^{\pi,\nu}[\mathrm{TV}(\nu_h(\cdot\mid s_h), \mu_h(\cdot\mid s_h))]. \tag{121}$$

Proof of Lemma Q.1. Let $W_h^{\pi,\mu}$ be the leader's value under (π,μ), and let $W_{H+1}^{\pi,\mu} \equiv 0$. By the Bellman equation,

$$W_h^{\pi,\mu}(s) = \sum_{b\in\mathcal{B}} \mu_h(b\mid s) Q_h^{\pi,\mu}(s,b).$$

Under the trajectory generated by (π,ν), the tower property gives

$$\mathbb{E}^{\pi,\nu}[u_h(s_h,a_h,b_h) + W_{h+1}^{\pi,\mu}(s_{h+1}) - W_h^{\pi,\mu}(s_h)] = \mathbb{E}^{\pi,\nu}\left[\sum_{b\in\mathcal{B}} (\nu_h-\mu_h)(b\mid s_h) Q_h^{\pi,\mu}(s_h,b)\right].$$

Summing over h telescopes the value terms, and the left-hand side becomes

$$J^{\mathrm{pol}}(\pi,\nu) - \mathbb{E}_{\rho_0}[W_1^{\pi,\mu}(s_1)] = J^{\mathrm{pol}}(\pi,\nu) - J^{\mathrm{pol}}(\pi,\mu),$$

which proves (120). Moreover, $0 \le Q_h^{\pi,\mu}(s,b) \le M_h$. For any distributions p, q on a finite set X and any function $f: X \to [0,M]$, one has $\sum_{x\in X}[p(x)-q(x)]f(x) \le M\,\mathrm{TV}(p,q)$. Applying this inequality to every summand in (120) proves (121). □

Proof of Theorem E.7. For episode t, define the leader suboptimality, ER policy, and QR policy by

$$\Delta^t := J^* - J(\pi^t, y^t), \qquad \nu^t := \nu_{\theta^*}^{\pi^t, y^t}, \qquad \mu^t := \mu_{\theta^*}^{\pi^t}.$$

By Theorem 3.3, the pair (π^t, μ^t) satisfies the preservation constraint under the true response model, and hence it is feasible for the benchmark in (8). Therefore,

$$J^{\mathrm{pol}}(\pi^t, \mu^t) \le J^*.$$

Using $J(\pi^t, y^t) = J^{\mathrm{pol}}(\pi^t, \nu^t)$ and Lemma Q.1, we obtain

$$\begin{aligned}[-\Delta^t]_+ = \left[J(\pi^t, y^t) - J^*\right]_+ &\le \left[J^{\mathrm{pol}}(\pi^t, \nu^t) - J^{\mathrm{pol}}(\pi^t, \mu^t)\right]_+ \\ &\le \sum_{h=1}^{H} M_h \mathbb{E}^t\left[\mathrm{TV}\big(\nu_h^t(\cdot \mid s_h^t), \mu_h^t(\cdot \mid s_h^t)\big)\right].\end{aligned}$$

PROM (Algorithm 1) uses the estimated-preservation recommendation

$$y_h^t(\cdot \mid s) = \mu_{\hat{\theta}_h^t, h}^{\pi^t}(\cdot \mid s) = \nu_{\hat{\theta}_h^t, h}^{\pi^t, y^t}(\cdot \mid s).$$

On the response-confidence event in Proposition M.4, the direct preservation certificate gives

$$\mathrm{TV}\big(\nu_h^t(\cdot \mid s_h^t), \mu_h^t(\cdot \mid s_h^t)\big) \le b_h^{t,\mathrm{P}}(s_h^t, \pi_h^t).$$

Since $|x| = x + 2[-x]_+$, summing over episodes proves (50).

Finally, the certificate definition in (32) and the matrix-potential bound (116) imply

$$\sum_{t=1}^{T}\sum_{h=1}^{H} M_h \mathbb{E}^t\left[b_h^{t,\mathrm{P}}(s_h^t, \pi_h^t)\right] \le \kappa L_\eta R_\Theta \mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h.$$

Combining this inequality and (40) with (50) gives

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le 2\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} \beta_h^{\mathrm{U}} + R_\Theta[L_\varepsilon + (1+2\kappa)L_\eta]\, \mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h,$$

which is (51). □

Proof of Theorem E.9. Fix an episode and abbreviate the ER as $\nu := \nu_{\theta^*}^{\pi, y}$ and the QR as $\mu := \mu_{\theta^*}^{\pi}$. The pair that recommends μ satisfies the preservation constraint and, by Theorem 3.3, has ER μ. It is therefore feasible for the benchmark in (8), which implies $J^{\mathrm{pol}}(\pi, \mu) \le J^*$. Consequently,

$$[J(\pi, y) - J^*]_+ \le [J^{\mathrm{pol}}(\pi, \nu) - J^{\mathrm{pol}}(\pi, \mu)]_+.$$

Every recommendation satisfies $y_h(b \mid s) \ge \underline{y}$. By Lemma C.2, the executed response satisfies $\nu_h(b \mid s) \ge \underline{n} := |\mathcal{B}|^{-1}\, \underline{y}^\kappa e^{-\varepsilon}$. Thus Proposition C.3 applies pointwise and gives

$$\mathrm{TV}(\nu_h(\cdot \mid s), \mu_h(\cdot \mid s)) \le L_{\mathrm{dev}}\, \mathrm{TV}(y_h(\cdot \mid s), \nu_h(\cdot \mid s)).$$

Combining this inequality with (121) yields

$$[J(\pi, y) - J^*]_+ \le L_{\mathrm{dev}}\, \mathrm{RRD}_{\mathrm{w}}(\pi, y; \theta^*).$$

Write the leader suboptimality as $\Delta^t := J^* - J(\pi^t, y^t)$. Since $|x| = x + 2[-x]_+$,

$$\left|\Delta^t\right| \le \Delta^t + 2L_{\mathrm{dev}}\, \mathrm{RRD}_{\mathrm{w}}(\pi^t, y^t; \theta^*).$$

Summing over t proves $\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2L_{\mathrm{dev}}\, \mathrm{RRD}_{\mathrm{w,cum}}(T)$. Finally, $M_h \le H$ implies $\mathrm{RRD}_{\mathrm{w}} \le H\, \mathrm{RRD}$, which proves the second inequality in (52). □

Proof of Corollary E.6. We first simplify the deterministic quantities in the explicit bounds. By (35), $\mathfrak{C}_{T,d,H}(\delta) = \tilde{\mathcal{O}}(\sqrt{Td})$. The horizon sums satisfy

$$\sum_{h=1}^{H} M_h = \frac{H(H+1)}{2}, \qquad \sum_{h=1}^{H} (H-h) = \frac{H(H-1)}{2}. \tag{122}$$

We verify the entropy and confidence-radius orders with constants uniform in h and H. The backward definitions are finite and noncircular because β_h^{U} uses only the already constructed $\mathfrak{L}_{h+1}$. Define the logarithmic factor $\ell := \log(eTHd/\delta)$ and the maximum logarithmic covering bound $\mathfrak{L}_{\max} := \max_{1\le h\le H+1} \mathfrak{L}_h$. Under the fixed-parameter convention, $R_\Theta^2 = O(d\ell)$, and (33)–(34) give

$$K_h^{\mathrm{cov}} \le C_0 H\big(\sqrt{T} + \sqrt{d\ell} + \sqrt{\mathfrak{L}_{\max}}\big),$$
$$\mathfrak{L}_{\max} \le C_1 d^2\big[\ell + \log(1 + \mathfrak{L}_{\max})\big],$$

where the constants depend only on the fixed problem parameters. The factor d^2 accounts for the two symmetric-matrix nets, while the vector-net prefactors are of order d. For the normalized covering bound $x := \mathfrak{L}_{\max}/d^2$, the inequality $\log(1 + d^2x) \le 2\log d + \log(1 + x)$ yields $x \le C_2\big[\ell + \log(1 + x)\big]$. Using $\log(1 + x) \le x/(2C_2) + C_3$ gives $x = O(\ell)$. Consequently, uniformly in h and H,

$$\mathfrak{L}_h = O(d^2\ell), \qquad \beta_h^{\mathrm{U}} = O\big((H - h)d\sqrt{\ell} + B_\vartheta + 1\big).$$

At the terminal step, the exact identity $\beta_H^{\mathrm{U}} = B_\vartheta$ remains valid.

Substituting these estimates and (122) into (40)–(42) gives

$$\mathrm{Reg}(T) \le \tilde{\mathcal{O}}(H^2d^{3/2}\sqrt{T}), \qquad \mathrm{RRD}_{\mathrm{cum}}(T) = \tilde{\mathcal{O}}(Hd\sqrt{T}), \qquad \mathrm{RRD}_{\mathrm{w,cum}}(T) = \tilde{\mathcal{O}}(H^2d\sqrt{T}).$$

Under the same convention, $R_\Theta^2 = O(d\log(eTH/\delta))$, $\mathfrak{C}_{T,d,H}(\delta) = \tilde{\mathcal{O}}(\sqrt{Td})$, and $\mathfrak{S}_{T,d,H}(\delta) = O(d\log(eTH/\delta))$. Substitution into Corollaries E.3 and E.4 therefore gives

$$\mathrm{CPV}(T) = \tilde{\mathcal{O}}(Hd\sqrt{T}), \qquad \mathrm{CG}(T) = O\big(Hd^2\log^2(eTH/\delta)\big).$$

It remains to verify the displayed behavioral factors. Using $L_x = xe^x/2$ and $c_{\mathrm{H}}(x) = (1 - e^{-x})^2/8$, direct algebra gives

$$L_\varepsilon \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}} = \frac{\sqrt{2}\,\varepsilon e^{2\varepsilon}}{1 - e^{-\varepsilon}}, \qquad L_\eta \frac{e^\varepsilon}{\sqrt{c_{\mathrm{H}}(\varepsilon)}} = \frac{\sqrt{2}\,\eta e^{\eta+\varepsilon}}{1 - e^{-\varepsilon}}.$$

These identities yield (49) and the two response contributions stated in the corollary. When $\eta \le c_0\log(e\,|\mathcal{B}|)$ and $\tau \le \tau_{\max}$, every remaining nonlogarithmic behavioral quantity is polynomial in $|\mathcal{B}|$ with the stated fixed dependence, which proves the final claim. □

Proof of Corollary E.8. Apply (51). Since $0 \le \kappa < 1$,

$$L_\varepsilon + (1 + 2\kappa)L_\eta \le 3(L_\varepsilon + L_\eta).$$

The simplifications in the proof of Corollary E.6 give

$$2\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}\beta_h^{\mathrm{U}} = \tilde{\mathcal{O}}(H^2d^{3/2}\sqrt{T})$$

and

$$3R_\Theta(L_\eta + L_\varepsilon)\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}M_h = \tilde{\mathcal{O}}(H^2d\sqrt{T})$$

under the fixed-parameter convention. Since $d \le d^{3/2}$ for $d \ge 1$, the second term is absorbed by the first, which proves the claim. □

Proof of Corollary E.13. Because I is independent of $\mathcal{F}_0^{T+1}$ and uniform on $[T]$, one has

$$\mathbb{P}(I = t \mid \mathcal{F}_0^{T+1}) = \frac{1}{T} \quad \text{almost surely for every } t \in [T].$$

Let $Z_1, \dots, Z_T$ be any integrable $\mathcal{F}_0^{T+1}$-measurable random variables. Since $Z_I = \sum_{t=1}^{T}\mathbf{1}\{I = t\}Z_t$, conditional linearity and the preceding identity give

$$\mathbb{E}[Z_I \mid \mathcal{F}_0^{T+1}] = \sum_{t=1}^{T}\mathbb{E}[\mathbf{1}\{I = t\}Z_t \mid \mathcal{F}_0^{T+1}] = \sum_{t=1}^{T}Z_t\mathbb{P}(I = t \mid \mathcal{F}_0^{T+1}) = \frac{1}{T}\sum_{t=1}^{T}Z_t$$

almost surely. Apply this identity first with $Z_t = |J^* - J(\pi^t, y^t)|$ to obtain

$$\mathbb{E}_I \left|J^* - J(\pi^I, y^I)\right| = \frac{1}{T}\sum_{t=1}^{T} \left|J^* - J(\pi^t, y^t)\right| = \frac{\mathrm{Reg}^{\mathrm{abs}}(T)}{T}.$$

Applying the same identity with $Z_t = \mathrm{RRD}(\pi^t, y^t; \theta^*)$ and with $Z_t = \mathrm{RRD}_{\mathrm{w}}(\pi^t, y^t; \theta^*)$ proves the other two equalities. The high-probability bounds follow by evaluating these almost-sure conditional identities on the $\mathcal{F}_0^{T+1}$-measurable training event. □

Proof of Corollary E.14. On the training event, Corollary E.13 gives

$$\mathbb{E}_I[|J^* - J(\hat{\pi}, \hat{y})| + \mathrm{RRD}(\hat{\pi}, \hat{y}; \theta^*)] = \frac{\mathrm{Reg}^{\mathrm{abs}}(T) + \mathrm{RRD}_{\mathrm{cum}}(T)}{T}.$$

Substituting (51) and (41) into the right-hand side gives

$$\begin{aligned}\mathrm{Reg}^{\mathrm{abs}}(T) + \mathrm{RRD}_{\mathrm{cum}}(T) \le 2\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} \beta_h^{\mathrm{U}} \\ + 3R_\Theta(L_\eta + L_\varepsilon)\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h + HL_\varepsilon R_\Theta \mathfrak{C}_{T,d,H}(\delta).\end{aligned}$$

The last display is precisely the quantity $\mathfrak{R}_T^{\mathrm{out}}(\delta)$ in the corollary, which proves (59) and (60).

Under the fixed-parameter convention, Corollaries E.6 and E.8 give

$$\frac{\mathfrak{R}_T^{\mathrm{out}}(\delta)}{T} = \tilde{\mathcal{O}}\left(\frac{H^2 d^{3/2}}{\sqrt{T}}\right).$$

Solving this inequality for an expected error level ϵ yields $T = \tilde{\mathcal{O}}(H^4 d^3/\epsilon^2)$.

