PESSIMISTIC FOLLOWER RESPONSES FOR ROBUST STACKELBERG REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Stackelberg and bilevel reinforcement learning optimize a leader while accounting for the adaptation of a follower, but typically assume that a particular optimal or entropy-regularized follower response will be realized. This can be fragile when approximately optimal follower behaviors have substantially different consequences for the leader. We introduce a locally pessimistic response model for two-player Stackelberg Markov games that exponentially tilts the follower's entropy-regularized best response toward actions that are less favorable to the leader, without modifying the follower's learning objective. We derive an off-policy hypergradient that differentiates through both follower adaptation and the pessimistic response transformation. We show that the proposed response is exactly worst-case within an induced local KL neighborhood, remains $O(\alpha^2)$-optimal for the follower, and that optimizing the leader against this pessimistic response sacrifices at most $O(\alpha)$ nominal leader performance.

1 INTRODUCTION

Many sequential decision-making problems are hierarchical: one agent commits to a policy while another adapts in response. Stackelberg and bilevel reinforcement learning model this asymmetry explicitly and have been applied to multi-agent coordination, reward and incentive design and policy alignment (Zhang et al., 2020; Gerstgrasser & Parkes, 2023; Chakraborty et al., 2024; Thoma et al., 2024; Shen et al., 2025). Because the follower's learned policy depends on the leader's decision, optimizing the leader requires accounting for this induced adaptation. Recent hypergradient methods exploit the structure of entropy-regularized reinforcement learning to do so without the convexity assumptions of generic bilevel optimization or direct control over the follower's learning procedure (Yang et al., 2025; Kudo et al., 2026). **We focus here on the underlying response-selection problem in finite Stackelberg Markov games rather than on any particular application domain.**

These methods address **how to differentiate through a nominal follower response, but not which follower response the leader should trust**. Policies that are optimal or nearly optimal for the follower can have sharply different consequences for the leader. This issue is explicit in classical Stackelberg games: strong Stackelberg equilibrium resolves best-response ambiguity in the leader's favor, whereas robust formulations protect the leader against unfavorable approximately optimal responses; even small follower suboptimality can substantially reduce leader utility (Gan et al., 2025). Related sensitivity occurs in sequential reward design under nonunique responses, reward uncertainty, and bounded rationality (Wu et al., 2025). Directly optimizing against the worst approximately optimal response, however, introduces another optimization problem inside an already bilevel sequential problem.

We introduce a locally pessimistic response model for bilevel learning in two-player Stackelberg Markov games. Starting from the follower's entropy-regularized best response, we exponentially shift probability toward actions that are less favorable to the leader while leaving the follower's own learning objective unchanged. We derive the corresponding off-policy leader hypergradient, which differentiates through both the nominal follower adaptation and the pessimistic modification. We further show that the modified response is exactly worst-case within a local KL neighborhood and remains $O(\alpha^2)$-optimal for the follower. Moreover, optimizing the leader against this response incurs at most an $O(\alpha)$ loss relative to the nominally optimal leader. Experiments show that the method

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.

behaves similarly to nominal bilevel learning when response ambiguity is benign, while providing substantially greater robustness when nearby follower responses have different consequences for the leader.

2 Related Work

Stackelberg and bilevel reinforcement learning. Unlike standard MARL formulations in which agents adapt at the same strategic level, Stackelberg RL models commitment by a leader followed by adaptation of a follower (Zhang et al., 2020; Gerstgrasser & Parkes, 2023; He et al., 2025). Bilevel RL makes the resulting dependence explicit and uses hypergradients to account for how the lower-level policy changes with the leader decision. Recent methods address policy alignment and incentive design (Chakraborty et al., 2024; Thoma et al., 2024), remove generic lower-level convexity requirements by exploiting entropy-regularized RL structure (Yang et al., 2025), or estimate this dependence directly from interaction data in decentralized two-player Markov games (Kudo et al., 2026). A common feature is that the leader is optimized against a specified nominal best response; these methods characterize adaptation of that response rather than uncertainty over which approximately optimal response will occur (Gerstgrasser & Parkes, 2023; He et al., 2025).

Entropy regularization and response ambiguity. A natural way to regularize the follower-response map is entropy regularization, which yields a smooth Boltzmann policy and is therefore particularly useful for hypergradient-based bilevel RL (Thoma et al., 2024; Yang et al., 2025; Kudo et al., 2026). However, smoothness does not imply robustness from the leader's perspective: response probabilities are determined by follower values and need not distinguish near-optimal actions with very different leader consequences. Robust Stackelberg equilibrium addresses this issue directly by evaluating the leader against unfavorable approximately optimal follower responses; Gan et al. (2025) show that strategies obtained under optimistic response selection can be highly sensitive even to small follower suboptimality. Wu et al. (2025) similarly study robustness to response nonuniqueness, reward-model error, and bounded rationality in sequential reward design. Crucially, however, their leader **controls the follower reward function**, whereas here the follower objective and learning procedure remain external to the leader. Their specialized robust optimization therefore does not directly provide a differentiable response model that can be inserted into decentralized, model-free bilevel policy learning.

KL-robust and risk-sensitive RL. Exponential tilting also arises in distributionally robust and risk-sensitive reinforcement learning. KL-constrained robust policy evaluation yields adversarial action distributions obtained by exponentially reweighting a nominal policy according to its action values (Smirnova et al., 2019), while KL-regularized robust MDPs are closely connected to entropic risk-sensitive control (Zhang et al., 2024). Our use of exponential tilting is therefore not itself the source of novelty. In our setting, the reference distribution is instead the entropy-regularized best response of a strategic follower, which depends on the leader policy; the leader must differentiate through both this bilevel response and its pessimistic modification. Moreover, the admissible deviation is tied to the follower's own optimality, which lets us quantify how much follower optimality is sacrificed by the local perturbation.

Relation to the present work. Response ambiguity can be represented globally through set-valued follower-response correspondences or by directly minimizing the leader objective over approximately optimal follower policies, as in robust Stackelberg and pessimistic bilevel formulations (Gan et al., 2025; Lignola & Morgan, 2017; 2019). Such approaches capture unfavorable response selection explicitly, but in sequential settings introduce an additional policy-dependent optimization problem whose candidate responses may induce different state occupancies. Our goal is therefore not hypergradient estimation alone, nor a new form of exponential tilting, but a differentiable mechanism for incorporating unfavorable near-optimal follower behavior into trajectory-based bilevel learning without solving an additional global response problem at every leader update. We use a state-action-local KL tilt of the endogenous entropy-regularized follower response and differentiate through both the nominal response and its pessimistic transformation. **Our guarantees characterize the resulting hedge–performance trade-off: the response is locally worst-case within an induced KL neighborhood while remaining approximately optimal for the follower, and optimizing against it preserves nominal leader optimality up to an additive $O(\alpha)$ loss.**

3 PROBLEM FORMULATION

3.1 TWO-PLAYER STACKELBERG MARKOV GAMES

We consider a two-player Stackelberg Markov game with a leader L and a follower F. The game is defined by the tuple

$$\mathcal{G} = (\mathcal{S}, \mathcal{A}_L, \mathcal{A}_F, P, r_L, r_F, \rho_0, \gamma_L, \gamma_F),$$

where $\mathcal{S}$ is a finite state space, $\mathcal{A}_L$ and $\mathcal{A}_F$ are the finite action spaces of the leader and follower, respectively, P is the transition kernel, r_L and r_F are the players' reward functions, ρ_0 is the initial-state distribution, and $\gamma_L, \gamma_F \in [0, 1)$ are their respective discount factors. At each time step t, the leader observes the current state s_t and samples an action $a_t \sim \pi_L(\cdot \mid s_t)$. The follower then observes both the state and the leader action and samples $b_t \sim \pi_F(\cdot \mid s_t, a_t)$. The environment subsequently transitions according to

$$s_{t+1} \sim P(\cdot \mid s_t, a_t, b_t).$$

Thus, $\pi_L : \mathcal{S} \to \Delta(\mathcal{A}_L)$ denotes the leader policy, while $\pi_F : \mathcal{S} \times \mathcal{A}_L \to \Delta(\mathcal{A}_F)$ denotes the follower policy. The follower thus conditions its response on both the state and the leader action.

For a fixed policy pair (π_L, π_F), the discounted return of player $i \in \{L, F\}$ is

$$J_i(\pi_L, \pi_F) = \mathbb{E}_{\rho_0, \pi_L, \pi_F, P}\left[\sum_{t=0}^{\infty} \gamma_i^t r_i(s_t, a_t, b_t)\right].$$

We denote the corresponding state and state-action value functions by $V_i^{\pi_L, \pi_F}$ and $Q_i^{\pi_L, \pi_F}$, respectively.

3.2 ENTROPY-REGULARIZED FOLLOWER RESPONSE

For a fixed leader policy π_L, the follower solves an entropy-regularized reinforcement learning problem. The follower objective is then

$$J_F^{\lambda}(\pi_L, \pi_F) = \mathbb{E}_{\rho_0, \pi_L, \pi_F, P}\left[\sum_{t=0}^{\infty} \gamma_F^t \left(r_F(s_t, a_t, b_t) - \lambda \log \pi_F(b_t \mid s_t, a_t)\right)\right],$$

where $\lambda > 0$ controls the strength of the entropy regularization. The leader objective is not entropy regularized.

Let $Q_F^{\pi_L, \pi_F^*}$ denote the follower's optimal soft action-value function. For finite action spaces, the entropy-regularized response has the Boltzmann form

$$\pi_F^*(b \mid s, a) = \frac{\exp\left(Q_F^{\pi_L, \pi_F^*}(s, a, b)/\lambda\right)}{\sum_{b' \in \mathcal{A}_F} \exp\left(Q_F^{\pi_L, \pi_F^*}(s, a, b')/\lambda\right)} \tag{1}$$

Entropy regularization yields a full-support response that is smooth in the follower action values, facilitating differentiation through π_F^*.

3.3 BILEVEL OBJECTIVE

We parameterize the leader policy by θ_L and write $\pi_L = \pi_L(\theta_L)$. Given the entropy-regularized follower response defined above, the leader solves the bilevel problem

$$\max_{\theta_L} J(\theta_L) := J_L(\pi_L(\theta_L), \pi_F^*(\pi_L(\theta_L))),$$

with

$$\pi_F^*(\pi_L) \in \arg\max_{\pi_F} J_F^{\lambda}(\pi_L, \pi_F).$$

Because π_F^* depends on $\pi_L(\theta_L)$, the leader update must account for both the direct dependence on θ_L and the induced change in the follower response. We write $\nabla_L := \nabla_{\theta_L}$.

4 PROPOSED APPROACH

Our approach builds on the bilevel reinforcement learning formulation of Stackelberg Markov games, in which the leader optimizes its policy while anticipating the response of a follower that independently optimizes its own objective. As in recent hypergradient-based methods, we explicitly account for the dependence of the follower's behavior on the leader policy rather than treating the follower as fixed during the leader update. However, our objective differs from the standard formulation in one important respect. Instead of assuming that the entropy-regularized follower best response π_F^* provides a sufficiently reliable prediction of how the follower will behave, we use it as a reference from which a second, more conservative response π_F^α is constructed. This modified response remains close to the follower's original best response, while shifting probability toward actions that are less favorable to the leader.

$$\pi_F^\alpha(b \mid s, a) = \frac{\pi_F^*(b \mid s, a) \exp\left(-\alpha Q_L^{\pi_L, \pi_F^*}(s, a, b)\right)}{Z_\alpha(s, a)}$$

where $Z_\alpha(s, a) = \sum_{b'} \pi_F^*(b' \mid s, a) \exp(-\alpha Q_L^{\pi_L, \pi_F^*}(s, a, b'))$ and $\alpha \geq 0$ determines how strongly "optimism" is penalized and $Q_L^{\pi_L, \pi_F^*}$ denotes the leader action-value function under the nominal best follower response.

The use of $Q_L^{\pi_L, \pi_F^*}$ in the tilt is deliberate. This quantity is not myopic in the temporal sense: it already captures the long-term consequences of the current follower action under the nominal continuation policy. The construction is instead local in follower-policy space, because future continuation behavior is evaluated under π_F^* rather than recursively recomputing the leader value under the modified response. A self-consistent alternative based on $Q_L^{\pi_L, \pi_F^\alpha}$ would require solving an additional fixed-point or lower-level optimization problem and would lose the closed-form response used here. Our analysis therefore treats the nominal response as the reference continuation model; Theorem 4.3 and Corollary 4.4 quantify the resulting value discrepancy, while increasingly large α may make this local approximation less informative.

For $\alpha = 0$, the response reduces to the nominal entropy-regularized best response. Increasing α downweights follower actions with high leader value and therefore produces a locally less favorable response. Importantly, the follower is not retrained under this criterion: π_F^α is used only to evaluate and update the leader.

4.1 HYPERGRADIENT UNDER THE MODIFIED FOLLOWER RESPONSE

We consider the leader update after the follower has converged to its entropy-regularized best response in section 3.2. As in standard bilevel reinforcement learning, the leader cannot regard this response as fixed: changing the leader policy modifies the trajectories experienced by the follower and, consequently, the follower policy obtained at the lower level. The appropriate update is therefore a hypergradient rather than an ordinary policy gradient $\nabla_L J_L(\pi_L, \pi_F) = \frac{d\pi_F}{d\pi_L} \frac{\partial J_L}{\partial \pi_F} + \frac{\partial J_L}{\partial \pi_L}$.

Following the policy-gradient interpretation used for two-player Markov games, the leader policy and the follower response can be viewed jointly when differentiating the leader objective (Kudo et al., 2026; Thoma et al., 2024). This separates the update into a direct contribution, corresponding to the change in the leader's own action probabilities and an indirect contribution arising from the dependence of the follower response on the leader policy. Kudo et al. (2026) use precisely this distinction for the nominal entropy-regularized best response; our setting retains it, but evaluates the leader against the modified follower response introduced above.