For the tail statement, define the combined output error by

$$X := |J^* - J(\hat{\pi}, \hat{y})| + \mathrm{RRD}(\hat{\pi}, \hat{y}; \theta^*) \ge 0.$$

Let $\mathcal{E}_T$ be the joint training event used in the proof of Theorem E.2, so $\mathcal{E}_T \in \mathcal{F}_0^{T+1}$ and $\mathbb{P}(\mathcal{E}_T^c) \le \delta$. On $\mathcal{E}_T$, the preceding conditional-expectation bound and Markov's inequality give, almost surely,

$$\mathbb{P}\left(X > \frac{\mathfrak{R}_T^{\mathrm{out}}(\delta)}{T\delta_{\mathrm{out}}} \,\middle|\, \mathcal{F}_0^{T+1}\right) \le \frac{T\mathbb{E}[X \mid \mathcal{F}_0^{T+1}]}{\mathfrak{R}_T^{\mathrm{out}}(\delta)}\delta_{\mathrm{out}} \le \delta_{\mathrm{out}}.$$

Define the output-error tail event by

$$\mathcal{A}_T := \left\{X > \frac{\mathfrak{R}_T^{\mathrm{out}}(\delta)}{T\delta_{\mathrm{out}}}\right\}.$$

Because $\mathcal{E}_T$ is $\mathcal{F}_0^{T+1}$-measurable, the law of total probability yields

$$\mathbb{P}(\mathcal{A}_T) \le \mathbb{P}(\mathcal{E}_T^c) + \mathbb{P}(\mathcal{A}_T \cap \mathcal{E}_T) = \mathbb{P}(\mathcal{E}_T^c) + \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_T}\mathbb{P}(\mathcal{A}_T \mid \mathcal{F}_0^{T+1})\right] \le \delta + \delta_{\mathrm{out}}.$$

This proves (61). Solving the threshold in (61) for T gives the stated $\delta_{\mathrm{out}}^{-2}$ dependence. □

Proof of Corollary E.15. Condition on the complete interaction data. The uniform index I is then the only remaining random variable, so Corollary E.13 gives

$$\mathbb{E}_I[|J^* - J(\hat{\pi}, \hat{y})| + \mathrm{RRD}_{\mathrm{w}}(\hat{\pi}, \hat{y}; \theta^*)] = \frac{\mathrm{Reg}^{\mathrm{abs}}(T) + \mathrm{RRD}_{\mathrm{w,cum}}(T)}{T}.$$

On the training event, (51) and (42) imply

$$\begin{aligned}\mathrm{Reg}^{\mathrm{abs}}(T) + \mathrm{RRD}_{\mathrm{w,cum}}(T) \le 2\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} \beta_h^{\mathrm{U}} + 3R_\Theta(L_\eta + L_\varepsilon)\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h \\ + L_\varepsilon R_\Theta \mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} M_h.\end{aligned}$$

The right-hand side is the value-sensitive output bound defined in the corollary. This proves the deterministic and conditional-expectation conclusions. For the tail statement, define the value-sensitive output error by

$$X_{\mathrm{w}} := |J^* - J(\hat{\pi}, \hat{y})| + \mathrm{RRD}_{\mathrm{w}}(\hat{\pi}, \hat{y}; \theta^*) \geq 0.$$

Let $\mathcal{E}_T$ be the joint training event used in the proof of Theorem E.2. On $\mathcal{E}_T$,

$$\mathbb{E}[X_{\mathrm{w}} \mid \mathcal{F}_0^{T+1}] \leq \frac{\mathfrak{R}^{\mathrm{out}}_{T,\mathrm{w}}(\delta)}{T}.$$

Markov's inequality therefore gives, almost surely on $\mathcal{E}_T$,

$$\mathbb{P}\left(X_{\mathrm{w}} > \frac{\mathfrak{R}^{\mathrm{out}}_{T,\mathrm{w}}(\delta)}{T\delta_{\mathrm{out}}} \,\middle|\, \mathcal{F}_0^{T+1} \right) \leq \delta_{\mathrm{out}}.$$

Define the value-sensitive output-error tail event by

$$\mathcal{A}_{T,\mathrm{w}} := \left\{ X_{\mathrm{w}} > \frac{\mathfrak{R}^{\mathrm{out}}_{T,\mathrm{w}}(\delta)}{T\delta_{\mathrm{out}}} \right\}.$$

Since $\mathcal{E}_T \in \mathcal{F}_0^{T+1}$ and $\mathbb{P}(\mathcal{E}_T^c) \leq \delta$,

$$\mathbb{P}(\mathcal{A}_{T,\mathrm{w}}) \leq \mathbb{P}(\mathcal{E}_T^c) + \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_T} \mathbb{P}(\mathcal{A}_{T,\mathrm{w}} \mid \mathcal{F}_0^{T+1})\big] \leq \delta + \delta_{\mathrm{out}}.$$

This proves the stated joint tail bound. □

Proof of Corollary G.11. When $\tau = 0$, (4) reduces to (2) for every admissible full-support recommendation. Therefore $J(\pi, y) = J^{\mathrm{pol}}(\pi, \mu^{\pi}_{\theta^*})$, which is at most the maximum in (8). This proves nonnegativity of every episodic suboptimality and hence (81). Finally, $\varepsilon = \eta$, while the current covariance and radius are shared by (31); the two bonuses therefore coincide. □

Proof of Corollary E.11. Substitute $T = N/H$ into Corollaries E.6 and E.8. In particular, $H^2\sqrt{N/H} = H^{3/2}\sqrt{N}$ and $H\sqrt{N/H} = H^{1/2}\sqrt{N}$. The logarithmic factors are absorbed by $\tilde{\mathcal{O}}$. □

R Behavioral-Parameter Misspecification

Suppose the true follower uses (η, τ), while the learner uses fixed fitted parameters $(\bar{\eta}, \bar{\tau})$. Define the fitted effective inverse temperature and recommendation exponent by

$$\bar{\varepsilon} := \frac{\bar{\eta}}{1 + \bar{\eta}\bar{\tau}}, \qquad \bar{\kappa} := \frac{\bar{\eta}\bar{\tau}}{1 + \bar{\eta}\bar{\tau}}.$$

Assumption R.1 (Compact behavioral domain)**.** *The true and fitted parameters satisfy*

$$\eta, \bar{\eta} \in [\eta_{\min}, \eta_{\max}], \qquad \tau, \bar{\tau} \in [0, \tau_{\max}], \qquad \eta_{\min} > 0.$$

The learner is given a valid calibration bound $\zeta_0 \in [0, 1]$ for the behavioral-parameter error ζ, where

$$\zeta := \max\{|\eta - \bar{\eta}|, |\tau - \bar{\tau}|\} \leq \zeta_0.$$

The robust algorithm uses ζ_0, not the unknown actual error ζ.

Every fitted QR recommendation satisfies the uniform probability lower bound

$$y_h(b \mid s) \geq \underline{y}_{\mathrm{mis}} := B^{-1} e^{-\eta_{\max}}.$$

For a reward vector $q \in [0, 1]^B$, define the response score

$$z_b(\eta, \tau) := \frac{\eta\tau}{1 + \eta\tau} \log y_b + \frac{\eta}{1 + \eta\tau} q_b.$$

Define the response-score sensitivity constant and the response-score error bound by

$$C_{\mathrm{beh}} := 1 + \eta_{\max}^2 + (\tau_{\max} + \eta_{\max}) \log(1/\underline{y}_{\mathrm{mis}}), \quad D_{\mathrm{beh}} := C_{\mathrm{beh}} \zeta_0.$$

We analyze the local robustness regime $D_{\mathrm{beh}} \le 1$.

The misspecified algorithm takes as inputs $(\bar{\eta}, \bar{\tau})$, the calibration certificate ζ_0, the bounds in Definition 2.1, and a fixed measurable exact or certified approximate likelihood solver. It replaces $(\eta, \tau, \varepsilon)$ by $(\bar{\eta}, \bar{\tau}, \bar{\varepsilon})$ in the likelihood, recommendation, and planning rule. Write $\bar{\nu}$ and $\bar{\mu}$ for fitted response maps. Define the fitted negative log-likelihood function, its MLE, the historical covariance matrix, and the feature covariance matrices by

$$\bar{\ell}_h^t(\theta) := -\sum_{i=1}^{t-1} \log \bar{\nu}_{\theta,h}^{\pi^i, y^i}(b_h^i \mid s_h^i), \qquad \hat{\theta}_h^t \in \operatorname*{argmin}_{\theta \in \Theta_h} \bar{\ell}_h^t(\theta), \tag{123}$$

$$\bar{V}_h^t := I_d + \sum_{i=1}^{t-1} \bar{\Sigma}_h^{i, \hat{\theta}_h^i}, \qquad \bar{\Sigma}_h^{i,\theta} := \mathrm{Cov}_{b \sim \bar{\nu}_{\theta,h}^{\pi^i, y^i}(\cdot \mid s_h^i)}[\phi_h^{\pi^i}(s_h^i, b)].$$

For the exact variant, a fixed Borel measurable minimizer of the fitted loss is obtained by the same measurable-selection argument used after (15). Because $(\bar{\eta}, \bar{\tau})$ are fixed, the fitted loss has the same pre-episode history measurability and parameter continuity as the well-specified loss in (15); only the numerical response map changes. The certified approximate variant uses a Borel measurable solver satisfying the stated negative-log-likelihood-gap certificate, and all of its required radius substitutions are specified in Corollary R.5. For a candidate p, define the estimated QR recommendation, its feature covariance matrix, and the response width by

$$\begin{aligned} \bar{y}_h^{t,p}(\cdot \mid s) &:= \bar{\mu}_{\hat{\theta}_h^t, h}^p(\cdot \mid s), \\ \bar{\Gamma}_h^t(s,p) &:= \mathrm{Cov}_{b \sim \bar{y}_h^{t,p}(\cdot \mid s)}[\phi_h^p(s,b)], \\ \bar{\rho}_h^t(s,p) &:= \sqrt{\mathrm{Tr}((\bar{V}_h^t)^{-1} \bar{\Gamma}_h^t(s,p))}. \end{aligned}$$

Define the fitted likelihood confidence threshold, the constant in the covariance error bound, and the response confidence radius by

$$\begin{aligned} \bar{\beta}_\Theta(\delta) &:= 4 \left[d \log(1 + 16 B_\Theta \bar{\varepsilon} T) + \log \frac{32HT}{\delta} + 4 \right], \\ C_{\mathrm{mis}} &:= \frac{8}{c_{\mathrm{H}}(\bar{\varepsilon})}, \\ R_{\mathrm{mis}}^t &:= \sqrt{B_\Theta^2 + e^{2\bar{\varepsilon}} C_{\mathrm{mis}} [\bar{\beta}_\Theta(\delta) + (t-1) D_{\mathrm{beh}}^2]}. \end{aligned}$$

The robust follower-response bonus and QR bonus are

$$\begin{aligned} b_h^{t,\mathrm{F},mis}(s,p) &:= \min \left\{ 1, L_{\bar{\varepsilon}} R_{\mathrm{mis}}^t \bar{\rho}_h^t(s,p) + \frac{1}{2} D_{\mathrm{beh}} \right\}, \\ b_h^{t,\mathrm{Q},mis}(s,p) &:= \min \left\{ 1, L_{\bar{\eta}} R_{\mathrm{mis}}^t \bar{\rho}_h^t(s,p) + \frac{1}{2} \zeta_0 \right\}. \end{aligned}$$

Define the robust envelope backward by setting $\bar{\mathfrak{L}}_{H+1} := 0$. For $h = H, H-1, \ldots, 1$, first define the Bellman confidence radius by

$$\bar{\beta}_h^{\mathrm{U}} := \frac{H-h}{2} \sqrt{2 \left[d \log(1 + T/d) + \bar{\mathfrak{L}}_{h+1} + \log(32H/\delta) \right]} + B_\vartheta + (H-h) B_P + \chi_{T,h}. \tag{124}$$

Then define the covering constant, net scale, and logarithmic covering bound by

$$\begin{aligned} \bar{K}_h^{\mathrm{cov}} &:= 1 + \sqrt{T} M_h + B_\Theta + 2M_h + \bar{\beta}_h^{\mathrm{U}} + M_h \bar{\eta} + M_h L_{\bar{\eta}} R_{\mathrm{mis}}^T \sqrt{1 + 3\bar{\eta}} + M_h L_{\bar{\eta}} + M_h, \\ \bar{q}_h &:= (8T \bar{K}_h^{\mathrm{cov}})^{-1}, \\ \bar{\mathfrak{L}}_h &:= d \log \left(1 + \frac{2\sqrt{T} M_h}{\bar{q}_h} \right) + d \log \left(1 + \frac{2B_\Theta}{\bar{q}_h^2} \right) + 2m_d \log \left(1 + \frac{2\sqrt{d}}{\bar{q}_h^2} \right) \\ &\quad + \log \left(1 + \frac{2R_{\mathrm{mis}}^T}{\bar{q}_h} \right) + \log \left(1 + \frac{\zeta_0}{\bar{q}_h} \right). \end{aligned} \tag{125}$$

After these constants have been fixed, set the terminal value class $\bar{\mathcal{W}}_{H+1} := \{0\}$. For each fixed $h \in [H]$, consider parameters satisfying

$$\omega \in \mathbb{R}^d, \qquad A, C \in \mathbb{S}^d_+, \qquad \theta \in \Theta_h, \qquad \|\omega\|_2 \le \sqrt{T} M_h,$$
$$0 \preceq A, C \preceq I_d, \qquad r \in [0, R^T_{\mathrm{mis}}], \qquad q_0 \in [0, \zeta_0/2].$$

For these parameters, define the robust optimistic action-value function, fitted QR policy, and its feature covariance matrix by

$$\bar{U}_{\omega,A}(s,a,b) := \mathrm{Trun}_{[0,M_h]}\left(\phi_h(s,a,b)^\top \omega + \bar{\beta}^{\mathrm{U}}_h \sqrt{\phi_h(s,a,b)^\top A \phi_h(s,a,b)}\right),$$
$$\bar{\mu}^p_{\theta,h}(b \mid s) := \frac{\exp\{\bar{\eta}\langle \phi^p_h(s,b), \theta\rangle\}}{\sum_{c\in\mathcal{B}} \exp\{\bar{\eta}\langle \phi^p_h(s,c), \theta\rangle\}},$$
$$\bar{\Gamma}^p_{\theta,h}(s) := \mathrm{Cov}_{b\sim\bar{\mu}^p_{\theta,h}(\cdot|s)}[\phi^p_h(s,b)].$$

For $p \in \Delta(\mathcal{A})$, define the robust planning objective and the corresponding value function by

$$\bar{G}_{\omega,A,\theta,C,r,q_0}(s,p) := \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a)\bar{\mu}^p_{\theta,h}(b \mid s)\bar{U}_{\omega,A}(s,a,b) + M_h \min\left\{1, L_{\bar{\eta}} r \sqrt{\mathrm{Tr}(C\bar{\Gamma}^p_{\theta,h}(s))} + q_0\right\},$$
$$\bar{F}_{\omega,A,\theta,C,r,q_0}(s) := \mathrm{Trun}_{[0,M_h]}\left(\sup_{p\in\Delta(\mathcal{A})} \bar{G}_{\omega,A,\theta,C,r,q_0}(s,p)\right).$$

Let the robust generated value class $\bar{\mathcal{W}}_h$ be the collection of all functions $\bar{F}_{\omega,A,\theta,C,r,q_0}$ over the displayed parameter ranges. The scalar parameters r and q_0 each require a one-dimensional cover.

For completeness, define the fitted robust recursion only after its noncircular confidence radii have been fixed. In this paragraph all hatted values, policies, and recommendations refer to the fitted recursion. Initialize the terminal value function $\hat{W}^{t,\mathrm{mis}}_{H+1} \equiv 0$, and, backward in h, define the ridge-regression estimator, optimistic action-value function, and planning objective by

$$\hat{\omega}^{t,\mathrm{mis}}_h := (\Lambda^t_h)^{-1} \sum_{i=1}^{t-1} \phi^i_h \left(u_h(s^i_h, a^i_h, b^i_h) + \hat{W}^{t,\mathrm{mis}}_{h+1}(s^i_{h+1})\right), \tag{126}$$

$$\hat{U}^{t,\mathrm{mis}}_h(s,a,b) := \mathrm{Trun}_{[0,M_h]}\left(\phi_h(s,a,b)^\top \hat{\omega}^{t,\mathrm{mis}}_h + \bar{\beta}^{\mathrm{U}}_h \sqrt{\phi_h(s,a,b)^\top (\Lambda^t_h)^{-1} \phi_h(s,a,b)}\right), \tag{127}$$

$$\bar{\mathcal{Q}}^t_h(s,p) := \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a)\bar{y}^{t,p}_h(b \mid s)\hat{U}^{t,\mathrm{mis}}_h(s,a,b) + M_h b^{t,\mathrm{Q},mis}_h(s,p). \tag{128}$$

Specify the leader policy, recommendation policy, and optimistic value function by

$$\pi^{t,\mathrm{mis}}_h(\cdot \mid s) \in \operatorname*{argmax}_{p\in\Delta(\mathcal{A})} \bar{\mathcal{Q}}^t_h(s,p),$$
$$y^{t,\mathrm{mis}}_h(\cdot \mid s) := \bar{y}^{t,\pi^{t,\mathrm{mis}}_h}_h(\cdot \mid s), \tag{129}$$
$$\hat{W}^{t,\mathrm{mis}}_h(s) := \mathrm{Trun}_{[0,M_h]}\left(\bar{\mathcal{Q}}^t_h(s, \pi^{t,\mathrm{mis}}_h(\cdot \mid s))\right). \tag{130}$$

The fitted objective $\bar{\mathcal{Q}}^t_h(s,p)$ has the same history measurability and continuity properties as $\mathcal{Q}^t_h(s,p)$, and the robust recursion fixes a Borel measurable maximum selector by the same construction used in the well-specified backward recursion. To reduce notation in the robust theorem and its proof, we suppress the superscript "mis" after this display.

The next two results are the fitted, misspecification-robust counterparts of Proposition N.1 and lemmas O.1 and P.4. They support Theorem R.4 and its corollaries.

Proposition R.2 (Robust value-class cover and Bellman confidence)**.** *For every $h \in [H]$, the deterministic class $\bar{\mathcal{W}}_h$ satisfies*

$$\log N_\infty(\bar{\mathcal{W}}_h, T^{-1}) \le \bar{\mathfrak{L}}_h.$$

For every $t \in [T]$ and $h \in [H]$, the value function $\hat{W}_h^t$ generated by the fitted robust recursion belongs to $\bar{\mathcal{W}}_h$. Moreover, with probability at least $1-\delta/4$, simultaneously for every $t \in [T]$, $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $b \in \mathcal{B}$,

$$\left|\phi_h(s,a,b)^\top \hat{\omega}_h^t - [u_h(s,a,b) + (P_h \hat{W}_{h+1}^t)(s,a,b)]\right| \le \bar{\beta}_h^{\mathrm{U}} \sqrt{\phi_h(s,a,b)^\top (\Lambda_h^t)^{-1} \phi_h(s,a,b)}. \tag{131}$$

Corollary R.3 (Expected potentials for the fitted recursion)**.** *With probability at least $1-\delta/4$, simultaneously for every $h \in [H]$,*

$$\sum_{t=1}^{T} \mathbb{E}^t \sqrt{\phi_h(s_h^t, a_h^t, b_h^t)^\top (\Lambda_h^t)^{-1} \phi_h(s_h^t, a_h^t, b_h^t)} \le \mathfrak{C}_{T,d,H}(\delta),$$

$$\sum_{t=1}^{T} \mathbb{E}^t \bar{\rho}_h^t(s_h^t, \pi_h^t) \le \mathfrak{C}_{T,d,H}(\delta). \tag{132}$$

Also define the baseline response confidence radius R_{mis}^0, its growth coefficient C_R, and the potential bounds $\mathfrak{D}_{T,d,H}(\delta)$ and $\mathfrak{E}_{T,d,H}^{\mathrm{mis}}(\delta)$ by

$$R_{\mathrm{mis}}^0 := \sqrt{B_\Theta^2 + e^{2\bar{\varepsilon}} C_{\mathrm{mis}} \bar{\beta}_\Theta(\delta)}, \qquad C_R := e^{\bar{\varepsilon}} \sqrt{C_{\mathrm{mis}}},$$
$$\mathfrak{D}_{T,d,H}(\delta) := \sqrt{T}\, \mathfrak{C}_{T,d,H}(\delta), \qquad \mathfrak{E}_{T,d,H}^{\mathrm{mis}}(\delta) := R_{\mathrm{mis}}^0 \mathfrak{C}_{T,d,H}(\delta) + C_R D_{\mathrm{beh}} \mathfrak{D}_{T,d,H}(\delta).$$

Then $R_{\mathrm{mis}}^t \le R_{\mathrm{mis}}^0 + C_R D_{\mathrm{beh}} \sqrt{t-1}$. For transparent behavioral-constant bookkeeping, also define the behavioral factors for the regret and RRD bounds by

$$\mathfrak{B}_{\mathrm{mis}}^{\mathrm{R}} := C_{\mathrm{beh}}(L_{\bar{\eta}} + L_{\bar{\varepsilon}}) e^{\bar{\varepsilon}} \sqrt{\frac{8}{c_{\mathrm{H}}(\bar{\varepsilon})}} + \frac{C_{\mathrm{beh}}+1}{2}, \quad \mathfrak{B}_{\mathrm{mis}}^{\mathrm{F}} := C_{\mathrm{beh}} L_{\bar{\varepsilon}} e^{\bar{\varepsilon}} \sqrt{\frac{8}{c_{\mathrm{H}}(\bar{\varepsilon})}} + \frac{C_{\mathrm{beh}}}{2}. \tag{133}$$

Theorem R.4 (Regret under behavioral misspecification)**.** *Let $T \ge 1$ and $\delta \in (0,1)$. Suppose the environment is the linear LFG in Definition 2.1 and the supplied bounds $B_\Theta, B_\vartheta, B_P$ are valid. Assume that Assumption R.1 holds, $D_{\mathrm{beh}} \le 1$, the fitted negative-log-likelihood minimizers in (123) are exact and selected by the fixed measurable fitted-loss convention, and the learner uses the fitted recursion (126)–(130) with the fixed measurable exact full-simplex planning convention. With probability at least $1-\delta$,*

$$\mathrm{Reg}(T) \le 2\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^{H} \bar{\beta}_h^{\mathrm{U}} + \sum_{h=1}^{H} M_h \left\{ (L_{\bar{\eta}} + L_{\bar{\varepsilon}}) \mathfrak{E}_{T,d,H}^{\mathrm{mis}}(\delta) + \frac{T}{2}(D_{\mathrm{beh}} + \zeta_0) \right\}, \tag{134}$$

$$\mathrm{RRD}_{\mathrm{cum}}(T) \le H \left\{ L_{\bar{\varepsilon}} \mathfrak{E}_{T,d,H}^{\mathrm{mis}}(\delta) + \frac{T}{2} D_{\mathrm{beh}} \right\}, \tag{135}$$

$$\mathrm{RRD}_{\mathrm{w,cum}}(T) \le S_H \times \left\{ L_{\bar{\varepsilon}} \mathfrak{E}_{T,d,H}^{\mathrm{mis}}(\delta) + \frac{T}{2} D_{\mathrm{beh}} \right\}. \tag{136}$$

Under the fixed-parameter convention used in Corollary E.6, the preceding explicit bounds imply

$$\mathrm{Reg}(T) \le \tilde{\mathcal{O}}\left(H^2 d^{3/2} \sqrt{T} + H^2 \mathfrak{B}_{\mathrm{mis}}^{\mathrm{R}} \zeta_0 T \sqrt{d} \right),$$
$$\mathrm{RRD}_{\mathrm{cum}}(T) \le \tilde{\mathcal{O}}\left(H d \sqrt{T} + H \mathfrak{B}_{\mathrm{mis}}^{\mathrm{F}} \zeta_0 T \sqrt{d} \right).$$

The result is a graceful-degradation guarantee. More precisely, consider a family of T-episode runs in which the fitted behavioral coefficients and a valid calibration certificate $\zeta_0(T)$ are fixed before each run and may depend on T. Keep the remaining problem parameters, δ, and the compact behavioral domain fixed. The displayed bounds are sublinear whenever

$$\max\{\mathfrak{B}_{\mathrm{mis}}^{\mathrm{R}}, \mathfrak{B}_{\mathrm{mis}}^{\mathrm{F}}\}\, \zeta_0(T) \frac{\mathfrak{C}_{T,d,H}(\delta)}{\sqrt{T}} = o(1).$$

Since the behavioral factors are uniformly bounded on the fixed compact domain, it suffices that $\zeta_0(T)\sqrt{\log(eT)} \to 0$. A fixed nonzero calibration certificate does not yield a sublinear guarantee from these bounds. The same conclusion applies to the absolute regret of the robust variant of PROM (Algorithm 1) through Corollary R.6; the direct bridge removes the inverse-map factor but does not remove the calibration-bias term.

For the optimization-tolerant robust variant, suppose a fitted solver is fixed before interaction and returns an $\mathcal{F}_0^t$-measurable point $\tilde{\theta}_h^t \in \Theta_h$. For a known tolerance $\bar{\gamma}_{\mathrm{MLE}} \geq 0$, assume that

$$\bar{\ell}_h^t(\tilde{\theta}_h^t) \leq \inf_{\theta \in \Theta_h} \bar{\ell}_h^t(\theta) + \bar{\gamma}_{\mathrm{MLE}} \qquad \text{for every } t \in [T],\ h \in [H]. \tag{137}$$

We use a superscript γ to denote quantities generated by the approximate fitted solver with tolerance $\bar{\gamma}_{\mathrm{MLE}}$. Form the fitted historical covariance matrix

$$\bar{V}_h^{t,\gamma} := I_d + \sum_{i=1}^{t-1} \bar{\Sigma}_h^{i,\tilde{\theta}_h^i}.$$

For a candidate p, also define the estimated QR recommendation, its feature covariance matrix, and the response width by

$$\begin{aligned}
\bar{y}_h^{t,p,\gamma}(\cdot \mid s) &:= \bar{\mu}^p_{\tilde{\theta}_h^t,h}(\cdot \mid s), \\
\bar{\Gamma}_h^{t,\gamma}(s,p) &:= \mathrm{Cov}_{b \sim \bar{y}_h^{t,p,\gamma}(\cdot|s)}[\phi_h^p(s,b)], \\
\bar{\rho}_h^{t,\gamma}(s,p) &:= \sqrt{\mathrm{Tr}((\bar{V}_h^{t,\gamma})^{-1}\bar{\Gamma}_h^{t,\gamma}(s,p))}.
\end{aligned}$$

Define the response confidence radius for the approximate fitted MLE by

$$R_{\mathrm{mis}}^{t,\gamma} := \sqrt{B_\Theta^2 + e^{2\bar{\varepsilon}} C_{\mathrm{mis}} \left[\bar{\beta}_\Theta(\delta) + (t-1)D_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}}\right]}.$$

Also define the baseline and terminal response confidence radii by

$$\begin{aligned}
R_{\mathrm{mis}}^{0,\gamma} &:= \sqrt{B_\Theta^2 + e^{2\bar{\varepsilon}} C_{\mathrm{mis}} [\bar{\beta}_\Theta(\delta) + \bar{\gamma}_{\mathrm{MLE}}]}, \\
R_{\mathrm{mis}}^{T,\gamma} &:= \sqrt{B_\Theta^2 + e^{2\bar{\varepsilon}} C_{\mathrm{mis}} [\bar{\beta}_\Theta(\delta) + (T-1)D_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}}]}.
\end{aligned}$$

Recompute the complete robust envelope, rather than only the online bonus: start from $\bar{\mathfrak{L}}_{H+1,\gamma} = 0$, replace R_{mis}^T by $R_{\mathrm{mis}}^{T,\gamma}$ in (124)–(125), and denote the recursively obtained quantities by $\bar{\beta}_{h,\gamma}^{\mathrm{U}}, \bar{K}_{h,\gamma}^{\mathrm{cov}}, \bar{q}_{h,\gamma}, \bar{\mathfrak{L}}_{h,\gamma}$. Make the same endpoint-radius substitution in the range of the scalar parameter r. Construct the fitted historical covariance matrices, recommendations, response bonuses, and Bellman recursion from $\tilde{\theta}_h^t$, $\bar{V}_h^{t,\gamma}$, $R_{\mathrm{mis}}^{t,\gamma}$, and this recomputed envelope, using the fixed Borel measurable maximum-selection convention for the resulting fitted planning objective. Within this variant and its analysis, unsuperscripted policy variables denote the outputs of this approximate robust recursion.