For brevity's sake we introduce the following joint policies

$$w(b, a, s) = \frac{\pi_F^\alpha(b \mid a, s)}{\pi_F^*(b \mid a, s)}, \; v(b, a \mid s) = \pi_L(a \mid s)\pi_F^*(b \mid a, s), \; \beta(b, a \mid s) = \pi_L(a \mid s)\pi_F^\alpha(b \mid a, s)$$

We also adopt the Benefit nomenclature from Kudo et al. (2026) to designate the quantity that measures whether a particular follower action is more or less favorable to the leader than the average follower action at the same state/leader-action pair.

$$B_L^{\pi_L, \pi_F}(s, a, b) = \mathbf{Q}_\mathbf{L}^{\pi_\mathbf{L}, \pi_\mathbf{F}}(s, a, b) - \mathbb{E}_{b \sim \pi_F(\cdot \mid s, a)}\left[\mathbf{Q}_\mathbf{L}^{\pi_\mathbf{L}, \pi_\mathbf{F}}(s, a, b)\right]$$

We use the same centering principle. This is especially natural in our formulation because the modification of the follower policy is itself determined by relative differences in the leader's action value. Consequently, follower actions that are indistinguishable from the leader's perspective make no contribution through this term, whereas actions whose consequences differ substantially for the leader have a larger effect on the update.

Theorem 4.1 (Off-Policy Hypergradient for Perturbed MGs)**.** *Let d^v and d^β denote the normalized γ_L-discounted state occupancies induced by the nominal and modified joint policies, respectively, and define $\zeta(s) = \frac{d^\beta(s)}{d^v(s)}$.*

The hypergradient of the leader's discounted sum of rewards with respect to the leader's policy parameters is given by

$$\begin{aligned}\nabla_L J(\theta_L, \pi_F^\alpha(\theta_L)) = &\frac{1}{1-\gamma_L}\mathbb{E}_{\substack{s\sim d_v \\ a,b\sim v}}\Big[\zeta(s)w(b,a,s)\Big(\nabla_L \log \pi_L(a \mid s)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b) \\ &+ B_L^{\pi_L,\pi_F^\alpha}(s,a,b)\big(\frac{1}{\lambda}\nabla_L \mathbf{Q_F}^{\pi_\mathbf{L},\pi_\mathbf{F}^*}(s,a,b) - \alpha\nabla_L \mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^*}(s,a,b)\big)\Big)\Big]\end{aligned} \tag{2}$$

Proof is detailed in Appendix A.4. The resulting hypergradient has three conceptually distinct components. The first is the usual policy-gradient contribution and captures the improvement in the leader's return obtained by changing its own policy while holding the response mechanism locally fixed. The second accounts for the change in the follower's entropy-regularized best response induced by the leader update. This is the same anticipatory effect that motivates hypergradient methods for bilevel reinforcement learning: an update that omits this term optimizes against the current follower rather than against the follower that will result from the updated leader policy.

The third component is specific to our construction. Since the leader is evaluated against the modified response rather than directly against the nominal best response, changing the leader policy also changes the transformation from the nominal follower policy to the modified one. In particular, the leader affects both the follower values determining the original entropy-regularized response and the leader values used to construct the pessimistic modification. The hypergradient must therefore differentiate through both stages. This is the main structural difference from the hypergradient derived by Kudo et al. (2026), where the indirect term only describes how the nominal follower best response changes with the leader policy. In our case, the same bilevel dependence is preserved, but an additional response-selection dependence is introduced by the optimism penalty.

An important consequence is that our method does not simply train the follower on an adversarial objective. The lower-level learner continues to optimize its own entropy-regularized return and produces the nominal best response. The modified response is constructed only for the leader optimization step. This distinction is important to the interpretation of the method: the leader does not assume control over the follower's learning objective, but instead asks how its own policy would perform under a nearby response that is less favorable to it. The modified response therefore acts as a robustness mechanism around the behavior generated by the original lower-level problem rather than as a replacement for that problem.

4.2 Off-Policy Estimation

This separation between the nominal and modified responses also creates an estimation problem. Interaction data are naturally collected under the leader policy together with the follower's nominal best response $v(b, a \mid s)$, whereas the leader objective of interest is evaluated under the modified follower $\beta(b, a \mid s)$. The two joint policies therefore induce different follower-action distributions and, through their effect on the transition dynamics, potentially different state-visitation distributions. The term $\zeta(s_t)w(b_t, a_t, s_t)$ in Theorem 4.1 explicitly accounts for both changes through importance weighting. In our tabular experiments, w is computed directly from the two follower policies, while the discounted state-occupancy ratio ζ is computed exactly from the induced transition matrices. **Appendix B provides the full off-policy estimator and implementation details**.

4.3 Implementation

Training alternates between computing the follower's entropy-regularized response, collecting nominal-response trajectories, updating the leader critics, constructing π_F^α, and applying the hy-

pergradient in Theorem 4.1. In the finite tabular experiments, the two value sensitivities $\nabla_L Q_F^*$ and $\nabla_L Q_L^v$ are obtained by differentiating their Bellman fixed points using the known transition model, while the modified leader value is learned off-policy from nominal-response data. Full details of the critics, importance corrections, exact value-sensitivity solves, and actor surrogate are given in Appendix B.

4.4 Robustness

Introducing a pessimistic modification is useful only if the resulting policy remains a plausible response for the follower and if optimizing against it does not impose an uncontrolled cost on the leader. We first show that the exponential tilt is exactly a locally worst-case response under a KL constraint. We then establish that the modification remains approximately optimal for the follower while inducing at most a linear change in leader value. Finally, we show that optimizing the leader against this locally pessimistic response preserves nominal leader optimality up to the same $O(\alpha)$ loss.

Proposition 4.2 (Exact local KL-robustness). *Fix a leader policy π_L, a state–action pair (s, a), and for $\alpha > 0$ let*

$$p(\cdot) = \pi_F^*(\cdot \mid s, a) \quad \text{and} \quad q_\alpha(\cdot) = \pi_F^\alpha(\cdot \mid s, a)$$

Then

$$\boxed{q_\alpha = \arg \min_{q \in \Delta(\mathcal{A}_F)} \left\{ \mathbb{E}_{b \sim q} \left[Q_L^{\pi_L, p}(s, a, b) \right] + \frac{1}{\alpha} D_{\mathrm{KL}}(q \| p) \right\}}$$

Moreover, letting

$$\kappa_\alpha(s, a) := D_{\mathrm{KL}}(q_\alpha \| p),$$

the same policy solves

$$\boxed{q_\alpha \in \arg \min_{\substack{q \in \Delta(\mathcal{A}_F) \\ D_{\mathrm{KL}}(q \| p) \le \kappa_\alpha(s,a)}} \mathbb{E}_{b \sim q} \left[Q_L^{\pi_L, p}(s, a, b) \right]}$$

Hence, at every (s, a), π_F^α is the worst response for the leader within the KL neighborhood of the nominal entropy-regularized best response having radius $\kappa_\alpha(s, a)$.

Proof is detailed in Appendix A.5. The result follows from the optimality conditions of KL-regularized linear minimization; the constrained form follows by taking the attained KL divergence as the ambiguity radius.

Theorem 4.3 (Controlled response perturbation). *Assume finite state and action spaces and $\gamma_F, \gamma_L < 1$. Fix a leader policy π_L, and suppose that*

$$\mathrm{span}_b \, Q_L^{\pi_L, \pi_F^*}(s, a, b) := \max_b Q_L^{\pi_L, \pi_F^*}(s, a, b) - \min_b Q_L^{\pi_L, \pi_F^*}(s, a, b) \le C_L$$

for every (s, a). Then the modified response satisfies:

$$\boxed{0 \le J_F^\lambda(\pi_L, \pi_F^*(\pi_L)) - J_F^\lambda(\pi_L, \pi_F^\alpha(\pi_L)) \le \frac{\lambda \alpha^2 C_L^2}{8(1 - \gamma_F)}} \tag{3}$$

In particular, π_F^α is an ϵ_α-best response to π_L, where $\epsilon_\alpha = \frac{\lambda \alpha^2 C_L^2}{8(1-\gamma_F)}$. Moreover,

$$\boxed{0 \le J_L(\pi_L, \pi_F^*) - J_L(\pi_L, \pi_F^\alpha) \le \frac{\alpha C_L^2}{4(1 - \gamma_L)}} \tag{4}$$

Thus the follower's own objective deteriorates only quadratically in α, whereas the induced leader-value change is at most linear in α. *Proof is detailed in Appendix A.6.*

For brevity, let

$$p_{\pi_L} := \pi_F^*(\pi_L), \qquad q^\alpha_{\pi_L} := \pi_F^\alpha(\pi_L).$$

The previous result bounds the loss induced by the pessimistic response for any fixed leader policy. We next show that this control is preserved when the leader itself is optimized against the modified response.

Corollary 4.4 (Controlled nominal optimality loss)**.** *Assume the conditions of Theorem 4.3 hold uniformly over $\pi_L \in \Pi_L$, i.e.,*

$$\sup_{\pi_L \in \Pi_L} \sup_{s,a} \operatorname{span}_b Q_L^{\pi_L, p_{\pi_L}}(s,a,b) \le C_L.$$

Let

$$\pi_L^0 \in \arg\max_{\pi_L \in \Pi_L} J_L(\pi_L, p_{\pi_L})$$

denote a nominally optimal leader policy, and let

$$\pi_L^\alpha \in \arg\max_{\pi_L \in \Pi_L} J_L(\pi_L, q^\alpha_{\pi_L})$$

denote a leader policy optimized against the pessimistic response. Then, with

$$\rho_\alpha = \frac{\alpha C_L^2}{4(1-\gamma_L)},$$

we have

$$J_L(\pi_L^0, p_{\pi_L^0}) - \rho_\alpha \le J_L(\pi_L^\alpha, q^\alpha_{\pi_L^\alpha}) \le J_L(\pi_L^\alpha, p_{\pi_L^\alpha}) \le J_L(\pi_L^0, p_{\pi_L^0}).$$

In particular,

$$0 \le J_L(\pi_L^0, p_{\pi_L^0}) - J_L(\pi_L^\alpha, p_{\pi_L^\alpha}) \le \rho_\alpha.$$

Thus optimizing against the locally pessimistic follower response sacrifices at most $O(\alpha)$ nominal leader performance.

A proof is given in Appendix A.7.

Scope of the guarantees. Our analysis is stated for finite state and action spaces with an entropy-regularized follower response. In this setting, $\lambda > 0$ gives the Boltzmann response full support, which ensures that the action-level likelihood ratio used above is well defined. The span assumption is automatically finite under bounded leader rewards (for $r_L \in [r_{\min}, r_{\max}]$, a uniform bound of order $(r_{\max} - r_{\min})/(1-\gamma_L)$ follows) although such worst-case bounds may be loose. Extending the construction to continuous action spaces would additionally require a tractable representation and normalization of the tilted follower distribution, as well as occupancy-ratio estimation when the dynamics are unknown.

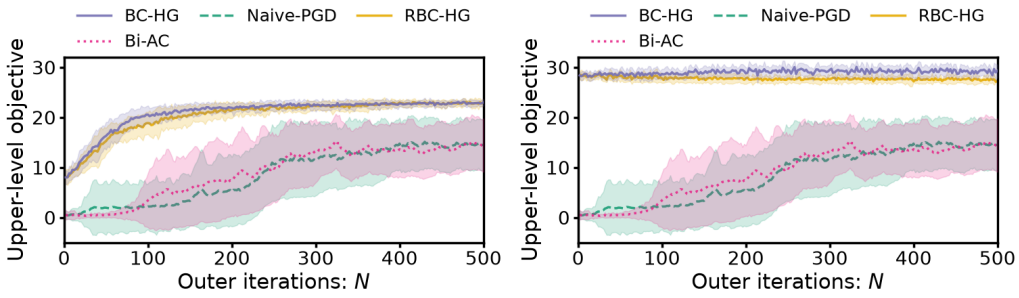
5 Experiments

5.1 Experimental Questions and Setup

Our experiments examine two aspects of the proposed robustness correction. First, we test whether introducing the pessimistic response modification degrades performance in settings where the nominal follower response is already sufficiently reliable. Second, we study environments in which small deviations from the nominal follower response can substantially reduce the leader's return, and evaluate whether the proposed method provides greater robustness in these settings.

We compare Robust BC-HG (RBC-HG) with BC-HG (Kudo et al., 2026), which optimizes the leader against the nominal entropy-regularized follower response. Where applicable, we also report Naive-PGD and Bi-AC following the experimental protocol of Kudo et al. (2026). All reported curves are averaged over 10 independent random seeds, with shaded regions indicating variation across seeds.

For robustness evaluation, the learned leader policies are held fixed while the follower response is progressively perturbed away from its nominal entropy-regularized best response. We consider three forms of perturbation: temperature scaling, an ϵ-random follower perturbation, and a near-best adversarial perturbation. These tests measure how rapidly leader performance deteriorates as the assumed follower behavior becomes less accurate.



(a) Limited response ambiguity.

(b) Response-sensitive benchmark.

Figure 1: Training performance on the two tabular Stackelberg benchmarks. Curves show the discounted leader objective over 10 random seeds; shading indicates variation across seeds. RBC-HG uses $\alpha = 0.5$ in this comparison.

5.2 Training Behavior Across Benchmarks

We evaluate on two three-state Stackelberg Markov games. The first is the benchmark of Kudo et al. (2026), where high-performing leader policies are comparatively insensitive to nearby follower responses. The second modifies the transition and reward structure so that follower actions with similar follower value can have substantially different consequences for the leader. The complete game specifications are given in Appendix C.1. All shared optimization settings follow the BC-HG setup.

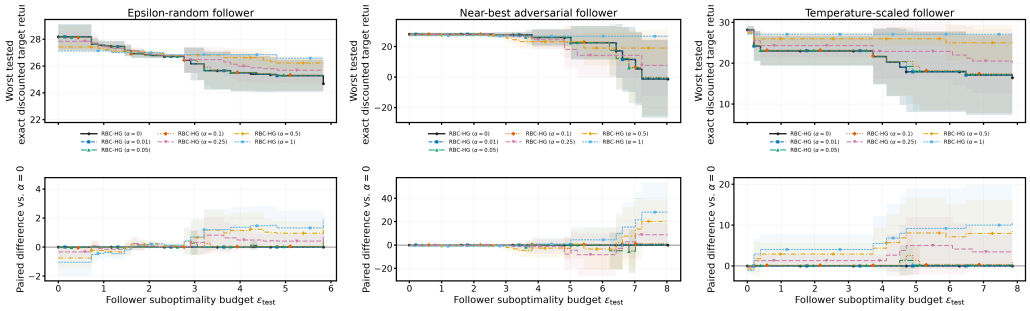
Figure 1 shows the discounted leader objective during training. On the limited-ambiguity benchmark, RBC-HG and BC-HG converge to similar returns, indicating that the pessimistic correction does not materially reduce nominal performance when response ambiguity is benign. On the response-sensitive benchmark, RBC-HG converges to a somewhat lower nominal return than BC-HG. The robustness experiments below test whether this nominal cost is accompanied by reduced sensitivity to follower response deviations. Both hypergradient methods substantially outperform Naive-PGD and Bi-AC on the two tasks.