Let $\bar{\mathcal{W}}_{h,\gamma}$ be the robust generated-value class obtained from $\bar{\mathcal{W}}_h$ in Proposition R.2 by replacing R_{mis}^T and $\bar{\beta}_h^{\mathrm{U}}$ with $R_{\mathrm{mis}}^{T,\gamma}$ and $\bar{\beta}_{h,\gamma}^{\mathrm{U}}$, respectively.

Corollary R.5 (Optimization-tolerant fitted MLE)**.** *For the optimization-tolerant robust variant just defined, with probability at least $1 - \delta/4$, simultaneously for all $t \in [T]$ and $h \in [H]$,*

$$\left\|\tilde{\theta}_h^t - \theta_h^*\right\|_{\bar{V}_h^{t,\gamma}} \leq R_{\mathrm{mis}}^{t,\gamma}. \tag{138}$$

On the same MLE event, the two robust response bounds hold for the approximate fitted objects. The estimated-preservation identity implies the determinant-potential identity deterministically for every realized covariance sequence. After intersecting the MLE event with the robust Bellman-confidence and potential-martingale events, all conclusions of Theorem R.4 remain valid with probability at least $1 - \delta$ after replacing R_{mis}^t, R_{mis}^0, and $\bar{\beta}_h^{\mathrm{U}}$ by $R_{\mathrm{mis}}^{t,\gamma}$, $R_{\mathrm{mis}}^{0,\gamma}$, and $\bar{\beta}_{h,\gamma}^{\mathrm{U}}$, respectively. With the same substitutions, the direct robust absolute-regret conclusion in Corollary R.6 also remains valid on this joint event.

The term $C_R D_{\mathrm{beh}} \mathfrak{D}_{T,d,H}(\delta)$ is unchanged because $R^{t,\gamma}_{\mathrm{mis}} \le R^{0,\gamma}_{\mathrm{mis}} + C_R D_{\mathrm{beh}} \sqrt{t-1}$. In particular, a tolerance $\bar{\gamma}_{\mathrm{MLE}} = \tilde{\mathcal{O}}(d)$ preserves the displayed rates up to logarithmic factors and fixed behavioral constants.

Corollary R.6 (Direct absolute-regret control for robust PROM)**.** *Under the conditions of Theorem R.4, on the same joint event of probability at least* $1-\delta$*,*

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2\sum_{t=1}^{T}\sum_{h=1}^{H} M_h \mathbb{E}^t\left[b_h^{t,\mathrm{F},mis}(s_h^t,\pi_h^t) + b_h^{t,\mathrm{Q},mis}(s_h^t,\pi_h^t)\right] \tag{139}$$

$$\le 2\mathfrak{C}_{T,d,H}(\delta)\sum_{h=1}^{H}\bar{\beta}_h^{\mathrm{U}} + 3S_H\left\{(L_{\bar{\eta}} + L_{\bar{\varepsilon}})\mathfrak{E}^{\mathrm{mis}}_{T,d,H}(\delta) + \frac{T}{2}(D_{\mathrm{beh}} + \zeta_0)\right\}. \tag{140}$$

Under the fixed-parameter convention and exact planning, this implies, with probability at least $1-\delta$*,*

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \tilde{\mathcal{O}}\Big(H^2 d^{3/2}\sqrt{T} + H^2 \mathfrak{B}^{\mathrm{R}}_{\mathrm{mis}}\zeta_0 T\sqrt{d}\Big). \tag{141}$$

The preceding corollary is the algorithm-specific absolute-regret result for the robust fitted recursion. For completeness, the following structural bridge controls arbitrary robust recommendation sequences through weighted RRD under a common positivity bound.

For the model-level absolute-regret consequence under behavioral misspecification, define the upper bounds on the recommendation exponent and effective inverse temperature, the uniform lower bound on ER probabilities, and the comparison constant by

$$\kappa_{\max} := \frac{\eta_{\max}\tau_{\max}}{1+\eta_{\max}\tau_{\max}}, \qquad \varepsilon_{\max} := \eta_{\max},$$

$$\underline{n}_{\mathrm{mis}} := B^{-1}\underline{y}_{\mathrm{mis}}^{\kappa_{\max}} e^{-\varepsilon_{\max}}, \qquad L^{\mathrm{mis}}_{\mathrm{dev}} := 1 + \frac{1+\eta_{\max}\tau_{\max}}{2\min\{\underline{y}_{\mathrm{mis}}, \underline{n}_{\mathrm{mis}}\}}.$$

Corollary R.7 (Model-level absolute-regret bridge under behavioral misspecification)**.** *Suppose Assumption R.1 holds. Let* $\{(\pi^t, y^t)\}_{t=1}^T$ *be any sequence of policy pairs that may depend on past data. Assume that the recommendations satisfy*

$$y_h^t(b \mid s) \ge \underline{y}_{\mathrm{mis}} \qquad \text{for every } t\in[T],\ h\in[H],\ s\in\mathcal{S},\ b\in\mathcal{B}.$$

Then the following deterministic bridge holds:

$$\mathrm{Reg}^{\mathrm{abs}}(T) \le \mathrm{Reg}(T) + 2L^{\mathrm{mis}}_{\mathrm{dev}}\,\mathrm{RRD}_{\mathrm{w,cum}}(T) \le \mathrm{Reg}(T) + 2HL^{\mathrm{mis}}_{\mathrm{dev}}\,\mathrm{RRD}_{\mathrm{cum}}(T). \tag{142}$$

Every sequence generated by the robust fitted recursion satisfies the displayed premise because each fitted QR recommendation has the common lower bound $\underline{y}_{\mathrm{mis}}$*.*

S Proofs for Behavioral Misspecification

Proof of Proposition R.2. Fix two admissible parameter tuples $(\omega, A, \theta, C, r, q_0)$ and $(\omega', A', \theta', C', r', q_0')$ in the definition of $\bar{\mathcal{W}}_h$, and let $\bar{F}_{\omega,A,\theta,C,r,q_0}$ and $\bar{F}_{\omega',A',\theta',C',r',q_0'}$ denote the corresponding robust value functions. The fitted softmax and covariance maps obey (105)–(106) with η replaced by $\bar{\eta}$.

Truncation and the supremum over $p\in\Delta(\mathcal{A})$ are nonexpansive in the sup norm. Consequently, the continuity bounds used in Proposition N.1 give

$$\big\|\bar{F}_{\omega,A,\theta,C,r,q_0} - \bar{F}_{\omega',A',\theta',C',r',q_0'}\big\|_\infty \le \|\omega-\omega'\|_2 + \bar{\beta}_h^{\mathrm{U}}\sqrt{\|A-A'\|_{\mathrm{op}}} + \frac{M_h\bar{\eta}}{2}\|\theta-\theta'\|_2$$
$$+ M_h L_{\bar{\eta}}|r-r'| + M_h L_{\bar{\eta}} R^T_{\mathrm{mis}}\sqrt{\|C-C'\|_{\mathrm{op}} + 3\bar{\eta}\,\|\theta-\theta'\|_2} + M_h|q_0-q_0'|.$$

The fourth term uses $\mathrm{Tr}(C\bar{\Gamma}^p_{\theta,h}(s)) \le 1$, and the fifth term uses $r' \le R^T_{\mathrm{mis}}$ together with (106). For every $t\in[T]$, one has $\Lambda_h^t \succeq I_d$ and $\bar{V}_h^t \succeq I_d$, so $(\Lambda_h^t)^{-1}$ and $(\bar{V}_h^t)^{-1}$ belong to $\mathbb{S}_+^d$ and are bounded

above by I_d. The ridge objective defining $\hat{\omega}_h^{t,\mathrm{mis}}$ at its minimizer is no larger than its value at the zero vector. Every robust regression target belongs to $[0, M_h]$, and therefore

$$\left\|\hat{\omega}_h^{t,\mathrm{mis}}\right\|_2^2 \le \sum_{i=1}^{t-1}\left[u_h(s_h^i, a_h^i, b_h^i) + \hat{W}_{h+1}^{t,\mathrm{mis}}(s_{h+1}^i)\right]^2 \le TM_h^2.$$

Moreover, $\hat{\theta}_h^t \in \Theta_h$, $R_{\mathrm{mis}}^t \le R_{\mathrm{mis}}^T$, and $\zeta_0/2 \in [0, \zeta_0/2]$. Consequently, for every value generated by the fitted recursion, the actual parameters satisfy all defining ranges of $\bar{\mathcal{W}}_h$ and are

$$(A, C, \theta, r, q_0) = \left((\Lambda_h^t)^{-1}, (\bar{V}_h^t)^{-1}, \hat{\theta}_h^t, R_{\mathrm{mis}}^t, \frac{\zeta_0}{2}\right), \tag{143}$$

with $\omega = \hat{\omega}_h^{t,\mathrm{mis}}$. Thus, the fitted recursion belongs to the claimed deterministic class, i.e., $\hat{W}_h^t \in \bar{\mathcal{W}}_h$. Use mesh $\bar{q}_h$ for ω, r, q_0 and mesh $\bar{q}_h^2$ for θ, A, C. Since $\bar{q}_h \le 1$, the preceding continuity bound makes the induced function error at most

$$\bar{q}_h + \bar{\beta}_h^{\mathrm{U}}\bar{q}_h + \frac{M_h\bar{\eta}}{2}\bar{q}_h^2 + M_hL_{\bar{\eta}}\bar{q}_h + M_hL_{\bar{\eta}}R_{\mathrm{mis}}^T\sqrt{1+3\bar{\eta}}\,\bar{q}_h + M_h\bar{q}_h \le \bar{K}_h^{\mathrm{cov}}\bar{q}_h = \frac{1}{8T} \le T^{-1}.$$

The six parameter nets have total cardinality at most

$$\left(1 + \frac{2\sqrt{T}M_h}{\bar{q}_h}\right)^d \left(1 + \frac{2B_\Theta}{\bar{q}_h^2}\right)^d \left(1 + \frac{2\sqrt{d}}{\bar{q}_h^2}\right)^{2m_d} \left(1 + \frac{2R_{\mathrm{mis}}^T}{\bar{q}_h}\right)\left(1 + \frac{\zeta_0}{\bar{q}_h}\right).$$

Taking logarithms gives exactly the five contributions in (125), where the two symmetric-matrix nets are combined in the coefficient $2m_d$.

We next prove the Bellman statement. Fix a deterministic robust continuation value $W \in \bar{\mathcal{W}}_{h+1}$ and abbreviate the observed feature vector as

$$\phi_h^i := \phi_h(s_h^i, a_h^i, b_h^i).$$

Define the ridge-regression estimator, Bellman parameter, and centered transition noise by

$$\begin{aligned}
\hat{\omega}_h^t[W] &:= (\Lambda_h^t)^{-1}\sum_{i=1}^{t-1}\phi_h^i\left[u_h(s_h^i, a_h^i, b_h^i) + W(s_{h+1}^i)\right],\\
\omega_h[W] &:= \vartheta_h^* + \sum_{s'\in\mathcal{S}}\mathbf{m}_h^*(\{s'\})W(s'),\\
\zeta_h^i(W) &:= W(s_{h+1}^i) - (P_hW)(s_h^i, a_h^i, b_h^i).
\end{aligned}$$

The bounded-variation condition in Definition 2.1 implies

$$\|\omega_h[W]\|_2 \le B_\vartheta + (H-h)B_P.$$

By the linear reward and transition identities,

$$u_h(s,a,b) + (P_hW)(s,a,b) = \phi_h(s,a,b)^\top\omega_h[W].$$

Substituting this identity into the ridge estimator gives

$$\hat{\omega}_h^t[W] - \omega_h[W] = (\Lambda_h^t)^{-1}\left[\sum_{i=1}^{t-1}\phi_h^i\zeta_h^i(W) - \omega_h[W]\right]. \tag{144}$$

The misspecified behavioral model does not change either the transition kernel or the deterministic leader reward. Consequently, (29) gives

$$\mathbb{E}[\zeta_h^i(W) \mid \mathcal{H}_h^i] = 0.$$

Because $0 \le W \le H - h$, the centered variable $\zeta_h^i(W)$ has conditional support in an interval of length at most $H - h$ and is therefore conditionally $(H-h)/2$-sub-Gaussian. For fixed h, define the inter-episode filtration $\bar{\mathcal{K}}_i^{(h)} := \mathcal{H}_h^{i+1}$ for $i \ge 0$. The inclusions $\mathcal{H}_h^i \subseteq \mathcal{F}_h^i \subseteq \mathcal{F}_H^i = \mathcal{F}_0^{i+1} \subseteq \mathcal{H}_h^{i+1}$ show that $\{\bar{\mathcal{K}}_i^{(h)}\}_{i\ge0}$ is a filtration. For every deterministic continuation function W, the design

vector ϕ_h^i is $\bar{\mathcal{K}}_{i-1}^{(h)}$-measurable, while $\zeta_h^i(W)$ is $\bar{\mathcal{K}}_i^{(h)}$-measurable and has conditional mean zero given $\bar{\mathcal{K}}_{i-1}^{(h)}$. Thus the predictable-design self-normalized inequality applies to $\{(\phi_h^i, \zeta_h^i(W))\}_{i\geq 1}$. For each h, fix a deterministic T^{-1}-net $\bar{\mathcal{N}}_{h+1}$ of the robust step-$(h+1)$ value class $\bar{\mathcal{W}}_{h+1}$ such that $|\bar{\mathcal{N}}_{h+1}| \leq \exp\{\bar{\mathfrak{L}}_{h+1}\}$. Set the failure probability for each net point to

$$\bar{\delta}'_h := \frac{\delta \exp\{-\bar{\mathfrak{L}}_{h+1}\}}{32H}.$$

For every $W \in \bar{\mathcal{N}}_{h+1}$, the self-normalized inequality gives, uniformly over $t \in [T]$,

$$\left\| \sum_{i=1}^{t-1} \phi_h^i \zeta_h^i(W) \right\|_{(\Lambda_h^t)^{-1}} \leq \frac{H-h}{2} \sqrt{2\left[d\log(1+T/d) + \log(1/\bar{\delta}'_h)\right]}$$

except on an event of probability at most $\bar{\delta}'_h$. A union bound over all net points and all $h \in [H]$ has failure probability at most

$$\sum_{h=1}^{H} |\bar{\mathcal{N}}_{h+1}| \bar{\delta}'_h \leq \frac{\delta}{32},$$

which is within the stated $\delta/4$ budget. Since $\log(1/\bar{\delta}'_h) = \bar{\mathfrak{L}}_{h+1} + \log(32H/\delta)$, the resulting simultaneous event satisfies

$$\left\| \sum_{i=1}^{t-1} \phi_h^i \zeta_h^i(W) \right\|_{(\Lambda_h^t)^{-1}} \leq \frac{H-h}{2} \sqrt{2\left[d\log(1+T/d) + \bar{\mathfrak{L}}_{h+1} + \log(32H/\delta)\right]}$$

for every t, every step, and every net point, on an event of failure probability at most $\delta/4$. Taking the Λ_h^t-norm in (144), using $\Lambda_h^t \succeq I_d$, and applying the triangle inequality show that the ridge regularizer contributes at most $B_\vartheta + (H-h)B_P$.

For an arbitrary robust W, choose a nearest net point W^0. The estimator extension (111) contributes at most $T^{-1/2}$, while the linear-transition target extension (112) contributes at most $\sqrt{T+1}B_P/T$. Their sum is $\chi_{T,h}$ when $h < H$ and is zero at the terminal step. Combining these four terms gives

$$\left\|\hat{\omega}_h^t[W] - \omega_h[W]\right\|_{\Lambda_h^t} \leq \bar{\beta}_h^{\mathrm{U}}.$$

Cauchy–Schwarz gives the pointwise confidence bound, and only after the uniform event has been established do we substitute the random continuation value generated by the robust recursion using (143). This proves (131).

Finally, the backward definitions starting from $\bar{\mathfrak{L}}_{H+1} = 0$ are finite and noncircular. Define the logarithmic factor $\ell := \log(eTHd/\delta)$ and the maximum logarithmic covering bound $\bar{\mathfrak{L}}_{\max} := \max_{1\leq h\leq H+1} \bar{\mathfrak{L}}_h$. On the fixed compact behavioral domain, $D_{\mathrm{beh}} \leq 1$ implies $R_{\mathrm{mis}}^T = O(\sqrt{d\ell + T})$. Equations (124)–(125) therefore yield

$$\bar{K}_h^{\mathrm{cov}} \leq C_0 H\big(\sqrt{T} + \sqrt{d\ell} + \sqrt{\bar{\mathfrak{L}}_{\max}}\big),$$
$$\bar{\mathfrak{L}}_{\max} \leq C_1 d^2\big[\ell + \log(1 + \bar{\mathfrak{L}}_{\max})\big],$$

with constants independent of h, H, T, d, δ. The two additional scalar nets contribute only logarithmic terms. Applying the scalar bound from the proof of Corollary E.6 gives, uniformly in h and H,

$$\bar{\mathfrak{L}}_h = O(d^2\ell), \qquad \bar{\beta}_h^{\mathrm{U}} = O\big((H-h)d\sqrt{\ell} + B_\vartheta + 1\big).$$

□

Proof of Corollary R.3. Fix $h \in [H]$ and define the realized scalar widths

$$\bar{z}_h^t := \sqrt{\phi_h(s_h^t, a_h^t, b_h^t)^\top (\Lambda_h^t)^{-1} \phi_h(s_h^t, a_h^t, b_h^t)}.$$

The scalar Gram-matrix recursion is unchanged by behavioral misspecification. Therefore the determinant argument in the proof of Lemma P.4 gives

$$\sum_{t=1}^{T} (\bar{z}_h^t)^2 \leq 2\log \frac{\det \Lambda_h^{T+1}}{\det I_d} \leq 2d\log(1+T/d),$$

and hence

$$\sum_{t=1}^{T} \bar{z}_h^t \le \sqrt{2Td\log(1+T/d)}.$$

Applying Lemma P.3 with $X^t = \bar{z}_h^t$ and $\alpha = \delta/(16H)$ gives

$$\sum_{t=1}^{T} \mathbb{E}^t \bar{z}_h^t \le \mathfrak{C}_{T,d,H}(\delta)$$

with failure probability at most $\delta/(16H)$.

For the fitted response widths, define the feature covariance matrix and its normalized version by

$$\bar{X}_h^t := \bar{\Gamma}_h^t(s_h^t, \pi_h^t), \qquad \bar{Y}_h^t := (\bar{V}_h^t)^{-1/2} \bar{X}_h^t (\bar{V}_h^t)^{-1/2}.$$

The estimated-preservation identity gives

$$\bar{V}_h^{t+1} = \bar{V}_h^t + \bar{X}_h^t.$$

Since $\bar{X}_h^t \succeq 0$, $\mathrm{Tr}(\bar{X}_h^t) \le 1$, and $\bar{V}_h^t \succeq I_d$, every eigenvalue of $\bar{Y}_h^t$ lies in $[0,1]$. Thus

$$\log\det(I_d + \bar{Y}_h^t) \ge \frac{1}{2}\mathrm{Tr}(\bar{Y}_h^t).$$

Telescoping the log determinants gives

$$\sum_{t=1}^{T} [\bar{\rho}_h^t(s_h^t, \pi_h^t)]^2 = \sum_{t=1}^{T} \mathrm{Tr}((\bar{V}_h^t)^{-1}\bar{X}_h^t) \le 2\log\det \bar{V}_h^{T+1} \le 2d\log(1+T/d).$$

Cauchy–Schwarz therefore yields

$$\sum_{t=1}^{T} \bar{\rho}_h^t(s_h^t, \pi_h^t) \le \sqrt{2Td\log(1+T/d)}.$$

Applying Lemma P.3 with $X^t = \bar{\rho}_h^t(s_h^t, \pi_h^t)$ and $\alpha = \delta/(16H)$ proves the second inequality in (132) with failure probability at most $\delta/(16H)$. A union bound over both width families and all $h \in [H]$ has failure probability at most $\delta/8$, which is within the stated $\delta/4$ budget. □

Lemma S.1 (Sensitivity to behavioral parameters)**.** *Fix $q \in [0,1]^B$ and $y \in \mathrm{int}\,\Delta(\mathcal{B})$ such that $y_b \ge \underline{y}_{\mathrm{mis}}$ for every $b \in \mathcal{B}$. Under Assumption R.1,*

$$\|z(\eta,\tau) - z(\bar{\eta},\bar{\tau})\|_\infty \le D_{\mathrm{beh}}. \tag{145}$$

If p and $\bar{p}$ are the resulting ERs with the same reward vector and recommendation, then

$$\max_{b\in\mathcal{B}} |\log[p(b)/\bar{p}(b)]| \le 2D_{\mathrm{beh}}, \quad \mathrm{KL}(p\|\bar{p}) \le \frac{1}{2}D_{\mathrm{beh}}^2, \quad \mathrm{TV}(p,\bar{p}) \le \frac{1}{2}D_{\mathrm{beh}}.$$

For quantal responses with the same reward vector,

$$\mathrm{TV}(\mu_\eta, \mu_{\bar{\eta}}) \le \zeta/2 \le \zeta_0/2. \tag{146}$$

Proof of Lemma S.1. Write the effective inverse temperature and recommendation exponent as functions of the behavioral parameters:

$$e(\eta,\tau) = \frac{\eta}{1+\eta\tau}, \qquad k(\eta,\tau) = \frac{\eta\tau}{1+\eta\tau}.$$

On the compact domain,

$$|\partial_\eta e| \le 1, \quad |\partial_\tau e| \le \eta_{\max}^2, \quad |\partial_\eta k| \le \tau_{\max}, \quad |\partial_\tau k| \le \eta_{\max}.$$

Since $|\log y_b| \le \log(1/\underline{y}_{\mathrm{mis}})$ and $q_b \in [0,1]$, the mean-value theorem proves (145). The response bounds follow from Lemma L.3. For QRs, the score vectors are ηq and $\bar{\eta} q$, whose sup-norm difference is at most $\zeta \le \zeta_0$. □

For episode i and step h, let $p_h^{i,*}$ be the true conditional follower distribution and let $\bar{p}_h^{i,\theta}$ be the fitted conditional distribution with reward parameter θ. Let $\bar{\Sigma}_h^{i,\theta}$ be the feature covariance under $\bar{p}_h^{i,\theta}$.

Lemma S.2 (Cumulative near-model log loss)**.** *Define the log-likelihood ratio by*

$$Z_h^i := \log \frac{p_h^{i,*}(b_h^i)}{\bar{p}_h^{i,\theta_h^*}(b_h^i)}.$$

If $D_{\mathrm{beh}} \le 1$, then with probability at least $1-\delta/8$, simultaneously for every $h \in [H]$ and $n \le T$,

$$\sum_{i=1}^{n} Z_h^i \le 2nD_{\mathrm{beh}}^2 + 4\log\frac{32HT}{\delta}. \tag{147}$$

Proof of Lemma S.2. Fix h and reuse the fixed-step inter-episode filtration $\mathcal{J}_i^{(h)} := \mathcal{G}_h^{i+1}$ for $i \ge 0$. Then Z_h^i is $\mathcal{J}_i^{(h)}$-measurable, while its conditional mean is taken with respect to $\mathcal{J}_{i-1}^{(h)} = \mathcal{G}_h^i$. The announced context and the conditional distributions $p_h^{i,*}$ and $\bar{p}_h^{i,\theta_h^*}$ are $\mathcal{J}_{i-1}^{(h)}$-measurable. Conditionally on that pre-action field,

$$\mathbb{E}[Z_h^i \mid \mathcal{J}_{i-1}^{(h)}] = \mathrm{KL}(p_h^{i,*} \| \bar{p}_h^{i,\theta_h^*}) \le D_{\mathrm{beh}}^2/2,$$

and $\left|Z_h^i\right| \le 2D_{\mathrm{beh}}$ by Lemma S.1. For $X_h^i := Z_h^i - \mathbb{E}[Z_h^i \mid \mathcal{J}_{i-1}^{(h)}]$, the sequence $\{X_h^i\}_{i\ge1}$ is a martingale-difference sequence relative to $\{\mathcal{J}_i^{(h)}\}_{i\ge0}$, with $\left|X_h^i\right| \le 3D_{\mathrm{beh}}$ and $\sum_{i=1}^n \mathbb{E}[(X_h^i)^2 \mid \mathcal{J}_{i-1}^{(h)}] \le 4nD_{\mathrm{beh}}^2$. The conditional Bernstein–Freedman inequality for martingale differences bounded above by b, with deterministic conditional-variance bound v, states that

$$S_n \le \sqrt{2vL} + \frac{bL}{3}$$

with failure probability at most e^{-L}. Applying it with $b = 3D_{\mathrm{beh}}$ and the displayed variance bound, the inequality

$$\sum_{i=1}^{n} X_h^i \le \sqrt{8nD_{\mathrm{beh}}^2 L} + D_{\mathrm{beh}} L$$

holds with failure probability at most e^{-L}. Taking $L = \log(32HT/\delta)$ and union bounding over (h,n) spends at most $\delta/32$. Consequently,

$$\sum_{i=1}^{n} Z_h^i \le \frac{nD_{\mathrm{beh}}^2}{2} + \sqrt{8nD_{\mathrm{beh}}^2 L} + D_{\mathrm{beh}} L, \qquad L = \log\frac{32HT}{\delta}.$$

Let $x = nD_{\mathrm{beh}}^2$. Young's inequality gives

$$\sqrt{8xL} \le \frac{3}{2}x + \frac{4}{3}L, \qquad D_{\mathrm{beh}} L \le L.$$

Substitution yields $\sum_{i=1}^n Z_h^i \le 2x + (7/3)L$, which is stronger than (147). □

Lemma S.3 (MLE concentration under behavioral misspecification)**.** *If $D_{\mathrm{beh}} \le 1$, then with probability at least $1-\delta/4$, simultaneously for every $t \in [T]$ and $h \in [H]$,*

$$\sum_{i=1}^{t-1} \left\| \hat{\theta}_h^t - \theta_h^* \right\|_{\bar{\Sigma}_h^{i,\hat{\theta}_h^t}}^2 \le C_{\mathrm{mis}}[\bar{\beta}_\Theta(\delta) + (t-1)D_{\mathrm{beh}}^2]. \tag{148}$$

Consequently,

$$\left\| \hat{\theta}_h^t - \theta_h^* \right\|_{\bar{V}_h^t} \le R_{\mathrm{mis}}^t. \tag{149}$$

Proof of Lemma S.3. The square-root likelihood-ratio martingale used in Lemma M.2 does not require the reference distribution to belong to the fitted family. For fixed h, θ, it gives, uniformly in t,