5.3 Comparison with the Exact δ-RSE

Because the response-sensitive benchmark is sufficiently small, we can additionally compare the learned behavior with its exact stochastic robust Stackelberg solution (Gan et al., 2025). Under the nominal follower best response, the leader prefers the risky cycle $S \to B \to S$, which yields a discounted return of 38.73. However, the follower has a near-optimal response at B that reduces its reward only from 1 to 0.9 while changing the leader reward from 1 to -1. Accounting for δ-optimal follower responses therefore rapidly decreases the robust value of this route. For stochastic follower responses, the robust values of the two routes cross at $\delta^\star \simeq 0.585$, after which the exact δ-RSE selects the safer cycle $S \to A \to S$, whose nominal leader return is 27.11. The complete derivation is given in Appendix C.2. This provides a ground-truth reference for evaluating whether our pessimistic response model induces the qualitative behavior of the true robust equilibrium.

5.4 Robustness and the Effect of α

We isolate the contribution of the pessimistic response by sweeping $\alpha \in \{0, 0.01, 0.05, 0.1, 0.25, 0.5, 1\}$ on the response-sensitive benchmark, using 10 matched seeds and holding all other training settings fixed. After training, the leader is frozen and the nominal follower is recomputed to convergence. We evaluate epsilon-random, temperature-scaled, and near-best adversarial follower responses. Since



(a) Epsilon-random follower (b) Near-best adversarial follower (c) Temperature-scaled follower

Figure 2: Effect of the pessimism strength α on the response-sensitive benchmark. Each panel reports the worst exact discounted leader return among the tested follower responses whose follower suboptimality does not exceed ϵ_{test} (see Appendix C.3). Stronger pessimistic training yields substantially greater robustness for moderate and large response deviations, especially under temperature-scaled and near-best adversarial responses. Curves show representative values of α. Shading denotes variation over 10 matched random seeds.

their raw perturbation parameters are not directly comparable, the horizontal axis reports the induced follower suboptimality

$$\epsilon_{\text{test}}(q) = J_F^{\lambda}(\pi_L, \pi_F^*) - J_F^{\lambda}(\pi_L, q),$$

and each curve gives the worst leader return among the tested responses within that budget. Full evaluation details are given in Appendix C.3.

Figure 2 shows a clear robustness–performance trade-off. Small values $\alpha \leq 0.1$ behave similarly to the matched $\alpha = 0$ control, while robustness begins to improve at $\alpha = 0.25$ and becomes pronounced for $\alpha = 0.5$ and 1. The separation is strongest for temperature-scaled and near-best adversarial responses: at large ϵ_{test}, the nominal-style policies deteriorate sharply whereas $\alpha = 1$ retains a high leader return. Epsilon-random perturbations are less damaging overall but exhibit the same trend. This robustness comes with a modest reduction in nominal performance (Appendix C.3). The response induced during training is not monotone in α after re-optimizing the leader: its KL divergence and follower optimality gap peak around $\alpha = 0.5$ and decrease at $\alpha = 1$. Thus the strongest robust solution is not simply the one trained against the largest realized follower displacement.

6 Conclusion

We introduced a locally pessimistic follower response for bilevel learning in Stackelberg Markov games. The response is state-action-wise worst-case within an induced KL neighborhood, remains approximately optimal for the follower, and optimizing against it sacrifices at most $O(\alpha)$ nominal leader performance under the stated span condition. Empirically, the method behaves similarly to nominal bilevel learning when response ambiguity is limited, while improving robustness when near-optimal follower responses have substantially different consequences for the leader. On the response-sensitive benchmark, the exact δ-RSE provides an analytic reference showing the same qualitative shift from the nominally superior risky route to the safer route as response ambiguity increases.

Limitations. Our analysis and experiments are restricted to finite-action games, and the tabular experiments use the known transition model to compute value sensitivities and occupancy corrections exactly. Under unknown dynamics these quantities would need to be estimated. Moreover, our construction is local around the nominal continuation value and does not solve a self-consistent global worst-response problem.

AI USE STATEMENT

In this work, we used generative AI tools to assist in providing feedback on the experimental design. Generative AI tools were also used to assist with software implementation and debugging, the presentation of experimental results, and the refinement of experimental settings.

Additionally, we used generative AI tools to draft and edit parts of the manuscript, improve readability and structure, and suggest titles, section organization, and figure or caption improvements.

All AI-assisted theoretical content was independently reviewed by the authors. AI-assisted code was reviewed, executed, and tested by the authors, including comparisons with exact computations in the tabular settings where applicable. All reported experimental results were obtained and evaluated by the authors. Generative AI was not used to generate experimental data or to replace the execution of experiments.

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

REPRODUCIBILITY STATEMENT.

We have taken several steps to facilitate reproduction of both the theoretical and empirical results presented in this work. The assumptions underlying our theoretical results are stated explicitly in the main text, and complete proofs of the propositions and supporting lemmas are provided in Appendix A. Detailed descriptions of the benchmark environments, training procedures, and additional experimental results are provided in the experimental appendix. We additionally provide the complete implementation of the proposed method, benchmark environments, and configuration files including the hyperparameters, used to generate the reported results as supplementary material. The accompanying code includes instructions for reproducing the experiments and figures presented in the paper.

REFERENCES

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. PARL: A Unified Framework for Policy Alignment in Reinforcement Learning from Human Feedback. In *International Conference on Learning Representations*, 2024. URL https://mlanthology.org/iclr/2024/chakraborty2024iclr-parl/.

Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. Robust Stackelberg Equilibria. *Mathematical Programming*, November 2025. ISSN 1436-4646. doi: 10.1007/s10107-025-02291-4. URL https://doi.org/10.1007/s10107-025-02291-4.

Matthias Gerstgrasser and David C. Parkes. Oracles & followers: Stackelberg equilibria in deep multi-agent reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11213–11236. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gerstgrasser23a.html.

Jun He, Andrew Liu, and Yihsu Chen. Learning in stackelberg markov games. In *NeurIPS 2025 Workshop on Tackling Climate Change with Machine Learning*, 2025. URL https://www.climatechange.ai/papers/neurips2025/30.

Mikoto Kudo, Takumi Tanabe, Akifumi Wachi, and Youhei Akimoto. Sample-efficient hypergradient estimation for decentralized bi-level reinforcement learning. In *Proceedings of the International Conference on Automated Planning and Scheduling, 36(1), 630–638.*, 2026. URL https://doi.org/10.1609/icaps.v36i1.42881.

M. Beatrice Lignola and Jacqueline Morgan. Inner Regularizations and Viscosity Solutions for Pessimistic Bilevel Optimization Problems. *Journal of Optimization Theory and Applications*, 173(1):183–202, April 2017. ISSN 1573-2878. doi: 10.1007/s10957-017-1085-4. URL https://doi.org/10.1007/s10957-017-1085-4.

M. Beatrice Lignola and Jacqueline Morgan. Further on Inner Regularizations in Bilevel Optimization. *Journal of Optimization Theory and Applications*, 180(3):1087–1097, March 2019. ISSN 1573-2878. doi: 10.1007/s10957-018-1438-7. URL https://doi.org/10.1007/s10957-018-1438-7.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. *Journal of Machine Learning Research*, 26(114):1–49, 2025. URL http://jmlr.org/papers/v26/24-0720.html.

Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning, 2019. URL https://openreview.net/forum?id=r1xfz_93oN.

Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/e66309ead63bc1410d2df261a28f602d-Abstract-Conference.html.

Shuo Wu, Haoxiang Ma, Jie Fu, and Shuo Han. Robust Reward Design for Markov Decision Processes. *Journal of Artificial Intelligence Research*, 84, 2025. doi: 10.1613/JAIR.1.19154. URL https://mlanthology.org/jair/2025/wu2025jair-robust/.

Yan Yang, Bin Gao, and Ya-xiang Yuan. Bilevel reinforcement learning via the development of hyper-gradient without lower-level convexity. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4780–4788. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/yang25g.html.

Haifeng Zhang, Weizhe Chen, Zeren Huang, Minne Li, Yaodong Yang, Weinan Zhang, and Jun Wang. Bi-level actor-critic for multi-agent coordination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7325–7332, 04 2020. doi: 10.1609/aaai.v34i05.6226.

Runyu Zhang, Yang Hu, and Na Li. Soft robust mdps and risk-sensitive mdps: Equivalence, policy gradient, and sample complexity. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 14655–14669, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3f7f788b479a26348605a9f97c465d22-Paper-Conference.pdf.

A Proofs and Technical Derivations

A.1 Formal definitions

We denote, where $i \in \{L, F\}$, the state-value function by

$$V_i^{\pi_L,\pi_F}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma_i^t r_i(s_t, a_t, b_t) \,\middle|\, s_0 = s\right]$$

and by

$$Q_i^{\pi_L,\pi_F}(s, a, b) = r_i(s, a, b) + \gamma_i \mathbb{E}_{s' \sim P(\cdot|s,a,b)}\left[V_i^{\pi_L,\pi_F}(s')\right]$$

the state-action value function defined by the standard Bellman recursion.

A.2 Entropy-Regularized Follower Response

For a fixed leader policy π_L, the follower solves an entropy-regularized reinforcement learning problem. The follower objective is then

$$J_F^{\lambda}(\pi_L, \pi_F) = \mathbb{E}_{\rho_0,\pi_L,\pi_F,P}\left[\sum_{t=0}^{\infty} \gamma_F^t \left(r_F(s_t, a_t, b_t) - \lambda \log \pi_F(b_t \mid s_t, a_t)\right)\right],$$

where $\lambda > 0$ controls the strength of the entropy regularization. The leader objective is not entropy regularized.

Since the follower selects its action after observing the leader action, we define its soft value at (s, a) by

$$V_F^{\pi_L,\pi_F}(s, a) = \mathbb{E}_{b \sim \pi_F(\cdot|s,a)}\left[Q_F^{\pi_L,\pi_F}(s, a, b) - \lambda \log \pi_F(b \mid s, a)\right],$$

with corresponding soft action-value function

$$Q_F^{\pi_L,\pi_F}(s, a, b) = r_F(s, a, b) + \gamma_F \mathbb{E}_{\substack{s' \sim P(\cdot|s,a,b) \\ a' \sim \pi_L(\cdot|s')}}\left[V_F^{\pi_L,\pi_F}(s', a')\right].$$

For a given leader policy, we denote the follower's entropy-regularized best response by

$$\pi_F^*(\pi_L) \in \arg\max_{\pi_F} J_F^{\lambda}(\pi_L, \pi_F).$$

Under the finite-action entropy-regularized setting, the optimal response admits the Boltzmann form

$$\pi_F^*(b \mid s, a) = \frac{\exp\left(Q_F^{\pi_L,\pi_F^*}(s, a, b)/\lambda\right)}{\sum_{b' \in \mathcal{A}_F} \exp\left(Q_F^{\pi_L,\pi_F^*}(s, a, b')/\lambda\right)}.$$

Entropy regularization therefore provides a stochastic response with full support and a smooth dependence on the follower action values. This property is particularly useful in bilevel reinforcement learning, where the leader update must account for the dependence of π_F^* on π_L (Thoma et al., 2024; Yang et al., 2025; Kudo et al., 2026). Importantly, π_F^* remains the response obtained solely from the follower's own objective. The pessimistic modification introduced in Section 4 is applied only after this nominal response has been obtained.

A.3 Derivative Identities for the Follower Responses

Lemma A.1 (Derivative of the entropy-regularized response)**.** *For the entropy-regularized follower best response,*

$$\pi_F^*(b \mid s,a) = \frac{\exp\Big(Q_F^{\pi_L,\pi_F^*}(s,a,b)/\lambda\Big)}{\sum_{b'}\exp\Big(Q_F^{\pi_L,\pi_F^*}(s,a,b')/\lambda\Big)},$$

we have

$$\begin{aligned}
\nabla_L \pi_F^*(b \mid s,a) &= \nabla_L \frac{\exp(Q_F^*(s,a,b)/\lambda)}{\sum_{b'}\exp(Q_F^*(s,a,b')/\lambda)} \\
&= \frac{\frac{\nabla_L Q_F^*(s,a,b)}{\lambda}\exp(Q_F^*(s,a,b)/\lambda)\sum_{b'}\exp(Q_F^*(s,a,b')/\lambda)}{} \\
&\quad - \frac{\exp(Q_F^*(s,a,b)/\lambda)\sum_{b'}\frac{\nabla_L Q_F^*(s,a,b')}{\lambda}\exp(Q_F^*(s,a,b')/\lambda)}{\big(\sum_{b'}\exp(Q_F^*(s,a,b'))\big)^2} \\
&= \pi_F^*(s,a,b)\frac{\nabla_L Q_F^*(s,a,b)}{\lambda} - \pi_F^*(s,a,b)\sum_{b'}\frac{\pi_F^*(s,a,b')\nabla_L Q_F^*(s,a,b')}{\lambda} \\
&= \frac{1}{\lambda}\Big(\nabla_L Q_F^*(s,a,b) - \mathbb{E}_{b'\sim\pi_F^*(\cdot|s,a)}[\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b')]\Big)\,\pi_F^*(b \mid s,a)
\end{aligned}$$

and so

$$\nabla_L \log \pi_F^*(b \mid s,a) = \frac{1}{\lambda}\Big(\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b) - \mathbb{E}_{b'\sim\pi_F^*(\cdot|s,a)}[\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b')]\Big).$$

Lemma A.2 (Derivative of the modified response)**.** *The logarithmic derivative of the modified follower response satisfies*

$$\begin{aligned}
\nabla_L \log \pi_F^\alpha(b \mid s,a) = \frac{1}{\lambda}&\Big(\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b) - \mathbb{E}_{b'\sim\pi_F^\alpha}[\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b')]\Big) \\
&- \alpha\Big(\nabla_L Q_L^{\pi_L,\pi_F^*}(s,a,b) - \mathbb{E}_{b'\sim\pi_F^\alpha}[\nabla_L Q_L^{\pi_L,\pi_F^*}(s,a,b')]\Big).
\end{aligned} \tag{5}$$

Proof. Recall that

$$\pi_F^\alpha(b \mid s,a) = \frac{\pi_F^*(b \mid s,a)\exp\Big(-\alpha Q_L^{\pi_L,\pi_F^*}(s,a,b)\Big)}{Z_\alpha(s,a)}$$

Since

$$\pi_F^*(b \mid s,a) = \frac{\exp\Big(Q_F^{\pi_L,\pi_F^*}(s,a,b)/\lambda\Big)}{\sum_{\bar b}\exp\Big(Q_F^{\pi_L,\pi_F^*}(s,a,\bar b)/\lambda\Big)},$$

the normalization of π_F^* is independent of b and can be absorbed into the normalization constant of π_F^α. Hence

$$\pi_F^\alpha(b \mid s,a) = \frac{\exp\Big(\frac{1}{\lambda}Q_F^{\pi_L,\pi_F^*}(s,a,b) - \alpha Q_L^{\pi_L,\pi_F^*}(s,a,b)\Big)}{\sum_{\bar b}\exp\Big(\frac{1}{\lambda}Q_F^{\pi_L,\pi_F^*}(s,a,\bar b) - \alpha Q_L^{\pi_L,\pi_F^*}(s,a,\bar b)\Big)}$$

Define

$$h_b := \frac{1}{\lambda}Q_F^{\pi_L,\pi_F^*}(s,a,b) - \alpha Q_L^{\pi_L,\pi_F^*}(s,a,b)$$

Then

$$\log \pi_F^\alpha(b \mid s,a) = h_b - \log\sum_{\bar b} e^{h_{\bar b}}$$

Differentiating with respect to the leader parameters gives

$$\nabla_L \log \pi_F^\alpha(b \mid s,a) = \nabla_L h_b - \sum_{\bar b}\pi_F^\alpha(\bar b \mid s,a)\nabla_L h_{\bar b}$$

Using

$$\nabla_L h_b = \frac{1}{\lambda}\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b) - \alpha\nabla_L Q_L^{\pi_L,\pi_F^*}(s,a,b),$$

we obtain

$$\begin{aligned}\nabla_L \log \pi_F^\alpha(b \mid s,a) = \frac{1}{\lambda}\Big(\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b) - \mathbb{E}_{\bar{b}\sim\pi_F^\alpha(\cdot|s,a)}\Big[\nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,\bar{b})\Big]\Big)\\ - \alpha\Big(\nabla_L Q_L^{\pi_L,\pi_F^*}(s,a,b) - \mathbb{E}_{\bar{b}\sim\pi_F^\alpha(\cdot|s,a)}\Big[\nabla_L Q_L^{\pi_L,\pi_F^*}(s,a,\bar{b})\Big]\Big)\end{aligned}$$

This proves the result. ■

Lemma A.3 (Boltzmann covariance trick)**.** *Given three policies π_L, π_F, π_F', any state $s \in S$, and the following definition of the Benefit (Kudo et al., 2026),*

$$B_L^{\pi_L,\pi_F}(s,a,b) = \mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}}(s,a,b) - \mathbb{E}_{b\sim\pi_F(\cdot|s,a)}\left[\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}}(s,a,b)\right]$$

The Boltzmann covariance trick (Kudo et al., 2026), in this paper, refers to the following relationship:

$$\begin{aligned}\mathbb{E}_{b\sim\pi_F'}\left[B_L^{\pi_L,\pi_F'}(s,a,b)\nabla_L\mathbf{Q}^{\pi_\mathbf{L},\pi_\mathbf{F}}(s,a,b)\right] &= \mathbb{C}ov_b\left[\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}'}(s,a,b)\nabla_L\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}}(s,a,b)\right]\\ &= \mathbb{E}_{b\sim\pi_F'}\left[\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}'}(s,a,b)\big(\nabla_L\mathbf{Q}^{\pi_\mathbf{L},\pi_\mathbf{F}}(s,a,b) - \mathbb{E}_{b'\sim\pi_F'}\left[\nabla_L\mathbf{Q}^{\pi_\mathbf{L},\pi_\mathbf{F}}(s,a,b)\right]\big)\right]\end{aligned}$$

A.4 Proof of Theorem 4.1

We can express the leader's value function under the follower's modified response as follows

$$\begin{aligned}\nabla_L\mathbf{V}_L^{\pi_L,\pi_F^\alpha}(s) &= \nabla_L\left[\sum\pi_L(a\mid s)\pi_F^\alpha(b\mid a,s)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b)\right]\\ &= \sum_{s'}\sum_{k=0}\gamma_L^k\mathcal{P}r\left(s\to s'\mid k,\pi_L,\pi_F^\alpha\right)\sum_a\sum_b\big(\nabla_L\pi_L(a\mid s')\pi_F^\alpha(b\mid a,s')\\ &\qquad + \pi_L(a\mid s')\nabla_L\pi_F^\alpha(b\mid a,s')\big)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s',a,b)\end{aligned}\tag{6}$$

given that $J(\theta_L,\pi_F^\alpha(\theta_L)) = \mathbf{V}_L^{\pi_L,\pi_F^\alpha}(s)$, for a fixed s_0; averaging over $s_0\sim\rho_0$, using equation 6 gives the result,

$$\begin{aligned}\nabla_L J(\theta_L,\pi_F^\alpha(\theta_L)) &= \nabla_L V_L^{\pi_L,\pi_F^\alpha}(s_0)\\ &= \sum_s\sum_{k=0}\gamma_L^k\mathcal{P}r\left(s_0\to s\mid k,\pi_L,\pi_F^\alpha\right)\sum_a\sum_b\big(\nabla_L\pi_L(a\mid s)\pi_F^\alpha(b\mid a,s)\\ &\qquad + \pi_L(a\mid s)\nabla_L\pi_F^\alpha(b\mid a,s)\big)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b)\\ &= \sum_s\eta(s)\sum_a\sum_b\big(\nabla_L\pi_L(a\mid s)\pi_F^\alpha(b\mid a,s)\\ &\qquad + \pi_L(a\mid s)\nabla_L\pi_F^\alpha(b\mid a,s)\big)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b)\\ &= \frac{1}{1-\gamma_L}\sum_s\frac{\eta(s)}{\sum_{s'}\eta(s')}\sum_a\sum_b\big(\nabla_L\pi_L(a\mid s)\pi_F^\alpha(b\mid a,s)\\ &\qquad + \pi_L(a\mid s)\nabla_L\pi_F^\alpha(b\mid a,s)\big)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b)\\ &= \frac{1}{1-\gamma_L}\sum_s d^\beta(s)\sum_a\sum_b\big(\nabla_L\pi_L(a\mid s)\pi_F^\alpha(b\mid a,s)\\ &\qquad + \pi_L(a\mid s)\nabla_L\pi_F^\alpha(b\mid a,s)\big)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b)\\ &= \frac{1}{1-\gamma_L}\mathbb{E}_{s\sim d^\beta}\left[\sum_a\sum_b\big(\nabla_L\pi_L(a\mid s)\pi_F^\alpha(b\mid a,s)\right.\\ &\qquad \left. + \pi_L(a\mid s)\nabla_L\pi_F^\alpha(b\mid a,s)\big)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b)\right]\end{aligned}$$

Where

$$\eta(s) = \sum_{k=0}^\infty\gamma_L^k\Pr(s_k = s\mid\pi_L,\pi_F^\alpha),\qquad d^\beta(s) = (1-\gamma_L)\eta(s)$$

The issue, however, is that this gradient requires trajectories under the joint policy $\beta(b, a \mid s) = \pi_L(a \mid s)\pi_F^{\alpha}(b \mid a, s)$ which itself requires learning the optimal follower policy π_F^* first. To avoid this, we use an importance sampling based approach to make use of the already available trajectory roll outs we obtained from learning π_F^*.

Let $v(b, a \mid s) = \pi_L(a \mid s)\pi_F^*(b \mid a, s)$, $w(b, a, s) = \frac{\pi_F^{\alpha}(b|a,s)}{\pi_F^*(b|a,s)}$ and $\zeta(s) = \frac{d_\beta(s)}{d_v(s)}$. We have:

$$
\begin{aligned}
\nabla_L J(\theta_L) &= \nabla_L V_L^{\pi_L, \pi_F^{\alpha}}(s_0) \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{s\sim d_\beta}\Bigg[\sum_a \sum_b \big(\nabla_L \pi_L(a \mid s)\pi_F^{\alpha}(b \mid a, s) \\
&\qquad + \pi_L(a \mid s)\nabla_L \pi_F^{\alpha}(b \mid a, s)\big)\mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\Bigg] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{s\sim d_\beta}\left[\sum_a \sum_b \nabla_L \left(\pi_L(a \mid s)\pi_F^{\alpha}(b \mid a, s)\right) \mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\right] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{s\sim d_\beta}\left[\sum_a \sum_b \nabla_L \left(\frac{\pi_F^{\alpha}(b \mid a, s)}{\pi_F^*(b \mid a, s)}\pi_L(a \mid s)\pi_F^*(b \mid a, s)\right) \mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\right] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{s\sim d_\beta}\left[\sum_a \sum_b \nabla_L \left(w(b, a, s)v(b, a \mid s)\right) \mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\right] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{s\sim d_\beta}\Bigg[\sum_a \sum_b \big(\nabla_L w(b, a, s)v(b, a \mid s) \\
&\qquad + w(b, a, s)\nabla_L v(b, a \mid s)\big)\mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\Bigg] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{s\sim d_\beta}\Bigg[\sum_a \sum_b \big(\nabla_L w(b, a, s)v(b, a \mid s) \\
&\qquad + w(b, a, s)v(b, a \mid s)\nabla_L \log v(b, a \mid s)\big)\mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\Bigg] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{s\sim d_\beta}\Bigg[\sum_a \sum_b v(b, a \mid s)\big(\nabla_L w(b, a, s) \\
&\qquad + w(b, a, s)\nabla_L \log v(b, a \mid s)\big)\mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\Bigg] \\
\nabla_L J(\theta_L) &= \frac{1}{1-\gamma_L} \mathbb{E}_{\substack{s\sim d_\beta \\ a,b\sim v}}\left[\left(\nabla_L w(b, a, s) + w(b, a, s)\nabla_L \log v(b, a \mid s)\right) \mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\right] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{\substack{s\sim d_v \\ a,b\sim v}}\left[\zeta(s)\left(\nabla_L w(b, a, s) + w(b, a, s)\nabla_L \log v(b, a \mid s)\right) \mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\right]
\end{aligned}
$$

Where $d_v = d^{\pi_L, \pi_F^*}$.

Developing further:

$$
\begin{aligned}
\nabla_L J(\theta_L) &= \frac{1}{1-\gamma_L} \mathbb{E}_{\substack{s\sim d_v \\ a,b\sim v}}\left[\zeta(s)w(b, a, s)\nabla_L \log \beta(b, a \mid s)\mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\right] \\
&= \frac{1}{1-\gamma_L} \mathbb{E}_{\substack{s\sim d_v \\ a,b\sim v}}\Bigg[\zeta(s)w(b, a, s)\big(\nabla_L \log \pi_L(a \mid s) \\
&\qquad + \nabla_L \log \pi_F^{\alpha}(b \mid a, s)\big)\mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\Bigg]
\end{aligned}
$$

Using equation 5 in Lemma A.2,

$$
\begin{aligned}
\nabla_L J(\theta_L) = \frac{1}{1-\gamma_L} \mathbb{E}_{\substack{s\sim d_v \\ a,b\sim v}}\Bigg[\zeta(s)w(b, a, s)\Bigg(&\frac{\nabla_L Q_F^{\pi_L, \pi_F^*}(s, a, b) - \mathbb{E}_{b'\sim\pi_F^{\alpha}}[\nabla_L Q_F^{\pi_L, \pi_F^*}(s, a, b')]}{\lambda} \\
&+\nabla_L \log \pi_L(a \mid s) - \alpha\big(\nabla_L Q_L^{\pi_L, \pi_F^*}(s, a, b) - \mathbb{E}_{b'\sim\pi_F^{\alpha}}[\nabla_L Q_L^{\pi_L, \pi_F^*}(s, a, b')]\big)\Bigg) \\
&\cdot \mathbf{Q_L}^{\pi_\mathbf{L}, \pi_\mathbf{F}^{\alpha}}(s, a, b)\Bigg]
\end{aligned}
$$

Following Kudo et al. (Kudo et al., 2026), we seek to estimate the hypergradient directly from interaction trajectories without requiring multiple follower actions to be sampled from the same state. Such repeated-state sampling can be impractical in large or continuous state spaces, where identical states may rarely be revisited without additional rollouts or arbitrary resets. The covariance-based formulation avoids this requirement by expressing the relevant centered terms as expectations that can be estimated from ordinary trajectory samples.