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, \bar{p}_h^{i,\theta}) \le \frac{1}{2}[\bar{\ell}_h^t(\theta) - \ell_h^{t,*}] + \log(1/\delta'),$$

where the negative log-likelihood under the follower's true ER is $\ell_h^{t,*} := -\sum_{i=1}^{t-1} \log p_h^{i,*}(b_h^i)$. Take an $r = (8\bar{\varepsilon}T)^{-1}$-net $\bar{\mathcal{N}}_h$ of Θ_h, whose cardinality is at most

$$\bar{N}_\Theta \le (1 + 16B_\Theta \bar{\varepsilon} T)^d,$$

and set the failure probability for each net point to $\delta' = \delta/(32HT\bar{N}_\Theta)$. Fix $\theta \in \Theta_h$ and choose $\theta^0 \in \bar{\mathcal{N}}_h$ with $\left\|\theta - \theta^0\right\|_2 \le r$. For every historical context and every $b \in \mathcal{B}$,

$$\left|\log \bar{p}_h^{i,\theta}(b) - \log \bar{p}_h^{i,\theta^0}(b)\right| \le 2\bar{\varepsilon} r =: D = \frac{1}{4T}.$$

Therefore,

$$\left|\bar{\ell}_h^t(\theta) - \bar{\ell}_h^t(\theta^0)\right| \le (t-1)D \le \frac{1}{4}.$$

For the log-probability differences d_b^i, write

$$\bar{p}_h^{i,\theta}(b) = \bar{p}_h^{i,\theta^0}(b) e^{d_b^i}, \qquad \left|d_b^i\right| \le D.$$

Then

$$\left|D_{\mathrm{H}}^2(p_h^{i,*}, \bar{p}_h^{i,\theta}) - D_{\mathrm{H}}^2(p_h^{i,*}, \bar{p}_h^{i,\theta^0})\right| \le \sum_{b\in\mathcal{B}} \sqrt{p_h^{i,*}(b)\bar{p}_h^{i,\theta^0}(b)}\left|e^{d_b^i/2} - 1\right| \le e^{D/2}\frac{D}{2}.$$

Summing over $i < t$ gives a cumulative Hellinger change of at most $Te^{D/2}D/2 \le e^{1/8}/8$. The change in one half of the fitted likelihood difference is at most $1/8$. Moreover,

$$\log(1/\delta') \le d\log(1 + 16B_\Theta\bar{\varepsilon}T) + \log\frac{32HT}{\delta}.$$

Thus the extension from the net point to θ costs less than one, which is absorbed by the $+4$ term in $\bar{\beta}_\Theta(\delta)$. A union bound over steps and net points therefore yields

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, \bar{p}_h^{i,\theta}) \le \frac{1}{2}[\bar{\ell}_h^t(\theta) - \ell_h^{t,*}] + \frac{1}{4}\bar{\beta}_\Theta(\delta) \tag{150}$$

simultaneously for every $t \in [T]$, $h \in [H]$, and $\theta \in \Theta_h$.

Set $n = t - 1$ and $\theta = \hat{\theta}_h^t$. MLE optimality and (150) give

$$\begin{aligned}\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, \bar{p}_h^{i,\hat{\theta}_h^t}) &\le \frac{1}{2}[\bar{\ell}_h^t(\theta_h^*) - \ell_h^{t,*}] + \frac{1}{4}\bar{\beta}_\Theta(\delta) = \frac{1}{2}\sum_{i=1}^{t-1} Z_h^i + \frac{1}{4}\bar{\beta}_\Theta(\delta) \\ &\le nD_{\mathrm{beh}}^2 + 2\log\frac{32HT}{\delta} + \frac{1}{4}\bar{\beta}_\Theta(\delta) \le nD_{\mathrm{beh}}^2 + \bar{\beta}_\Theta(\delta).\end{aligned}$$

The last inequality follows from the definition

$$\bar{\beta}_\Theta(\delta) = 4\left[d\log(1 + 16B_\Theta\bar{\varepsilon}T) + \log\frac{32HT}{\delta} + 4\right],$$

which implies

$$2\log\frac{32HT}{\delta} + \frac{1}{4}\bar{\beta}_\Theta(\delta) \le \bar{\beta}_\Theta(\delta).$$

Also

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, \bar{p}_h^{i,\theta_h^*}) \le \sum_{i=1}^{t-1} \mathrm{KL}(p_h^{i,*} \| \bar{p}_h^{i,\theta_h^*}) \le nD_{\mathrm{beh}}^2/2.$$

Because Hellinger distance is a metric,

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(\bar{p}_h^{i,\hat{\theta}_h^t}, \bar{p}_h^{i,\theta_h^*}) \le 2\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(\bar{p}_h^{i,\hat{\theta}_h^t}, p_h^{i,*}) + 2\sum_{i=1}^{t-1} D_{\mathrm{H}}^2(p_h^{i,*}, \bar{p}_h^{i,\theta_h^*}) \le 2\bar{\beta}_\Theta(\delta) + 3nD_{\mathrm{beh}}^2.$$

Apply Lemma L.4 inside the fitted family with inverse temperature $\bar{\varepsilon}$ and with $\bar{p}_h^{i,\hat{\theta}_h^t}$ as covariance reference. The choice $C_{\mathrm{mis}} = 8/c_{\mathrm{H}}(\bar{\varepsilon})$ proves (148).

Finally, fitted-family covariance comparison gives $\bar{\Sigma}_h^{i,\hat{\theta}_h^i} \preceq e^{2\bar{\varepsilon}}\bar{\Sigma}_h^{i,\hat{\theta}_h^t}$. Combine this with the diameter bound to obtain (149). □

Proof of Corollary R.5. Fix $t \in [T]$ and $h \in [H]$, and set $n = t - 1$. Apply (150) with $\theta = \tilde{\theta}_h^t$. The certificate (137) implies

$$\bar{\ell}_h^t(\tilde{\theta}_h^t) \le \bar{\ell}_h^t(\theta_h^*) + \bar{\gamma}_{\mathrm{MLE}}.$$

Substituting this inequality into (150) gives

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2\left(p_h^{i,*}, \bar{p}_h^{i,\tilde{\theta}_h^t}\right) \le \frac{1}{2}\left[\bar{\ell}_h^t(\theta_h^*) - \ell_h^{t,*}\right] + \frac{1}{2}\bar{\gamma}_{\mathrm{MLE}} + \frac{1}{4}\bar{\beta}_\Theta(\delta) = \frac{1}{2}\sum_{i=1}^{t-1} Z_h^i + \frac{1}{2}\bar{\gamma}_{\mathrm{MLE}} + \frac{1}{4}\bar{\beta}_\Theta(\delta).$$

Using (147) and the inequality

$$2\log\frac{32HT}{\delta} + \frac{1}{4}\bar{\beta}_\Theta(\delta) \le \bar{\beta}_\Theta(\delta)$$

therefore yields

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2\left(p_h^{i,*}, \bar{p}_h^{i,\tilde{\theta}_h^t}\right) \le nD_{\mathrm{beh}}^2 + \bar{\beta}_\Theta(\delta) + \frac{1}{2}\bar{\gamma}_{\mathrm{MLE}}.$$

The behavioral-sensitivity bound also gives

$$\sum_{i=1}^{t-1} D_{\mathrm{H}}^2\left(p_h^{i,*}, \bar{p}_h^{i,\theta_h^*}\right) \le \frac{n}{2}D_{\mathrm{beh}}^2.$$

Because Hellinger distance is a metric and $(a+b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned}\sum_{i=1}^{t-1} D_{\mathrm{H}}^2\left(\bar{p}_h^{i,\tilde{\theta}_h^t}, \bar{p}_h^{i,\theta_h^*}\right) &\le 2\sum_{i=1}^{t-1} D_{\mathrm{H}}^2\left(\bar{p}_h^{i,\tilde{\theta}_h^t}, p_h^{i,*}\right) + 2\sum_{i=1}^{t-1} D_{\mathrm{H}}^2\left(p_h^{i,*}, \bar{p}_h^{i,\theta_h^*}\right) \\ &\le 2\bar{\beta}_\Theta(\delta) + 3nD_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}}.\end{aligned}$$

Apply Lemma L.4 inside the fitted family, with inverse temperature $\bar{\varepsilon}$ and covariance reference $\bar{p}_h^{i,\tilde{\theta}_h^t}$. This gives

$$\sum_{i=1}^{t-1} \left\|\tilde{\theta}_h^t - \theta_h^*\right\|_{\bar{\Sigma}_h^{i,\tilde{\theta}_h^t}}^2 \le \frac{2\bar{\beta}_\Theta(\delta) + 3nD_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}}}{c_{\mathrm{H}}(\bar{\varepsilon})}.$$

Since

$$2\bar{\beta}_\Theta(\delta) + 3nD_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}} \le 8\left[\bar{\beta}_\Theta(\delta) + nD_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}}\right],$$

and $C_{\mathrm{mis}} = 8/c_{\mathrm{H}}(\bar{\varepsilon})$, the preceding display implies

$$\sum_{i=1}^{t-1} \left\|\tilde{\theta}_h^t - \theta_h^*\right\|_{\bar{\Sigma}_h^{i,\tilde{\theta}_h^t}}^2 \le C_{\mathrm{mis}}\left[\bar{\beta}_\Theta(\delta) + nD_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}}\right].$$

For every $i < t$, fitted-family covariance comparison gives

$$\bar{\Sigma}_h^{i,\tilde{\theta}_h^i} \preceq e^{2\bar{\varepsilon}}\bar{\Sigma}_h^{i,\tilde{\theta}_h^t}.$$

Using $\left\|\tilde{\theta}_h^t - \theta_h^*\right\|_2 \le B_\Theta$ therefore yields

$$\left\|\tilde{\theta}_h^t - \theta_h^*\right\|_{\bar{V}_h^{t,\gamma}}^2 \le B_\Theta^2 + e^{2\bar{\varepsilon}}C_{\mathrm{mis}}\left[\bar{\beta}_\Theta(\delta) + nD_{\mathrm{beh}}^2 + \bar{\gamma}_{\mathrm{MLE}}\right] = [R_{\mathrm{mis}}^{t,\gamma}]^2.$$

This proves (138) on the joint predictive-Hellinger and near-model log-loss event, whose failure probability is at most $\delta/4$.

The estimated-preservation identity gives the deterministic update

$$\bar{V}_h^{t+1,\gamma} = \bar{V}_h^{t,\gamma} + \bar{\Gamma}_h^{t,\gamma}(s_h^t, \pi_h^t).$$

Hence the scalar and matrix determinant-potential identities hold for every realized approximate fitted sequence. The ellipsoidal inequality just proved and the estimated-preservation identity imply

the two robust response bounds with the substitutions stated in the corollary. The robust generated-value envelope is recomputed with $R^{T,\gamma}_{\mathrm{mis}}$ and $\bar{\beta}^{\mathrm{U}}_{h,\gamma}$ before interaction and is therefore deterministic. Applying the proof of Proposition R.2 to this recomputed envelope gives a Bellman-confidence event with failure probability at most $\delta/4$. Applying Lemma P.3 to the approximate fitted width sequences gives the potential event with failure probability at most $\delta/4$. On the intersection of the MLE, Bellman, and potential events, the proof of Theorem R.4 applies with the replacements listed in the corollary. The proof of Corollary R.6 uses the same robust response bounds and therefore applies with the same replacements. The total failure probability is at most $3\delta/4 < \delta$, which completes the proof. □

Proposition S.4 (Validity of robust response bonuses)**.** *On the event of Lemma S.3, for every* $t \in [T]$, $h \in [H]$, $s \in \mathcal{S}$, *and* $p \in \Delta(\mathcal{A})$,

$$\mathrm{TV}\Big(\bar{y}^{t,p}_h(\cdot \mid s), \nu^{p,\bar{y}^{t,p}_h}_{\theta^*_h,h}(\cdot \mid s)\Big) \le b^{t,\mathrm{F},mis}_h(s,p), \tag{151}$$

$$\mathrm{TV}\Big(\bar{\mu}^{p}_{\hat{\theta}^t_h,h}(\cdot \mid s), \mu^{p}_{\theta^*_h,h}(\cdot \mid s)\Big) \le b^{t,\mathrm{Q},mis}_h(s,p). \tag{152}$$

Proof of Proposition S.4. Fix t, h, s, p. The estimated-preservation identity gives

$$\bar{y}^{t,p}_h(\cdot \mid s) = \bar{\nu}^{p,\bar{y}^{t,p}_h}_{\hat{\theta}^t_h,h}(\cdot \mid s).$$

Therefore,

$$\mathrm{TV}\Big(\bar{y}^{t,p}_h(\cdot \mid s), \nu^{p,\bar{y}^{t,p}_h}_{\theta^*_h,h}(\cdot \mid s)\Big) \le \mathrm{TV}\Big(\bar{\nu}^{p,\bar{y}^{t,p}_h}_{\hat{\theta}^t_h,h}(\cdot \mid s), \bar{\nu}^{p,\bar{y}^{t,p}_h}_{\theta^*_h,h}(\cdot \mid s)\Big) + \mathrm{TV}\Big(\bar{\nu}^{p,\bar{y}^{t,p}_h}_{\theta^*_h,h}(\cdot \mid s), \nu^{p,\bar{y}^{t,p}_h}_{\theta^*_h,h}(\cdot \mid s)\Big).$$

By Lemmas L.2, S.3 and L.5, the first term is at most $L_{\bar{\varepsilon}} R^t_{\mathrm{mis}} \bar{\rho}^t_h(s,p)$. By Lemma S.1, the second term is at most $D_{\mathrm{beh}}/2$. Hence

$$\mathrm{TV}(\bar{y}^{t,p}_h, \nu^{p,\bar{y}^{t,p}_h}_{\theta^*_h,h}) \le L_{\bar{\varepsilon}} R^t_{\mathrm{mis}} \bar{\rho}^t_h(s,p) + D_{\mathrm{beh}}/2.$$

Taking the minimum with one proves the follower-response inequality (151).