Our setting additionally requires off-policy correction, since trajectories are collected under the nominal joint policy v while the leader is evaluated under the modified policy β. The importance ratios w and ζ correct the follower-action and state-distribution shifts, respectively, allowing expectations under β to be estimated from trajectories generated under v. This preserves the sample-efficiency advantage of trajectory-based estimation without requiring additional rollouts from the modified follower. Hence, from Lemma A.3 we have:

$$\nabla_L J(\theta_L) = \frac{1}{1-\gamma_L}\mathbb{E}_{\substack{s\sim d_v \\ a,b\sim v}}\Big[\zeta(s)w(b,a,s)\Big(\nabla_L \log \pi_L(a \mid s)\mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^\alpha}(s,a,b) + B_L^{\pi_L,\pi_F^\alpha}(s,a,b)\big(\frac{1}{\lambda}\nabla_L \mathbf{Q_F}^{\pi_\mathbf{L},\pi_\mathbf{F}^*}(s,a,b) - \alpha\nabla_L \mathbf{Q_L}^{\pi_\mathbf{L},\pi_\mathbf{F}^*}(s,a,b)\big)\Big)\Big]$$

■

A.5 Proof of Proposition 4.2: Local KL robustness

Fix a leader policy π_L and a state–action pair (s,a). For brevity, let

$$p(b) = \pi_F^*(b \mid s,a), \qquad z_b = Q_L^{\pi_L,\pi_F^*}(s,a,b),$$

and define

$$q_\alpha(b) = \frac{p(b)e^{-\alpha z_b}}{Z_\alpha}, \qquad Z_\alpha = \sum_{b'} p(b')e^{-\alpha z_{b'}}.$$

Proof. We first show that q_α is the unique solution of

$$\min_{q\in\Delta(\mathcal{A}_F)} \left\{\mathbb{E}_{b\sim q}[z_b] + \frac{1}{\alpha}D_{\mathrm{KL}}(q\|p)\right\}.$$

For any $q \in \Delta(\mathcal{A}_F)$,

$$\log q_\alpha(b) = \log p(b) - \alpha z_b - \log Z_\alpha.$$

Hence

$$\begin{aligned} D_{\mathrm{KL}}(q\|q_\alpha) &= \sum_b q(b)\log\frac{q(b)}{q_\alpha(b)} \\ &= \sum_b q(b)\left[\log\frac{q(b)}{p(b)} + \alpha z_b + \log Z_\alpha\right] \\ &= D_{\mathrm{KL}}(q\|p) + \alpha\mathbb{E}_{b\sim q}[z_b] + \log Z_\alpha. \end{aligned}$$

Rearranging,

$$\boxed{\mathbb{E}_{b\sim q}[z_b] + \frac{1}{\alpha}D_{\mathrm{KL}}(q\|p) = -\frac{1}{\alpha}\log Z_\alpha + \frac{1}{\alpha}D_{\mathrm{KL}}(q\|q_\alpha).}$$

Since

$$D_{\mathrm{KL}}(q\|q_\alpha) \geq 0,$$

with equality if and only if $q = q_\alpha$, it follows immediately that

$$q_\alpha = \arg\min_{q\in\Delta(\mathcal{A}_F)} \left\{ \mathbb{E}_{b\sim q}[z_b] + \frac{1}{\alpha} D_{\mathrm{KL}}(q\|p) \right\}.$$

Substituting back the original notation gives

$$\boxed{\pi_F^\alpha(\cdot \mid s,a) = \arg\min_{q\in\Delta(\mathcal{A}_F)} \left\{ \mathbb{E}_{b\sim q}\left[Q_L^{\pi_L,\pi_F^*}(s,a,b) \right] + \frac{1}{\alpha} D_{\mathrm{KL}}\left(q\|\pi_F^*(\cdot \mid s,a)\right) \right\}.}$$

Now define the attained KL radius

$$\kappa_\alpha(s,a) = D_{\mathrm{KL}}\left(q_\alpha\|p\right).$$

We show that q_α also solves

$$\min_{\substack{q\in\Delta(\mathcal{A}_F) \\ D_{\mathrm{KL}}(q\|p)\leq\kappa_\alpha(s,a)}} \mathbb{E}_{b\sim q}[z_b].$$

Suppose, for contradiction, that there exists a feasible q' such that

$$D_{\mathrm{KL}}(q'\|p) \leq D_{\mathrm{KL}}(q_\alpha\|p)$$

and

$$\mathbb{E}_{q'}[z_b] < \mathbb{E}_{q_\alpha}[z_b].$$

Then

$$\mathbb{E}_{q'}[z_b] + \frac{1}{\alpha} D_{\mathrm{KL}}(q'\|p) < \mathbb{E}_{q_\alpha}[z_b] + \frac{1}{\alpha} D_{\mathrm{KL}}(q_\alpha\|p),$$

which contradicts the optimality of q_α for the KL-regularized problem established above. Therefore,

$$\boxed{q_\alpha \in \arg\min_{\substack{q\in\Delta(\mathcal{A}_F) \\ D_{\mathrm{KL}}(q\|p)\leq\kappa_\alpha(s,a)}} \mathbb{E}_{b\sim q}[z_b].}$$

Thus, for every (s,a), the modified follower response is exactly the response that minimizes the leader's nominal continuation value within the KL neighborhood of radius $\kappa_\alpha(s,a)$ around the entropy-regularized best response. ■

A.6 Proof of Theorem 4.3: Controlled response perturbation

For a policy π_L let $\pi_F^*(\pi_L) \in \arg\max_{\pi_F} J_F^\lambda(\pi_L,\pi_F)$ a best-response to a π_L. We say π_F^α is an ϵ_F-best response if and only if:

$$J_F^\lambda(\pi_L,\pi_F^\alpha) \geq J_F^\lambda(\pi_L,\pi_F^*) - \epsilon_F$$

We use the entropy-regularized follower best response formula:

$$\pi_F^*(b \mid s,a) = \frac{\exp\left(Q_F^{\pi_L,\pi_F^*}(s,a,b)/\lambda\right)}{\sum_{b'}\exp\left(Q_F^{\pi_L,\pi_F^*}(s,a,b')/\lambda\right)},$$

Lemma A.4 (Optimality gap). *Fix the leader policy π_L for any follower policy π_F.*

For the sake of simplicity we use $p(b \mid s,a)$ and $q_\alpha(b \mid s,a)$ instead of $\pi_F^*(b \mid s,a)$ and $\pi_F^\alpha(b \mid s,a)$ respectively. We use simply $p(b)$ and $q_\alpha(b)$ when there's no ambiguity about (s,a).

The entropy-regularized return takes the shape:

$$\begin{aligned}\mathbf{V}_F^{\pi_F}(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma_F^t\big(r_F(s_t,a_t,b_t) - \lambda\log\pi_F(b_t \mid s_t,a_t)\big) \;\middle|\; s_0 = s\right] \\ &= \mathbb{E}\Bigg[r_F(s,a,b) - \lambda\log\pi_F(b \mid s,a) \\ &\qquad + \sum_{t=1}^{\infty}\gamma_F^{t-1}\big(r_F(s_t,a_t,b_t) - \lambda\log\pi_F(b_t \mid s_t,a_t)\big) \;\Bigg|\; s_0 = s\Bigg] \\ &= \sum_a \pi_L(a \mid s)\sum_b \pi_F(b \mid s,a)\Big[r_f(s,a,b) - \lambda\log\pi_F(b \mid s,a) \\ &\qquad + \gamma_F\sum_{s'}P(s' \mid s,a,b)\mathbf{V}_F^{\pi_F}(s')\Big]\end{aligned}$$

The value functions under the modified follower policy q_α:

$$\mathbf{V}_F^{q_\alpha}(s) = \sum_a \pi_L(a \mid s)\sum_b q_\alpha(b)\Big[r_f(s,a,b) - \lambda\log q_\alpha(b) + \gamma_F\sum_{s'}P(s' \mid s,a,b)\mathbf{V}_F^{q_\alpha}(s')\Big] \tag{7}$$

Under the optimal follower policy p it takes the shape:

$$\mathbf{V}_F^*(s) = \mathbf{V}_F^p(s) = \sum_a \pi_L(a \mid s)\sum_b p(b)\Big[\mathbf{Q}_F^*(s,a,b) - \lambda\log p(b)\Big] \tag{8}$$

Subtracting the two values (8)–(7), we define the follower optimality-gap:

$$\Delta(s) := \mathbf{V}_F^*(s) - \mathbf{V}_F^{q_\alpha}(s)$$

Back to the formulation of $\mathbf{V}_F^{q_\alpha}(s)$ we can write:

$$\mathbf{V}_F^{q_\alpha}(s) = \sum_a \pi_L(a \mid s)\sum_b q_\alpha(b)\Big[\mathbf{Q}_F^*(s,a,b) - \lambda\log q_\alpha(b) - \gamma_F\sum_{s'}P(s' \mid s,a,b)\big(\underbrace{\mathbf{V}_F^*(s') - \mathbf{V}_F^{q_\alpha}(s')}_{\Delta(s')}\big)\Big]$$

Now subtract this from $\mathbf{V}_F^*(s)$:

$$\begin{aligned}\Delta(s) = \sum_a \pi_L(a \mid s)\Bigg[&\sum_b p(b)\big(\mathbf{Q}_F^*(s,a,b) - \lambda\log p(b)\big) \\ &- \sum_b q_\alpha(b)\big(\mathbf{Q}_F^*(s,a,b) - \lambda\log q_\alpha(b)\big)\Bigg] \\ &+ \gamma_F\sum_a \pi_L(a \mid s)\sum_b q_\alpha(b)\sum_{s'}P(s' \mid s,a,b)\Delta(s')\end{aligned}$$

We define only for convenience the first large bracket as the immediate regularized loss:

$$g(s) = \sum_a \pi_L(a \mid s)\,[regularized\ value\ under\ p - regularized\ value\ under\ q_\alpha]$$

Now we fully unroll the difference starting from s_0,

$$\Delta(s_0) = g(s_0) + \gamma_F\sum_{a_0,b_0,s_1}\pi_L(a_0 \mid s_0)q_\alpha(b_0 \mid s_0,a_0)P(s_1 \mid s_0,a_0,b_0)\Delta(s_1)$$

Substitute the same expression for $\Delta(s_1)$:

$$\begin{aligned}\Delta(s_0) =& g(s_0) \\ &+ \gamma_F\sum_{a_0,b_0,s_1}\pi_L(a_0 \mid s_0)q_\alpha(b_0 \mid s_0,a_0)P(s_1 \mid s_0,a_0,b_0)g(s_1) \\ &+ \gamma_F^2\sum_{\substack{a_0,b_0,s_1\\a_1,b_1,s_2}}\pi_L(a_0 \mid s_0)q_\alpha(b_0 \mid s_0,a_0)P(s_1 \mid s_0,a_0,b_0) \\ &\qquad\cdot\pi_L(a_1 \mid s_1)q_\alpha(b_1 \mid s_1,a_1)P(s_2 \mid s_1,a_1,b_1)\Delta(s_2)\end{aligned}$$

Substitute again

$$\begin{aligned}\Delta(s_0) =& g(s_0)\\ &+ \gamma_F \mathbb{E}_{\pi_L, q_\alpha}[g(s_1) \mid s_0]\\ &+ \gamma_F^2 \mathbb{E}_{\pi_L, q_\alpha}[g(s_2) \mid s_0]\\ &+ \gamma_F^3 \mathbb{E}_{\pi_L, q_\alpha}[g(s_3) \mid s_0]\\ &\vdots\end{aligned}$$

Thus the optimality gap can be written as:

$$\Delta(s) = \mathbb{E}_{\pi_L, q_\alpha}\left[\sum_{t=0}^{\infty} \gamma_F^t g(s_t) \mid s_0\right] \tag{9}$$

Proof. Let, for a fixed $(s, a) \in \mathcal{S} \times \mathcal{A}_L$, $p(b) = \pi_F^*(b \mid s, a)$, $q_\alpha(b) = \pi_F^\alpha(b \mid s, a)$ and $z_b = \mathbf{Q_L^{\pi_L, \pi_F^*}}(s, a, b)$.

Then $q_\alpha(b) = \frac{p(b)\exp(-\alpha z_b)}{\mathbf{Z}(\alpha)}$, where $\mathbf{Z}(\alpha) = \sum_b p(b)\exp(-\alpha z_b)$.

Notice that in Lemma A.4:

$$g(s) = \lambda \sum_a \pi_L(a \mid s) D_{KL}(q_\alpha(\cdot \mid s, a) \| p(\cdot \mid s, a)). \tag{10}$$

Let's define $\psi(\alpha) = \log(\mathbf{Z}(\alpha))$. Notice that $\psi'(\alpha) = -\mathbb{E}_{b\sim q_\alpha}[z_b]$ and $\psi''(\alpha) = Var_{b\sim q_\alpha}(z_b)$.

$$\begin{aligned}D_{KL}(q_\alpha \| p) &= \sum_b q_\alpha(b) \log \frac{q_\alpha(b)}{p(b)}\\ &= -\alpha \mathbb{E}_{b\sim q_\alpha}[z_b] - \log Z(\alpha)\\ &= \alpha\psi'(\alpha) - \psi(\alpha)\end{aligned} \tag{11}$$

Suppose further the leader's Q-function has uniformly bounded action span:

$$span_b Q_L^{\pi_L, \pi_F^*}(s, a, b) := \max_b Q_L^{\pi_L, \pi_F^*}(s, a, b) - \min_b Q_L^{\pi_L, \pi_F^*}(s, a, b) \le C_L$$

Popoviciu's inequality on variances gives us:

$$Var_{b\sim q_\alpha}(z_b) \le \frac{C_L^2}{4} \tag{12}$$

Notice $D_{KL}(q_0 \| p) = 0$. Given equation 11 and inequality 12 we can write:

$$\begin{aligned}D_{KL}(q_\alpha \| p) &= \int_0^\alpha t\psi''(t) d_t\\ &\le \int_0^\alpha t \frac{C_L^2}{4} d_t\\ &\le \frac{\alpha^2 C_L^2}{8} = \overline{g}\end{aligned} \tag{13}$$

Given inequality 13 and equation 10:

$$\begin{aligned}g(s) &\le \lambda \sum_a \pi_L(a \mid s)\overline{g}\\ &\le \lambda \overline{g}\end{aligned}$$

Then trivially, since

$$\begin{aligned}\Delta(s_0) &\le \lambda\overline{g} + \gamma_F \lambda \overline{g} + \gamma_F^2 \lambda \overline{g} \ldots\\ &\le \frac{\lambda \overline{g}}{1 - \gamma_F}\end{aligned}$$

We get:

$$0 \le \mathbf{V}_F^*(s_0) - \mathbf{V}_F^{\pi_F^\alpha}(s_0) \le \frac{\lambda \alpha^2 C_L^2}{8(1-\gamma_F)} \tag{14}$$

We now establish the second statement of Theorem 4.3. Keeping the notation above, let

$$p(b \mid s,a) = \pi_F^*(b \mid s,a), \qquad q_\alpha(b \mid s,a) = \pi_F^\alpha(b \mid s,a),$$

and define the leader-value gap

$$\Delta_L(s) := V_L^{\pi_L,p}(s) - V_L^{\pi_L,q_\alpha}(s).$$

Under the modified follower response,

$$V_L^{\pi_L,q_\alpha}(s) = \sum_a \pi_L(a \mid s) \sum_b q_\alpha(b \mid s,a) \left[r_L(s,a,b) + \gamma_L \sum_{s'} P(s' \mid s,a,b) V_L^{\pi_L,q_\alpha}(s') \right].$$

Using

$$Q_L^{\pi_L,p}(s,a,b) = r_L(s,a,b) + \gamma_L \sum_{s'} P(s' \mid s,a,b) V_L^{\pi_L,p}(s'),$$

we may rewrite this as

$$V_L^{\pi_L,q_\alpha}(s) = \sum_a \pi_L(a \mid s) \sum_b q_\alpha(b \mid s,a) \Bigg[Q_L^{\pi_L,p}(s,a,b) - \gamma_L \sum_{s'} P(s' \mid s,a,b) \Delta_L(s') \Bigg].$$

Meanwhile,

$$V_L^{\pi_L,p}(s) = \sum_a \pi_L(a \mid s) \sum_b p(b \mid s,a) Q_L^{\pi_L,p}(s,a,b).$$

Subtracting the two expressions gives

$$\Delta_L(s) = \sum_a \pi_L(a \mid s) \left[\sum_b p(b \mid s,a) Q_L^{\pi_L,p}(s,a,b) - \sum_b q_\alpha(b \mid s,a) Q_L^{\pi_L,p}(s,a,b) \right] + \gamma_L \sum_a \pi_L(a \mid s) \sum_b q_\alpha(b \mid s,a) \sum_{s'} P(s' \mid s,a,b) \Delta_L(s').$$

Define the immediate leader-value loss

$$h(s) := \sum_a \pi_L(a \mid s) \left[\mathbb{E}_{b \sim p(\cdot \mid s,a)} Q_L^{\pi_L,p}(s,a,b) - \mathbb{E}_{b \sim q_\alpha(\cdot \mid s,a)} Q_L^{\pi_L,p}(s,a,b) \right].$$

Hence,

$$\Delta_L(s) = h(s) + \gamma_L \sum_a \pi_L(a \mid s) \sum_b q_\alpha(b \mid s,a) \sum_{s'} P(s' \mid s,a,b) \Delta_L(s'). \tag{15}$$

For a fixed (s,a), let

$$z_b = Q_L^{\pi_L, p}(s, a, b)$$

and introduce the family

$$q_t(b) = \frac{p(b)e^{-tz_b}}{\sum_{b'} p(b')e^{-tz_{b'}}}, \qquad t \in [0, \alpha].$$

Thus $q_0 = p$ and $q_\alpha = \pi_F^\alpha$. Define

$$m(t) = \mathbb{E}_{b\sim q_t}[z_b].$$

Differentiating,

$$\begin{aligned} m'(t) &= \frac{d}{dt}\sum_b q_t(b) z_b \\ &= -\left(\mathbb{E}_{q_t}[z_b^2] - \mathbb{E}_{q_t}[z_b]^2\right) \\ &= -\operatorname{Var}_{b\sim q_t}(z_b). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}_p[z_b] - \mathbb{E}_{q_\alpha}[z_b] &= m(0) - m(\alpha) \\ &= \int_0^\alpha \operatorname{Var}_{b\sim q_t}(z_b)\, dt. \end{aligned} \tag{16}$$

In particular,

$$\mathbb{E}_p[z_b] - \mathbb{E}_{q_\alpha}[z_b] \geq 0.$$

By the assumed uniform action-span bound,

$$\max_b z_b - \min_b z_b \leq C_L.$$

Popoviciu's inequality therefore gives, for every $t \in [0, \alpha]$,

$$\operatorname{Var}_{b\sim q_t}(z_b) \leq \frac{C_L^2}{4}.$$

Using equation 16,

$$0 \leq \mathbb{E}_p[z_b] - \mathbb{E}_{q_\alpha}[z_b] \leq \int_0^\alpha \frac{C_L^2}{4}\, dt = \frac{\alpha C_L^2}{4}.$$

Since $h(s)$ is a $\pi_L(\cdot \mid s)$-weighted average of these quantities,

$$\boxed{0 \leq h(s) \leq \frac{\alpha C_L^2}{4}.} \tag{17}$$

We can now unroll 15. Starting from s_0,

$$\Delta_L(s_0) = h(s_0) + \gamma_L \sum_{a_0, b_0, s_1} \pi_L(a_0 \mid s_0) q_\alpha(b_0 \mid s_0, a_0) P(s_1 \mid s_0, a_0, b_0) \Delta_L(s_1).$$

Substituting recursively yields

$$\Delta_L(s_0) = h(s_0) + \gamma_L \mathbb{E}_{\pi_L, q_\alpha}[h(s_1) \mid s_0] + \gamma_L^2 \mathbb{E}_{\pi_L, q_\alpha}[h(s_2) \mid s_0] + \cdots,$$

and therefore

$$\boxed{\Delta_L(s_0) = \mathbb{E}_{\pi_L, q_\alpha}\left[\sum_{t=0}^{\infty} \gamma_L^t h(s_t) \,\middle|\, s_0\right].} \tag{18}$$

Using inequality 17,

$$0 \le \Delta_L(s_0) \le \sum_{t=0}^{\infty} \gamma_L^t \frac{\alpha C_L^2}{4} = \frac{\alpha C_L^2}{4(1-\gamma_L)}.$$

Hence,

$$\boxed{0 \le V_L^{\pi_L, p}(s_0) - V_L^{\pi_L, q_\alpha}(s_0) \le \frac{\alpha C_L^2}{4(1-\gamma_L)}.} \tag{19}$$

Finally, taking the expectation over $s_0 \sim \rho_0$ gives

$$\boxed{0 \le J_L(\pi_L, \pi_F^*) - J_L(\pi_L, \pi_F^\alpha) \le \frac{\alpha C_L^2}{4(1-\gamma_L)}.}$$

This completes the proof of Theorem 4.3. ■

A.7 Proof of Corollary 4.4

By Theorem 4.3, the uniform span assumption implies that for every leader policy π_L,

$$J_L(\pi_L, p_{\pi_L}) - \rho_\alpha \le J_L(\pi_L, q_{\pi_L}^\alpha) \le J_L(\pi_L, p_{\pi_L}).$$

Applying the left inequality at the nominal optimizer π_L^0 gives

$$J_L(\pi_L^0, q_{\pi_L^0}^\alpha) \ge J_L(\pi_L^0, p_{\pi_L^0}) - \rho_\alpha.$$

Since π_L^α maximizes the modified objective,

$$J_L(\pi_L^\alpha, q_{\pi_L^\alpha}^\alpha) \ge J_L(\pi_L^0, q_{\pi_L^0}^\alpha),$$

and therefore

$$J_L(\pi_L^\alpha, q_{\pi_L^\alpha}^\alpha) \ge J_L(\pi_L^0, p_{\pi_L^0}) - \rho_\alpha.$$

Theorem 4.3 also gives

$$J_L(\pi_L^\alpha, p_{\pi_L^\alpha}) \ge J_L(\pi_L^\alpha, q_{\pi_L^\alpha}^\alpha),$$

while the optimality of π_L^0 for the nominal objective implies

$$J_L(\pi_L^\alpha, p_{\pi_L^\alpha}) \le J_L(\pi_L^0, p_{\pi_L^0}).$$

Combining these inequalities yields

$$J_L(\pi_L^0, p_{\pi_L^0}) - \rho_\alpha \le J_L(\pi_L^\alpha, q_{\pi_L^\alpha}^\alpha) \le J_L(\pi_L^\alpha, p_{\pi_L^\alpha}) \le J_L(\pi_L^0, p_{\pi_L^0}),$$

which proves the result.

B Off-Policy Estimation and Implementation Details

Algorithm 1 summarizes the tabular RBC-HG procedure used in our experiments. The individual components of the update are described in detail in Sections B.1→ B.5.

B.1 Follower response and sampling

In the reported experiments, the follower is white-box: its tabular soft value function is available to the leader-side estimator. In settings where the follower is not directly accessible, $\widehat{V}_F$ would need to be estimated from follower trajectories, as in black-box variants of bilevel response estimation; we do not study this extension here.

The follower lower-level problem is solved using tabular soft Q-iteration. For a fixed leader policy, the soft Bellman operator is iterated until convergence and the nominal follower response is obtained as

$$\pi_F^*(b \mid s, a) = \mathrm{softmax}\left(Q_F^*(s, a, \cdot)/\lambda\right)_b .$$

We use $\lambda = 0.05$ and convergence tolerance 10^{-6}. Since $\lambda > 0$, the Boltzmann response has full support over the finite follower action space. Trajectories are generated under the nominal joint policy

$$v(a, b \mid s) = \pi_L(a \mid s)\pi_F^*(b \mid s, a).$$

For actor updates, transitions are sampled with probabilities proportional to γ_L^t, where t denotes the time step within the trajectory.

B.2 Leader-value estimation

We maintain separate critics for the nominal and modified follower responses. The nominal critic $\widehat{Q}_L^v$ approximates $Q_L^{\pi_L, \pi_F^*}$ and defines

$$\pi_F^\alpha(b \mid s, a) \propto \pi_F^*(b \mid s, a) \exp\left[-\alpha \widehat{Q}_L^v(s, a, b)\right].$$

The critic values used in this transformation are detached from the actor computation graph; the dependence of the transformation on the leader policy is accounted for separately through the hypergradient correction below.

A second critic $\widehat{Q}_L^\beta$ evaluates the modified joint policy

$$\beta(a, b \mid s) = \pi_L(a \mid s)\pi_F^\alpha(b \mid s, a).$$

It is fitted from nominal-policy transitions using the Bellman target

$$y^\beta = r_L + \gamma_L(1 - d)\mathbb{E}_{\substack{a' \sim \pi_L(\cdot \mid s') \\ b' \sim \pi_F^\alpha(\cdot \mid s', a')}} [\widehat{Q}_{L,\mathrm{targ}}^\beta(s', a', b')].$$

Thus the observed transition is reused off-policy while the continuation value is evaluated under the modified follower response. Target critics are updated by Polyak averaging.

B.3 Follower-action correction

For a nominal-policy transition (s, a, b), the change in follower-action distribution is corrected using

$$w(s, a, b) = \frac{\pi_F^\alpha(b \mid s, a)}{\pi_F^*(b \mid s, a)}.$$

The entropy-regularized nominal response provides the support required by this ratio. To reduce estimator variance, the implementation uses

$$\bar{w}(s,a,b) = \min\{w(s,a,b), w_{\max}\}, \qquad w_{\max} = 10.$$

Consequently, importance-weight clipping introduces the usual bias-variance trade-off relative to the exact identity in Theorem 4.1.

B.4 State-occupancy correction

Changing the follower response can also alter the distribution of visited states. In the tabular experiments, the transition kernel is available, allowing us to compute this correction directly rather than introducing a learned density-ratio estimator. For $\chi \in \{v, \beta\}$, we construct the induced state-transition matrix

$$P_\chi(s,s') = \sum_{a,b} \chi(a,b \mid s) P(s' \mid s,a,b).$$

For the infinite-horizon objective, the normalized discounted occupancy is

$$d^\chi = (1-\gamma_L) \sum_{t=0}^{\infty} \gamma_L^t \rho_0 P_\chi^t,$$

which can equivalently be obtained by solving

$$d^{\chi\top} = (1-\gamma_L)(I - \gamma_L P_\chi^\top)^{-1} \rho_0^\top.$$

However, the experimental trajectories have finite horizon H, and replay samples are weighted proportionally to γ_L^t. We therefore use the corresponding finite-horizon occupancy

$$d_H^\chi(s) = \frac{\sum_{t=0}^{H-1} \gamma_L^t [\rho_0 P_\chi^t](s)}{\sum_{t=0}^{H-1} \gamma_L^t},$$

and define

$$\zeta_H(s) = \frac{d_H^\beta(s)}{d_H^v(s)}.$$

This correction exactly matches the discounted state-sampling distribution used by the replay buffer in the finite tabular experiments. We additionally compute the infinite-horizon ratio during training as a diagnostic and record its maximum deviation from ζ_H.

B.5 Exact Tabular Value Sensitivities

The hypergradient in Theorem 4.1 depends on the two value sensitivities

$$G_F(s,a,b) := \nabla_L Q_F^{\pi_L, \pi_F^*}(s,a,b), \qquad G_L(s,a,b) := \nabla_L Q_L^{\pi_L, \pi_F^*}(s,a,b)$$

In the finite tabular experiments, we compute these quantities by differentiating their Bellman fixed points rather than by using a trajectory-suffix approximation.

This choice exploits the known transition model in the reported toy environments and allows us to isolate the effect of the pessimistic response mechanism from additional Monte Carlo error in the value sensitivities. The follower remains white-box, as described in Appendix B.1, and its current soft value table is used as the nominal follower fixed point.

Let

$$p(b \mid s, a) := \pi_F^*(b \mid s, a),$$

and define the nominal state-transition matrix

$$P_v(s, s') = \sum_{a,b} \pi_L(a \mid s) p(b \mid s, a) P(s' \mid s, a, b)$$

Follower-value sensitivity. Under the action-observing follower convention used in the experiments, the soft follower value is

$$V_F(s, a) = \lambda \log \sum_b \exp\left(\frac{Q_F(s, a, b)}{\lambda}\right),$$

and

$$U_F(s) = \sum_a \pi_L(a \mid s) V_F(s, a)$$

The follower Bellman equation is therefore

$$Q_F(s, a, b) = r_F(s, a, b) + \gamma_F \sum_{s'} P(s' \mid s, a, b) U_F(s')$$

Differentiating U_F with respect to the leader parameters gives the state-wise direct term

$$h_F(s) = \sum_a \nabla_L \pi_L(a \mid s)\, V_F(s, a),$$

or, equivalently,

$$h_F(s) = \sum_a \pi_L(a \mid s) \nabla_L \log \pi_L(a \mid s)\, V_F(s, a)$$

Because differentiating the soft value with respect to its action values produces the nominal follower probabilities, the remaining continuation term propagates through the nominal transition matrix P_v. Hence the state sensitivity

$$D_F(s) := \nabla_L U_F(s)$$

satisfies

$$D_F = h_F + \gamma_F P_v D_F,$$

and therefore

$$D_F = (I - \gamma_F P_v)^{-1} h_F$$

Since the initial tuple (s, a, b) is conditioned on in $Q_F(s, a, b)$, its leader and follower actions are not differentiated. The corresponding follower action-value sensitivity is consequently

$$\boxed{G_F(s, a, b) = \gamma_F \sum_{s'} P(s' \mid s, a, b) D_F(s')}$$

For the deterministic environments used in the experiments, this reduces to

$$G_F(s, a, b) = \gamma_F D_F(T(s, a, b)),$$

where T denotes the deterministic transition function.