For the QR inequality, insert the fitted QR at the true reward parameter:

$$\mathrm{TV}\Big(\bar{\mu}^{p}_{\hat{\theta}^t_h,h}(\cdot \mid s), \mu^{p}_{\theta^*_h,h}(\cdot \mid s)\Big) \le \mathrm{TV}\Big(\bar{\mu}^{p}_{\hat{\theta}^t_h,h}(\cdot \mid s), \bar{\mu}^{p}_{\theta^*_h,h}(\cdot \mid s)\Big) + \mathrm{TV}\Big(\bar{\mu}^{p}_{\theta^*_h,h}(\cdot \mid s), \mu^{p}_{\theta^*_h,h}(\cdot \mid s)\Big).$$

The first term is bounded by $L_{\bar{\eta}} R^t_{\mathrm{mis}} \bar{\rho}^t_h(s,p)$ through Lemmas L.2, S.3 and L.5 applied to the fitted QR family. The second term is at most $\zeta_0/2$ by (146). Taking the minimum with one proves the QR inequality (152). □

Proof of Theorem R.4. Let $\bar{\mathcal{E}}_{\mathrm{MLE}}$, $\bar{\mathcal{E}}_{\mathrm{Bell}}$, and $\bar{\mathcal{E}}_{\mathrm{pot}}$ be the simultaneous events in Lemma S.3, proposition R.2, and corollary R.3, respectively. Each event has failure probability at most $\delta/4$. Therefore,

$$\mathbb{P}(\bar{\mathcal{E}}_{\mathrm{MLE}} \cap \bar{\mathcal{E}}_{\mathrm{Bell}} \cap \bar{\mathcal{E}}_{\mathrm{pot}}) \ge 1 - 3\delta/4 \ge 1 - \delta.$$

We work on this intersection.

We first prove optimism for the fitted robust recursion. At step $H+1$, $\hat{W}^{t,\mathrm{mis}}_{H+1} = W^*_{H+1} = 0$. Assume that $\hat{W}^{t,\mathrm{mis}}_{h+1}(s') \ge W^*_{h+1}(s')$ for every $s' \in \mathcal{S}$. The transition kernel is positive, so

$$P_h \hat{W}^{t,\mathrm{mis}}_{h+1} \ge P_h W^*_{h+1}.$$

On $\bar{\mathcal{E}}_{\mathrm{Bell}}$, (131) and the definition of $\hat{U}^{t,\mathrm{mis}}_h$ give

$$\hat{U}^{t,\mathrm{mis}}_h(s,a,b) \ge u_h(s,a,b) + (P_h W^*_{h+1})(s,a,b) = U^*_h(s,a,b).$$

Fix s and $p \in \Delta(\mathcal{A})$. The robust QR inequality in Proposition S.4 gives

$$\mathrm{TV}\Big(\bar{y}^{t,p}_h(\cdot \mid s), \mu^{p}_{\theta^*_h,h}(\cdot \mid s)\Big) \le b^{t,\mathrm{Q},mis}_h(s,p).$$

Because $0 \le U_h^* \le M_h$, changing the follower distribution from the true QR to the fitted QR changes its expectation by at most $M_h b_h^{t,\mathrm{Q},mis}(s,p)$. Using $\hat{U}_h^{t,\mathrm{mis}} \ge U_h^*$, we obtain

$$\bar{\mathcal{Q}}_h^t(s,p) \ge \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} p(a)\mu^p_{\theta_h^*,h}(b \mid s) U_h^*(s,a,b).$$

The fitted recursion maximizes $\bar{\mathcal{Q}}_h^t(s,p)$ over $p \in \Delta(\mathcal{A})$. Monotone truncation therefore gives

$$\hat{W}_h^{t,\mathrm{mis}}(s) \ge W_h^*(s).$$

Backward induction proves robust optimism for every episode, step, and state.

Fix episode t and let $\nu_h^t := \nu^{\pi^t,y^t}_{\theta_h^*,h}$ be the true executed response. Robust optimism implies

$$J^* - J(\pi^t, y^t) \le \mathbb{E}_{\rho_0}[\hat{W}_1^{t,\mathrm{mis}}(s_1)] - J(\pi^t, y^t).$$

Apply Lemma P.1 with

$$\bar{\nu}_h = y_h^t, \qquad \bar{U}_h = \hat{U}_h^{t,\mathrm{mis}}, \qquad \bar{W}_h = \hat{W}_h^{t,\mathrm{mis}}.$$

We bound its three groups separately.

For the Bellman group, define the Bellman target function by

$$\bar{g}_h^t(s,a,b) := u_h(s,a,b) + (P_h \hat{W}_{h+1}^{t,\mathrm{mis}})(s,a,b).$$

The function $\bar{g}_h^t$ takes values in $[0, M_h]$. On $\bar{\mathcal{E}}_{\mathrm{Bell}}$, the robust Bellman confidence bound and truncation imply

$$\hat{U}_h^{t,\mathrm{mis}}(s,a,b) - \bar{g}_h^t(s,a,b) \le 2\bar{\beta}_h^{\mathrm{U}} \sqrt{\phi_h(s,a,b)^\top (\Lambda_h^t)^{-1} \phi_h(s,a,b)}.$$

Taking $\mathbb{E}^t$ controls the Bellman group.

For the policy-improvement group, the fitted recommendation is $y_h^t(\cdot \mid s) = \bar{y}_h^{t,\pi_h^t}(\cdot \mid s)$. By the definition of $\bar{\mathcal{Q}}_h^t$ and the monotone truncation inequality,

$$\hat{W}_h^{t,\mathrm{mis}}(s) - (\pi_h^t \otimes y_h^t) \circ \hat{U}_h^{t,\mathrm{mis}}(s) \le M_h b_h^{t,\mathrm{Q},mis}(s, \pi_h^t).$$

For the response group, $0 \le \hat{U}_h^{t,\mathrm{mis}} \le M_h$. The variational characterization of total variation and the robust follower-response inequality give

$$\pi_h^t \otimes (y_h^t - \nu_h^t) \circ \hat{U}_h^{t,\mathrm{mis}}(s) \le M_h \,\mathrm{TV}\big(y_h^t(\cdot \mid s), \nu_h^t(\cdot \mid s)\big) \le M_h b_h^{t,\mathrm{F},mis}(s, \pi_h^t).$$

Summing the three groups over episodes and steps, and then applying (132), gives

$$\mathrm{Reg}(T) \le 2\mathfrak{C}_{T,d,H}(\delta) \sum_{h=1}^H \bar{\beta}_h^{\mathrm{U}} + \sum_{h=1}^H M_h \sum_{t=1}^T \mathbb{E}^t\Big[b_h^{t,\mathrm{F},mis}(s_h^t, \pi_h^t) + b_h^{t,\mathrm{Q},mis}(s_h^t, \pi_h^t)\Big].$$

Since truncation at one only decreases a bonus,

$$b_h^{t,\mathrm{F},mis}(s,p) + b_h^{t,\mathrm{Q},mis}(s,p) \le (L_{\bar{\varepsilon}} + L_{\bar{\eta}}) R_{\mathrm{mis}}^t \bar{\rho}_h^t(s,p) + \frac{1}{2}(D_{\mathrm{beh}} + \zeta_0).$$

The radius bound

$$R_{\mathrm{mis}}^t \le R_{\mathrm{mis}}^0 + C_R D_{\mathrm{beh}} \sqrt{t-1}$$

and $\sqrt{t-1} \le \sqrt{T}$ imply

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}^t\big[R_{\mathrm{mis}}^t \bar{\rho}_h^t(s_h^t, \pi_h^t)\big] &\le R_{\mathrm{mis}}^0 \sum_{t=1}^T \mathbb{E}^t \bar{\rho}_h^t(s_h^t, \pi_h^t) + C_R D_{\mathrm{beh}} \sqrt{T} \sum_{t=1}^T \mathbb{E}^t \bar{\rho}_h^t(s_h^t, \pi_h^t) \\ &\le R_{\mathrm{mis}}^0 \mathfrak{C}_{T,d,H}(\delta) + C_R D_{\mathrm{beh}} \mathfrak{D}_{T,d,H}(\delta) = \mathfrak{E}^{\mathrm{mis}}_{T,d,H}(\delta).\end{aligned}$$

Substituting this inequality proves (134).

For cumulative RRD, the robust follower-response inequality gives, episode by episode,

$$\mathrm{RRD}(\pi^t, y^t; \theta^*) \le \sum_{h=1}^{H} \mathbb{E}^t\left[L_{\bar{\varepsilon}} R^t_{\mathrm{mis}} \bar{\rho}^t_h(s^t_h, \pi^t_h) + \frac{1}{2} D_{\mathrm{beh}}\right].$$

Summing over t and using the preceding bound for every h gives

$$\mathrm{RRD}_{\mathrm{cum}}(T) \le H\left\{L_{\bar{\varepsilon}} \mathfrak{E}^{\mathrm{mis}}_{T,d,H}(\delta) + \frac{T}{2} D_{\mathrm{beh}}\right\},$$

which is (135). Multiplying the step-h summand by M_h before summation gives

$$\mathrm{RRD}_{\mathrm{w,cum}}(T) \le S_H\left\{L_{\bar{\varepsilon}} \mathfrak{E}^{\mathrm{mis}}_{T,d,H}(\delta) + \frac{T}{2} D_{\mathrm{beh}}\right\},$$

which is (136).

Finally,

$$R^0_{\mathrm{mis}} = \tilde{\mathcal{O}}(\sqrt{d}), \qquad \bar{\beta}^{\mathrm{U}}_h = \tilde{\mathcal{O}}((H-h)d + B_\vartheta + 1), \qquad D_{\mathrm{beh}} = C_{\mathrm{beh}}\zeta_0.$$

Together with $\mathfrak{C}_{T,d,H}(\delta) = \tilde{\mathcal{O}}(\sqrt{Td})$, $\mathfrak{D}_{T,d,H}(\delta) = \tilde{\mathcal{O}}(T\sqrt{d})$, and $S_H = H(H+1)/2$, these relations give the simplified rates stated after the theorem. □

Proof of Corollary R.6. For episode t, define the leader suboptimality, ER policy, and QR policy by

$$\Delta^t := J^* - J(\pi^t, y^t), \qquad \nu^t := \nu^{\pi^t, y^t}_{\theta^*}, \qquad \mu^t := \mu^{\pi^t}_{\theta^*}.$$

By Definition 3.4, (π^t, μ^t) satisfies the preservation constraint, and (4) gives $\nu^{\pi^t,\mu^t}_{\theta^*} = \mu^t$ for every $\tau \ge 0$. It is therefore feasible for the benchmark in (8), so

$$J^{\mathrm{pol}}(\pi^t, \mu^t) \le J^*.$$

Using $J(\pi^t, y^t) = J^{\mathrm{pol}}(\pi^t, \nu^t)$ gives

$$[-\Delta^t]_+ = \left[J(\pi^t, y^t) - J^*\right]_+ \le \left[J^{\mathrm{pol}}(\pi^t, \nu^t) - J^{\mathrm{pol}}(\pi^t, \mu^t)\right]_+.$$

Applying (121) to the last display yields

$$[-\Delta^t]_+ \le \sum_{h=1}^{H} M_h \mathbb{E}^t\left[\mathrm{TV}\left(\nu^t_h(\cdot \mid s^t_h), \mu^t_h(\cdot \mid s^t_h)\right)\right].$$

The robust fitted recursion satisfies

$$y^t_h(\cdot \mid s) = \bar{\mu}^{\pi^t}_{\hat{\theta}^t_h, h}(\cdot \mid s) = \bar{\nu}^{\pi^t, y^t}_{\hat{\theta}^t_h, h}(\cdot \mid s).$$

For every visited state, the triangle inequality gives

$$\mathrm{TV}\left(\nu^t_h(\cdot \mid s^t_h), \mu^t_h(\cdot \mid s^t_h)\right) \le \mathrm{TV}\left(\nu^t_h(\cdot \mid s^t_h), y^t_h(\cdot \mid s^t_h)\right) + \mathrm{TV}\left(y^t_h(\cdot \mid s^t_h), \mu^t_h(\cdot \mid s^t_h)\right).$$

On the event of Proposition S.4, the two terms are bounded by $b^{t,\mathrm{F},mis}_h(s^t_h, \pi^t_h)$ and $b^{t,\mathrm{Q},mis}_h(s^t_h, \pi^t_h)$, respectively. Consequently,

$$[-\Delta^t]_+ \le \sum_{h=1}^{H} M_h \mathbb{E}^t\left[b^{t,\mathrm{F},mis}_h(s^t_h, \pi^t_h) + b^{t,\mathrm{Q},mis}_h(s^t_h, \pi^t_h)\right].$$

Since $|x| = x + 2[-x]_+$ for every real number x, summing the preceding inequality over $t \in [T]$ proves (139). The robust bonus definitions also imply

$$b^{t,\mathrm{F},mis}_h(s,p) + b^{t,\mathrm{Q},mis}_h(s,p) \le (L_{\bar{\varepsilon}} + L_{\bar{\eta}}) R^t_{\mathrm{mis}} \bar{\rho}^t_h(s,p) + \frac{1}{2}(D_{\mathrm{beh}} + \zeta_0).$$

For every $h \in [H]$, Corollary R.3 and the definition of $\mathfrak{E}^{\mathrm{mis}}_{T,d,H}(\delta)$ give

$$\sum_{t=1}^{T} \mathbb{E}^t\left[R^t_{\mathrm{mis}} \bar{\rho}^t_h(s^t_h, \pi^t_h)\right] \le \mathfrak{E}^{\mathrm{mis}}_{T,d,H}(\delta).$$

Therefore,

$$\sum_{t=1}^{T}\sum_{h=1}^{H} M_h \mathbb{E}^t\left[b_h^{t,\mathrm{F},mis}(s_h^t,\pi_h^t)+b_h^{t,\mathrm{Q},mis}(s_h^t,\pi_h^t)\right] \le S_H\left\{(L_{\bar{\varepsilon}}+L_{\bar{\eta}})\mathfrak{E}^{\mathrm{mis}}_{T,d,H}(\delta)+\frac{T}{2}(D_{\mathrm{beh}}+\zeta_0)\right\}.$$