Leader-value sensitivity. We next compute the total derivative of the nominal leader action value

$$Q_L^v := Q_L^{\pi_L, \pi_F^*}$$

The nominal leader state value is

$$V_L^v(s) = \sum_{a,b} \pi_L(a \mid s) p(b \mid s, a) Q_L^v(s, a, b)$$

Define the nominal follower-action Benefit

$$B_L^v(s, a, b) = Q_L^v(s, a, b) - \mathbb{E}_{\bar{b} \sim p(\cdot \mid s, a)} \left[Q_L^v(s, a, \bar{b}) \right]$$

The derivative of V_L^v contains two contributions. The first is the direct effect of changing the leader policy. The second is the induced change of the nominal follower response. Using the Boltzmann covariance identity underlying BC-HG, the latter can be expressed through the follower sensitivity G_F. We therefore define

$$h_L(s) = \sum_{a,b} \pi_L(a \mid s) p(b \mid s, a) \Big[Q_L^v(s,a,b) \nabla_L \log \pi_L(a \mid s) + \frac{1}{\lambda} B_L^v(s,a,b) G_F(s,a,b) \Big]$$

Let

$$D_L(s) := \nabla_L V_L^v(s)$$

The differentiated Bellman equation gives

$$D_L = h_L + \gamma_L P_v D_L,$$

and hence

$$D_L = (I - \gamma_L P_v)^{-1} h_L$$

Again, because the initial actions in $Q_L^v(s,a,b)$ are conditioned on, the derivative begins only through the successor state. Therefore

$$\boxed{G_L(s,a,b) = \gamma_L \sum_{s'} P(s' \mid s,a,b) D_L(s')}$$

For deterministic transitions,

$$G_L(s,a,b) = \gamma_L D_L(T(s,a,b))$$

This conditioning is important: in particular, when $\gamma_L = 0$, $Q_L^v(s,a,b) = r_L(s,a,b)$ and therefore $G_L(s,a,b) = 0$, as required.

Implementation. All tabular Bellman solves are performed in double precision. The transition matrices, nominal values, Benefits, and policy probabilities appearing as coefficients in the linear systems are treated as fixed quantities during the actor update. Rather than explicitly constructing a Jacobian with one column per leader-policy parameter, the implementation constructs differentiable score-function surrogates whose gradients equal the vectors G_F and G_L above.

The nominal leader values used in the sensitivity solve are obtained from the tabular Bellman equations under the known transition model. The learned nominal critic is still used to construct the pessimistically modified follower response,

$$\pi_F^\alpha(b \mid s,a) \propto \pi_F^*(b \mid s,a) \exp\Big(-\alpha \widehat{Q}_L^v(s,a,b)\Big),$$

while the separately learned modified-policy critic $\widehat{Q}_L^\beta$ is used in the outer leader objective.

The value-sensitivity solve is exact only conditional on the supplied nominal follower fixed point. In practice, incomplete lower-level convergence or reuse of a follower table across several leader updates can therefore introduce discrepancy. We monitor this through the soft Bellman residual of the follower and the mismatch between the supplied follower policy and the Boltzmann policy induced by its current Q-table.

B.6 Leader update

Define the modified-policy Benefit

$$\widehat{B}_L^\beta(s,a,b) = \widehat{Q}_L^\beta(s,a,b) - \mathbb{E}_{b' \sim \pi_F^\alpha(\cdot \mid s,a)} \Big[\widehat{Q}_L^\beta(s,a,b') \Big]. \tag{20}$$

As described in Appendix B.5, the exact tabular sensitivity solves are implemented through differentiable scalar surrogates $\mathcal{S}_F(s,a,b)$ and $\mathcal{S}_L(s,a,b)$ satisfying

$$\nabla_L \mathcal{S}_F(s,a,b) = G_F(s,a,b), \qquad \nabla_L \mathcal{S}_L(s,a,b) = G_L(s,a,b), \tag{21}$$

where

$$G_F(s,a,b) = \nabla_L Q_F^{\pi_L,\pi_F^*}(s,a,b), \qquad G_L(s,a,b) = \nabla_L Q_L^{\pi_L,\pi_F^*}(s,a,b)$$

We therefore define

$$\mathcal{S}_\alpha(s,a,b) = \frac{1}{\lambda}\mathcal{S}_F(s,a,b) - \alpha\mathcal{S}_L(s,a,b), \tag{22}$$

so that

$$\nabla_L \mathcal{S}_\alpha(s,a,b) = \frac{1}{\lambda}G_F(s,a,b) - \alpha G_L(s,a,b) \tag{23}$$

For a minibatch of N transitions sampled from the nominal-response replay buffer, the actor surrogate used in the experiments is

$$\widehat{\mathcal{L}}_L = -\frac{1}{N}\sum_{i=1}^{N} \mathrm{sg}(\zeta_H(s_i)\,\bar{w}(s_i,a_i,b_i)) \Big[\mathrm{sg}\Big(\widehat{Q}_L^\beta(s_i,a_i,b_i)\Big) \log \pi_L(a_i \mid s_i) + \mathrm{sg}\Big(\widehat{B}_L^\beta(s_i,a_i,b_i)\Big)\,\mathcal{S}_\alpha(s_i,a_i,b_i)\Big] \tag{24}$$

where sg denotes stop-gradient. Hence,

$$-\nabla_L \widehat{\mathcal{L}}_L = \frac{1}{N}\sum_{i=1}^{N} \zeta_H(s_i)\bar{w}(s_i,a_i,b_i)\Big[\widehat{Q}_L^\beta(s_i,a_i,b_i)\nabla_L \log \pi_L(a_i \mid s_i) + \widehat{B}_L^\beta(s_i,a_i,b_i)\left(\frac{1}{\lambda}G_F(s_i,a_i,b_i) - \alpha G_L(s_i,a_i,b_i)\right)\Big] \tag{25}$$

This is the finite-sample counterpart of the hypergradient in Theorem 4.1. The factor $1/(1-\gamma_L)$ is omitted because it is a positive parameter-independent scaling of the actor gradient. The exact identity in Theorem 4.1 uses the infinite-horizon occupancy ratio and unclipped follower-action likelihood ratios. In the reported experiments, ζ_H is instead the finite-horizon occupancy ratio defined in Appendix B.4, matching the discounted sampling distribution of the replay buffer, and the action ratio is clipped as $\bar{w} = \min\{w, w_{\max}\}$ for variance control. Furthermore, following the BCHG training protocol, several actor updates reuse the same recent replay buffer, so later updates may use transitions collected under a slightly earlier leader policy. We therefore regard the implemented update as a practical finite-horizon approximation to the exact hypergradient identity.

C Additional Experimental Details

C.1 Benchmark Environments

Figure 3 gives the transition and reward structures of the two tabular Stackelberg Markov games used in Section 5. Both contain three states, two leader actions, and three follower actions with deterministic transitions.

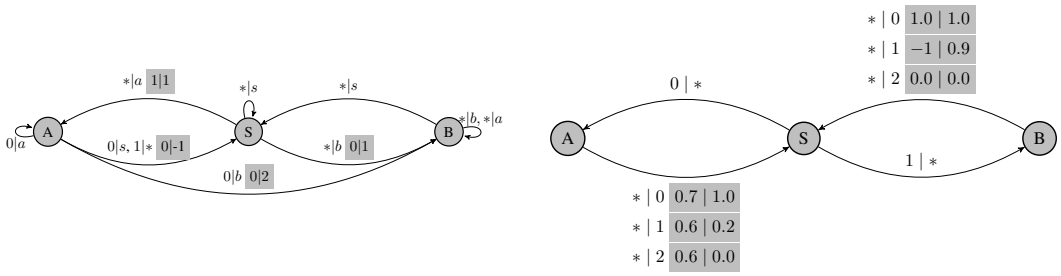
In the limited response ambiguity environment 3a, the leader's actions at states S and B do not affect either the transition dynamics or the rewards. The leader therefore only influences the follower through its policy at state A, and we can summarize this dependence by $p = \pi_L(0 \mid A)$.

When p is close to one, the follower prefers actions a, b, and s at states S, A, and B, respectively. At S, choosing a moves the process to A and gives both agents reward 1. Once in A, the follower chooses b, which moves the process to B and gives the follower reward 2, while the leader receives 0. At B, the follower chooses s, returning the process to S with zero reward for both agents. The resulting behavior therefore follows the cycle $S \to A \to B \to S$, with follower rewards $1, 2, 0$ and leader rewards $1, 0, 0$, respectively. The leader consequently obtains its positive reward whenever the follower selects a at S.

The attractiveness of this route to the follower depends on the leader continuing to choose action 0 sufficiently often at A. If the leader instead assigns high probability to action 1 at A, entering A becomes less attractive: under leader action 1, every follower response returns directly to S and gives the follower reward -1. As $\pi_L(1 \mid A)$ approaches one, the follower therefore prefers action

b already at S, bypassing A altogether. This produces the shorter cycle $S \to B \to S$, where the follower receives reward 1 on the transition to B, but the leader receives no positive reward anywhere along the cycle.

The leader therefore faces a strategic trade-off in its choice of p. Decreasing p changes the follower's incentives at A, but reducing it too far causes the follower to abandon action a at S and switch to the $S \to B \to S$ cycle, eliminating the leader's positive reward. The leader is thus encouraged to reduce p while keeping it sufficiently large for A to remain attractive to the follower. This creates a relatively narrow region in which changes to the follower response affect the leader's return, making the environment useful for testing the proposed pessimistic response model under limited response ambiguity.



(a) Limited response ambiguity game.

(b) Response-sensitive game.

Figure 3: **State Transition Diagrams.** Nodes represent states and edges represent transitions. Labels on edges indicate the respective leader and follower actions $\mathcal{A}_L \mid \mathcal{A}_F$ and their rewards $r_L \mid r_F$. The absence of shaded labels implies zero rewards. "$*$" represents a wildcard.

In the response-sensitive environment 3b, only the leader's action at state S determines which route is taken. Choosing action 0 leads to state A, while action 1 leads to state B; the follower action at S has no effect on this transition. From either A or B, the process returns to S after the follower responds. The leader therefore effectively chooses between the cycles $S \to A \to S$ and $S \to B \to S$.

Along the safe route $S \to A \to S$, the follower strictly prefers action 0, which yields follower reward 1 and leader reward 0.7. The alternative follower actions are substantially less attractive, with follower rewards 0.2 and 0, while the corresponding leader reward remains positive at 0.6. Hence, deviations from the nominal follower response at A have only a limited effect on the leader.

The risky route $S \to B \to S$ is instead highly response-sensitive. The follower's nominal best response is again action 0, yielding reward 1 to both agents. Under this response, the risky route is preferable to the leader over the safe route. However, action 1 is nearly as attractive to the follower, yielding follower reward 0.9, while reversing the leader reward from 1 to -1. Thus, a decrease of only 0.1 in follower utility corresponds to a decrease of 2 in leader utility. Action 2, by comparison, gives zero reward to both agents.

Consequently, a leader optimized solely against the nominal follower best response is encouraged to select action 1 at S and repeatedly induce the cycle $S \to B \to S$. A pessimistic response model instead accounts for the near-best follower action at B, making this route considerably less attractive. The alternative cycle $S \to A \to S$ sacrifices some nominal leader reward but is substantially less sensitive to uncertainty in the follower response. The environment therefore creates a direct trade-off between nominal performance and robustness to near-optimal follower behavior.

The two benchmarks differ in how strongly ambiguity in the follower's response affects the leader. In the limited-response-ambiguity game, alternative follower responses can alter the resulting transition and the follower's own reward, but their effect on the leader's reward is comparatively limited. In particular, several distinct responses from state A yield the same leader reward despite substantially different follower rewards. The response-sensitive game instead contains a deliberately risky route through state B. Thus, small ambiguity in the follower's response can have a disproportionately large effect on the leader in the second environment.

C.2 EXACT δ-RSE OF THE RESPONSE-SENSITIVE BENCHMARK

The response-sensitive benchmark is sufficiently small to compute its stochastic δ-robust Stackelberg solution analytically. This provides a ground-truth comparison for the behavior induced by our pessimistic response model. Following the δ-suboptimality robustness criterion Gan et al. (2025), we consider a stochastic extension in which the follower may use any stochastic policy whose discounted return is within δ of optimal.

The episode horizon is $H = 150$ and $\gamma = 0.99$. Since rewards are obtained at the route states A and B, which are visited every second time step, let

$$C_H = \sum_{n=0}^{74} \gamma^{2n+1} = \frac{\gamma(1-\gamma^{150})}{1-\gamma^2} \approx 38.7318 \tag{26}$$

Let

$$x = \pi_L(1 \mid S)$$

denote the probability with which the leader selects the route through B. Hence, $x = 0$ forces the cycle $S \to A \to S$, whereas $x = 1$ forces $S \to B \to S$. Under the nominal follower best response, action 0 is optimal at both A and B.

At A, action 0 yields $(r_L, r_F) = (0.7, 1)$, while action 1 yields $(0.6, 0.2)$. Action 2 need not be considered by a pessimistic follower: it gives the same leader reward as action 1 but incurs a larger follower loss. Similarly, at B, action 0 yields $(1, 1)$ and action 1 yields $(-1, 0.9)$. Action 2 is again dominated for the purpose of minimizing the leader return, since it yields a larger follower loss while being less damaging to the leader.