Combining this inequality with (134) and (139) proves (140). Finally, substitute $R^0_{\mathrm{mis}}=\tilde{\mathcal{O}}(\sqrt{d})$, $D_{\mathrm{beh}}=C_{\mathrm{beh}}\zeta_0$, $\mathfrak{C}_{T,d,H}(\delta)=\tilde{\mathcal{O}}(\sqrt{Td})$, and $S_H=H(H+1)/2$ into (140). The resulting statistical term is $\tilde{\mathcal{O}}(H^2d^{3/2}\sqrt{T})$, and the calibration term is the one displayed in (141). □

Proof of Corollary R.7. Fix any sequence of policy pairs satisfying the common recommendation lower bound in the corollary. For the true executed response, Assumption R.1 gives $\kappa\le\kappa_{\max}$ and $\varepsilon\le\varepsilon_{\max}$. Since $0<\underline{y}_{\mathrm{mis}}\le 1$ and $\kappa\le\kappa_{\max}$, one has $\underline{y}_{\mathrm{mis}}^{\kappa}\ge\underline{y}_{\mathrm{mis}}^{\kappa_{\max}}$. Since $\varepsilon\le\varepsilon_{\max}$, one also has $e^{-\varepsilon}\ge e^{-\varepsilon_{\max}}$. Therefore, Lemma C.2 gives

$$\nu^{\pi^t,y^t}_{\theta_h^*,h}(b\mid s)\ge|\mathcal{B}|^{-1}\underline{y}_{\mathrm{mis}}^{\kappa}e^{-\varepsilon}\ge|\mathcal{B}|^{-1}\underline{y}_{\mathrm{mis}}^{\kappa_{\max}}e^{-\varepsilon_{\max}}=\underline{n}_{\mathrm{mis}}$$

for every t,h,s,b. Moreover,

$$\lambda_{\mathrm{inv}}=1+\eta\tau\le 1+\eta_{\max}\tau_{\max}.$$

For episode t, define the leader suboptimality, ER policy, and QR policy by

$$\Delta^t:=J^*-J(\pi^t,y^t),\qquad \nu^t:=\nu^{\pi^t,y^t}_{\theta^*},\qquad \mu^t:=\mu^{\pi^t}_{\theta^*}.$$

By Definition 3.4, (π^t,μ^t) satisfies the preservation constraint, and (4) gives $\nu^{\pi^t,\mu^t}_{\theta^*}=\mu^t$ for every $\tau\ge 0$. It is therefore feasible for the benchmark in (8). Hence

$$J^{\mathrm{pol}}(\pi^t,\mu^t)\le J^*.$$

Using $J(\pi^t,y^t)=J^{\mathrm{pol}}(\pi^t,\nu^t)$ and Lemma Q.1,

$$[-\Delta^t]_+=[J^{\mathrm{pol}}(\pi^t,\nu^t)-J^*]_+\le[J^{\mathrm{pol}}(\pi^t,\nu^t)-J^{\mathrm{pol}}(\pi^t,\mu^t)]_+ \le\sum_{h=1}^{H}M_h\mathbb{E}^t\left[\mathrm{TV}\big(\nu_h^t(\cdot\mid s_h^t),\mu_h^t(\cdot\mid s_h^t)\big)\right].$$

The positivity bounds above and Proposition C.3 give, for every visited state,

$$\mathrm{TV}\big(\mu_h^t(\cdot\mid s_h^t),\nu_h^t(\cdot\mid s_h^t)\big)\le L^{\mathrm{mis}}_{\mathrm{dev}}\,\mathrm{TV}\big(y_h^t(\cdot\mid s_h^t),\nu_h^t(\cdot\mid s_h^t)\big).$$

Therefore,

$$[-\Delta^t]_+\le L^{\mathrm{mis}}_{\mathrm{dev}}\,\mathrm{RRD}_{\mathrm{w}}(\pi^t,y^t;\theta^*).$$

Since $|x|=x+2[-x]_+$ for every real number x,

$$\left|\Delta^t\right|\le\Delta^t+2L^{\mathrm{mis}}_{\mathrm{dev}}\,\mathrm{RRD}_{\mathrm{w}}(\pi^t,y^t;\theta^*).$$

Summing this inequality over $t\in[T]$ yields

$$\mathrm{Reg}^{\mathrm{abs}}(T)\le\mathrm{Reg}(T)+2L^{\mathrm{mis}}_{\mathrm{dev}}\,\mathrm{RRD}_{\mathrm{w,cum}}(T).$$

Finally, $\mathrm{RRD}_{\mathrm{w,cum}}(T)\le H\,\mathrm{RRD}_{\mathrm{cum}}(T)$ proves the second inequality in (142). □

T EXPERIMENTS

We examine how the estimated preservation constraint and its relaxation affect leader learning and follower protection. The following results relate the cumulative protection measures to the players' rewards.

T.1 Experimental setup

We use an empirical implementation of PROM (Algorithm 1) in a synthetic linear LFG with $H = 5$, $|\mathcal{S}| = |\mathcal{A}| = |\mathcal{B}| = 8$, $d = 25$, $\rho_0 = (0.2, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$, $\eta = 2$, and $\tau = 1$. The reward functions and transition kernel are held fixed across the tolerance levels $\alpha \in \{0, 0.10, 0.15\}$. Each level uses $T = 3000$ episodes and five random seeds on this common instance.

Before each episode, the leader estimates the follower's reward parameters from observed actions and computes her policy pair by numerical optimistic planning. At $\alpha = 0$, the recommendation equals the estimated QR at every state and step; positive tolerances permit deviations within the TV radius specified in (70).

T.2 Evaluation metrics

Let $\Pi_{\text{grid}} \subseteq \Pi$ denote the class of Markov commitment policies whose statewise action distributions belong to the numerical planner's grid. Define the common grid-restricted QSE benchmark by

$$J^*_{\text{grid}} := \max_{\pi \in \Pi_{\text{grid}}} J^{\text{pol}}(\pi, \mu^{\pi}_{\theta^*}) \,. \tag{153}$$

For episode i, write $\mu^i_h(\cdot \mid s) := \mu^{\pi^i}_{\theta^*_h,h}(\cdot \mid s)$ and $\nu^i_h(\cdot \mid s) := \nu^{\pi^i,y^i}_{\theta^*_h,h}(\cdot \mid s)$. Using the executed state occupancy $d^i_h(s)$ defined in (11), define the signed follower reward loss in episode i by

$$\Delta^r_i := \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d^i_h(s) \sum_{b \in \mathcal{B}} [\mu^i_h(b \mid s) - \nu^i_h(b \mid s)] r^{\pi^i}_{\theta^*_h,h}(s, b).$$

The four cumulative metrics in Figure 2 are evaluated as

$$\begin{aligned}
\mathrm{Reg}^{\text{abs}}_{\text{grid}}(t) &:= \sum_{i=1}^{t} \left| J^*_{\text{grid}} - J(\pi^i, y^i) \right| , \\
\mathrm{CPV}(t) &= \sum_{i=1}^{t} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d^i_h(s) \; \mathrm{TV}\big(\nu^i_h(\cdot \mid s), \mu^i_h(\cdot \mid s)\big) \,, \\
\mathrm{RG}(t) &= \sum_{i=1}^{t} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d^i_h(s) \left| \sum_{b \in \mathcal{B}} \big[\mu^i_h(b \mid s) - \nu^i_h(b \mid s)\big] r^{\pi^i}_{\theta^*_h,h}(s, b) \right| , \\
\mathrm{CG}(t) &= \frac{1}{\eta} \sum_{i=1}^{t} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d^i_h(s) \, \mathrm{KL}\big(\nu^i_h(\cdot \mid s) \big\| \mu^i_h(\cdot \mid s)\big) \,.
\end{aligned}$$

The cumulative average leader return and cumulative average signed follower reward loss in Figure 3 are

$$\overline{J}(t) := \frac{1}{t} \sum_{i=1}^{t} J(\pi^i, y^i), \qquad \overline{L}_r(t) := \frac{1}{t} \sum_{i=1}^{t} \Delta^r_i. \tag{154}$$

Each cumulative quantity or cumulative average is computed within each seed before averaging across the five seeds. Shaded bands show one sample standard deviation, with the lower edge clipped at zero for the nonnegative cumulative metrics.

T.3 Learning and follower protection

Figure 2 reports cumulative absolute leader regret, preservation violation, reward gap, and criterion gap. Under the estimated preservation constraint ($\alpha = 0$), the curves become progressively flatter, most visibly for CG. The relaxed runs accumulate larger deviations, with $\alpha = 0.15$ producing the largest final values for all four quantities. The continued accumulation in the relaxed runs is consistent with the fixed-tolerance terms in Theorem E.19: αT in the CPV and RG bounds and $\alpha^2 T$ in the CG bound. The return comparisons in Figure 3 further show that larger absolute regret can accompany higher leader returns, and that deviations from preservation can coexist with increases in the follower's raw expected reward at the same visited contexts.

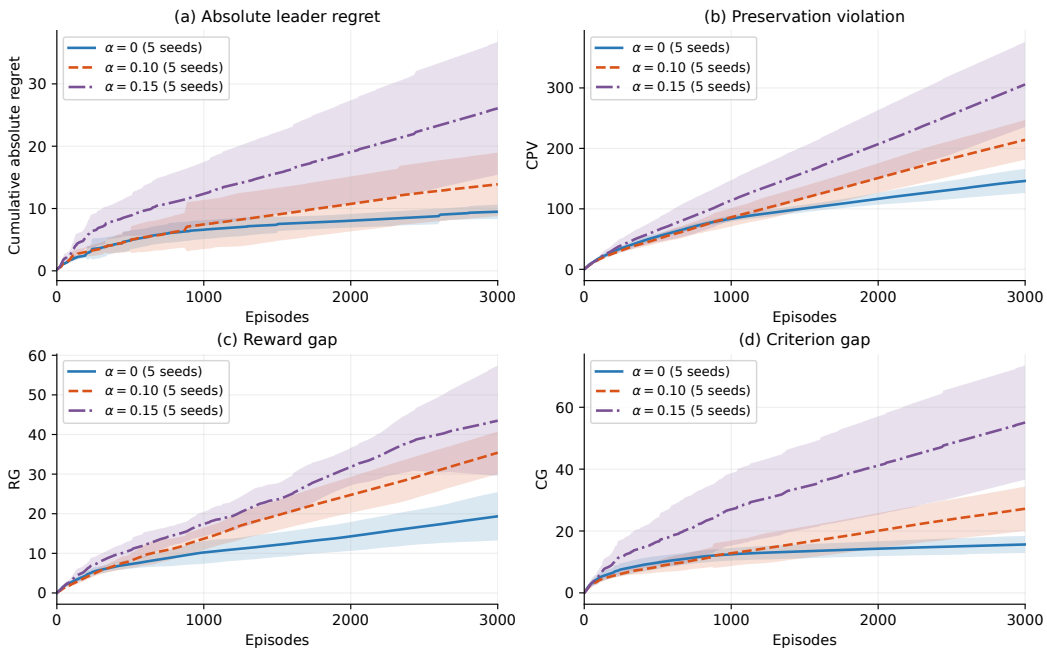

Figure 2: Learning and follower protection under estimated preservation and its fixed-tolerance relaxation. Panels show (a) cumulative absolute leader regret relative to the grid-restricted QSE benchmark, (b) CPV, (c) RG, and (d) CG. The follower-side quantities compare the ER with the true QR at the same state and step under the same commitment. The tolerance is $\alpha \in \{0, 0.10, 0.15\}$, with $\alpha = 0$ enforcing the estimated preservation constraint. Curves show means over five seeds, and shaded bands show one sample standard deviation.

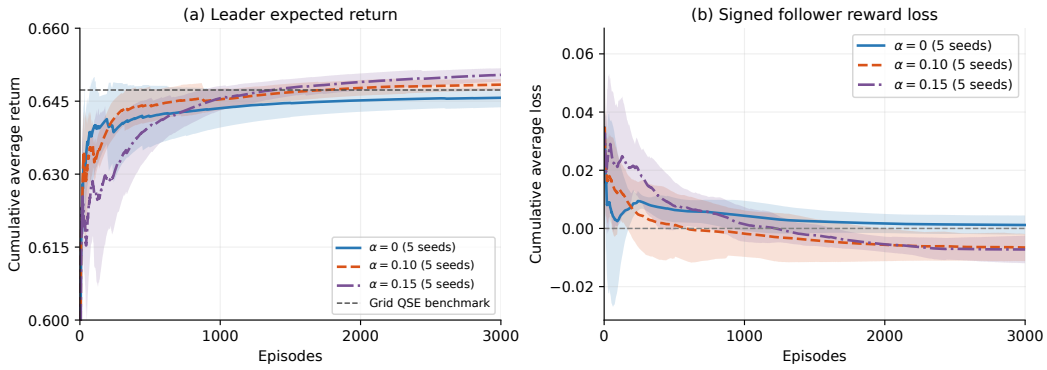

Figure 3: Leader returns and signed follower reward losses under estimated preservation and its relaxation. Panel (a) shows the cumulative average leader expected return; the dashed horizontal line marks the grid-restricted QSE benchmark. Panel (b) shows the cumulative average signed follower reward loss, summed over steps and weighted by the executed state occupancies; negative values indicate higher expected rewards under the ER than under the true QR at the same states, steps, and commitments. Curves show means over five seeds, and shaded bands show one sample standard deviation.

T.4 Leader return and follower reward

Figure 3 complements the cumulative measures in Figure 2 by showing the leader's return and the direction of the follower's reward change.

The leader's cumulative average return approaches the grid-restricted QSE benchmark under $\alpha = 0$ and eventually exceeds it under the two positive tolerances. The final average return increases across the three tolerance levels. These gains can accompany greater absolute regret because the absolute-regret metric accumulates deviations both above and below the common QSE benchmark.

The signed follower reward loss approaches zero under $\alpha = 0$ and becomes negative for the relaxed runs. Thus, in this instance, the greater CPV and CG under relaxation coexist with an average increase in the follower's raw expected reward relative to the QR at the same visited contexts and commitments. RG records the magnitude of the reward differences, while CG measures the loss in the recommendation-free decision criterion through (7). Together, the figures illustrate how behavioral preservation and preservation of the decision criterion provide information beyond the direction of the raw reward change.