Let q_A and q_B denote the probabilities of taking action 1 at A and B, respectively. Relative to the nominal follower response, the resulting loss in follower return is

$$\Delta_F = C_H \left[0.8(1-x)q_A + 0.1xq_B\right] \tag{27}$$

The corresponding leader return is

$$J_L(x, q_A, q_B) = C_H \left[(1-x)(0.7 - 0.1q_A) + x(1-2q_B)\right] \tag{28}$$

The pessimistic follower minimizes equation 28 subject to $\Delta_F \le \delta$. A deviation at B decreases the leader return by 2 for only 0.1 follower regret, whereas a deviation at A decreases the leader return by 0.1 for 0.8 follower regret. The corresponding leader-loss-to-follower-regret ratios are therefore

$$\frac{2}{0.1} = 20 \qquad \text{and} \qquad \frac{0.1}{0.8} = \frac{1}{8},$$

respectively. The pessimistic response therefore allocates its admissible suboptimality to deviations at B before deviations at A.

For the deterministic safe route $x = 0$, the robust leader value is

$$\underline{J}_L^A(\delta) = \begin{cases} 0.7C_H - \dfrac{\delta}{8}, & 0 \le \delta \le 0.8C_H, \\ 0.6C_H, & \delta > 0.8C_H \end{cases} \tag{29}$$

In particular, its nominal value is

$$\underline{J}_L^A(0) = 0.7C_H = \sum_{n=0}^{74} 0.99^{2n+1}(0.7) \approx 27.1123.$$

For the risky route $x = 1$, the follower can instead mix toward the near-optimal action 1 at B. Its robust value is

$$\underline{J}_L^B(\delta) = \begin{cases} C_H - 20\delta, & 0 \le \delta \le 0.1C_H, \\ -C_H, & \delta > 0.1C_H \end{cases} \tag{30}$$

Thus, although the risky route has the larger nominal value, $\underline{J}_L^B(0) = C_H \approx 38.7318$, its pessimistic value decreases much more rapidly as the follower response set expands. For completeness, allowing

the leader to randomize between the two routes does not yield a better robust solution. If $\delta \leq 0.1C_H x$, the entire admissible follower deviation can be allocated to state B, giving

$$\underline{J}_L(x;\delta) = C_H(0.7 + 0.3x) - 20\delta, \tag{31}$$

which is increasing in x. If instead

$$0.1C_H x < \delta \leq 0.1C_H x + 0.8C_H(1-x),$$

the response at B is fully shifted to action 1 and the remaining ambiguity is allocated to A, yielding

$$\underline{J}_L(x;\delta) = 0.7C_H - 1.6875C_H x - \frac{\delta}{8}, \tag{32}$$

which is decreasing in x. Once both deviations are fully admissible, the value is

$$\underline{J}_L(x;\delta) = 0.6C_H - 1.6C_H x, \tag{33}$$

which is again decreasing in x. Hence a robustly optimal leader policy is always attained at one of the two deterministic routes, $x = 0$ or $x = 1$.

The transition between them is therefore obtained by equating equation 29 and equation 30 in their first regimes:

$$0.7C_H - \frac{\delta^\star}{8} = C_H - 20\delta^\star \tag{34}$$

This gives

$$\delta^\star = \frac{0.3C_H}{20 - \frac{1}{8}} \approx 0.5846 \tag{35}$$

Consequently, for $\delta < \delta^\star$, the exact δ-RSE selects the nominally superior but response-sensitive cycle $S \to B \to S$. For $\delta > \delta^\star$, it instead selects the safer cycle $S \to A \to S$. At $\delta = \delta^\star$, the two deterministic routes have the same robust leader value. This benchmark therefore exhibits explicitly the behavior targeted by our method: increasing uncertainty over near-optimal follower responses eventually makes the leader sacrifice nominal performance in favor of a route whose value is substantially less sensitive to follower deviations.

C.3 Additional Robustness Evaluation and α Ablation

Matched α sweep. For the response-sensitive benchmark, we train RBC-HG with $\alpha \in \{0, 0.01, 0.05, 0.1, 0.25, 0.5, 1\}$ using 10 matched random seeds. All remaining environment, follower, leader, replay, optimization, and sensitivity-estimation settings are held fixed across the sweep. In particular, all runs use the action-conditioned follower Bellman operator and the exact tabular value sensitivities described in Appendix B.5. Before evaluation, we verify that the resulting snapshot grid contains exactly one trained model for every (α, seed) pair and that no configuration field other than α and the random seed differs across runs.

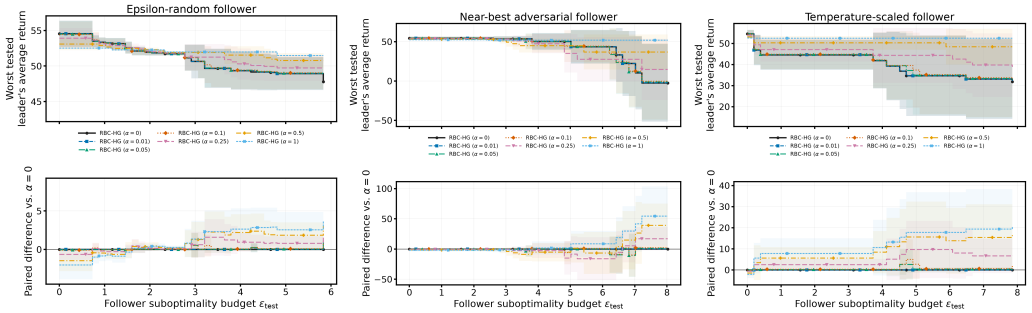
For completeness, Figure 4 reports the same seven-value α sweep using the undiscounted finite-horizon leader return. It exhibits the same qualitative ordering as the discounted results reported in the main text, showing that the observed robustness pattern is not specific to discounting the evaluation return.

Nominal follower used for evaluation. For each frozen trained leader policy π_L, we recompute the follower's entropy-regularized best response to convergence using the action-conditioned soft Bellman equations. Denoting this converged response by

$$p(b \mid s, a) = \pi_F^*(b \mid s, a),$$

all reported response perturbations are constructed relative to this re-equilibrated nominal follower rather than to a potentially stale follower stored in the training snapshot.

The nominal leader action-value table used to construct adversarial evaluation responses is also computed exactly from the known tabular model under the frozen pair (π_L, p). Learned critics are therefore not used to define the post-training perturbation families reported below.



(a) Epsilon-random follower (b) Near-best adversarial follower (c) Temperature-scaled follower

Figure 4: Undiscounted-return counterpart of Figure 2. Each panel reports the worst exact finite-horizon leader return among the tested follower responses whose follower suboptimality does not exceed ϵ_{test}. The qualitative dependence on the training pessimism strength α is consistent with the discounted evaluation reported in the main text. Shading denotes variation over 10 matched random seeds.

Perturbation families. We evaluate three follower-response shifts.

For the epsilon-random response,

$$q_\epsilon(b \mid s, a) = (1 - \epsilon)p(b \mid s, a) + \epsilon \frac{1}{|\mathcal{A}_F|}$$

For the temperature-scaled response, we retain the converged follower action values Q_F^* but change the evaluation temperature:

$$q_T(b \mid s, a) = \frac{\exp(Q_F^*(s, a, b)/T)}{\sum_{\bar{b}} \exp(Q_F^*(s, a, \bar{b})/T)}$$

For the near-best adversarial response, define

$$\mathcal{A}_\delta(s, a) = \left\{ b : Q_F^*(s, a, b) \geq \max_{\bar{b}} Q_F^*(s, a, \bar{b}) - \delta \right\}$$

Among these approximately optimal actions we select

$$b^-(s, a) \in \arg \min_{b \in \mathcal{A}_\delta(s,a)} Q_L^{\pi_L, p}(s, a, b).$$

To avoid a fully deterministic response, the experiments place probability $1 - \xi$ on b^- and distribute the remaining mass $\xi = 0.05$ uniformly over $\mathcal{A}_\delta(s, a)$.

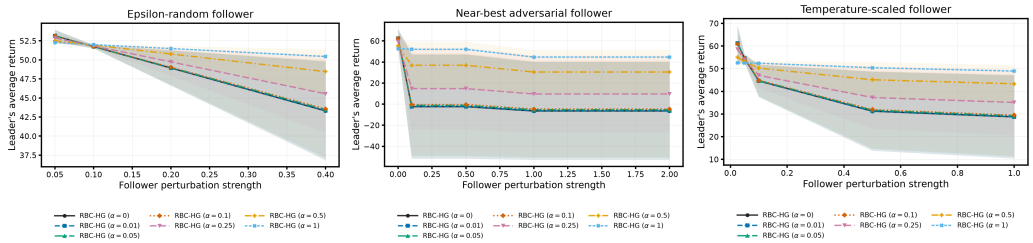
The evaluation implementation additionally supports an exponential pessimistic perturbation. On the present benchmark this family induces an effectively degenerate range of follower suboptimality, and we therefore do not include it in the reported robustness comparisons.

Alternative views of the perturbations. Figures 5 and 7 provide two complementary views of the same post-training perturbations. The first plots leader performance directly against the native strength parameter of each perturbation family. These parameters have different meanings across perturbations and should therefore only be interpreted within each panel. The second instead indexes the budgeted robustness curves by the discounted test-occupancy KL divergence from the nominal follower response.

Follower-suboptimality budget. Since the perturbation parameters ϵ, T, and δ have different meanings, we compare them using their induced loss in the follower's own entropy-regularized objective:

$$\epsilon_{\text{test}}(q) = J_F^\lambda(\pi_L, p) - J_F^\lambda(\pi_L, q)$$

The two quantities are evaluated exactly from the tabular transition and reward model.



(a) Epsilon-random follower (b) Near-best adversarial follower (c) Temperature-scaled follower

Figure 5: Leader performance under the three follower perturbation families as a function of their native perturbation strength. The horizontal parameter has a different interpretation for epsilon-random, near-best adversarial, and temperature-scaled responses, so comparisons of perturbation magnitude should be made within each panel rather than across panels. Curves show the seven training values of α, averaged over 10 matched random seeds; shading denotes variation across seeds.

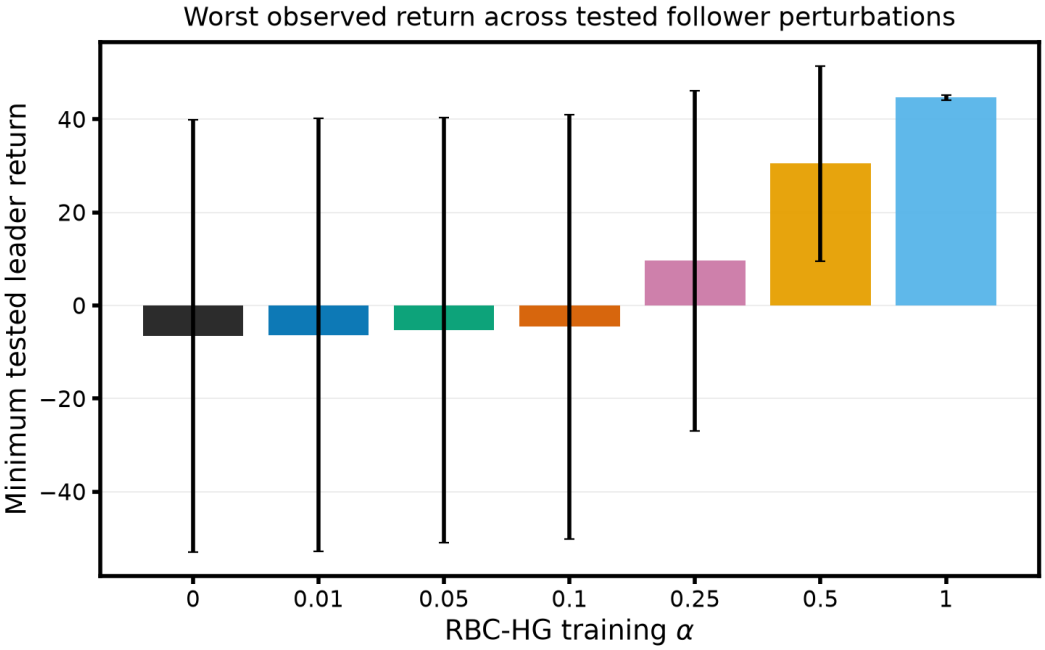


Figure 6: Worst observed leader return across the finite post-training perturbation suite on the response-sensitive benchmark. For each training value α and random seed, we take the minimum leader return over all tested epsilon-random, temperature-scaled, and near-best adversarial follower responses and perturbation strengths. Bars show the mean of these minima over 10 matched seeds and error bars show one standard deviation. This is a finite stress-test summary rather than a global worst-case follower value or a fixed-budget robustness measure.

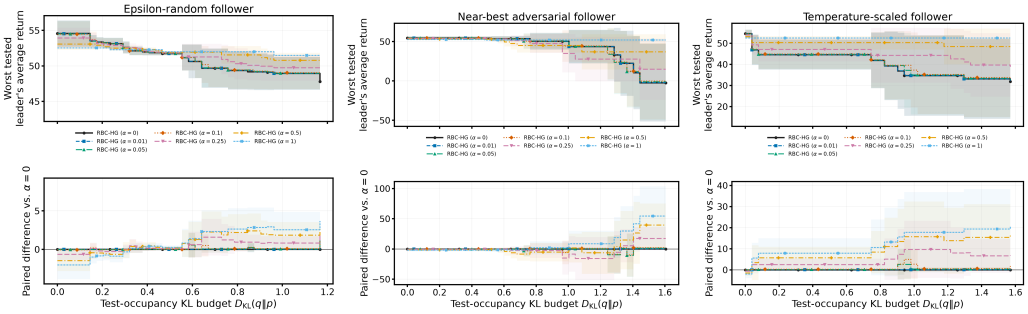
We additionally record the discounted test-occupancy KL divergence

$$D_{\mathrm{KL}}^{\mathrm{test}}(q\|p) = \sum_s d_q^{\gamma_F}(s) \sum_a \pi_L(a \mid s) D_{\mathrm{KL}}(q(\cdot \mid s,a)\|p(\cdot \mid s,a)),$$

where $d_q^{\gamma_F}$ is the normalized γ_F-discounted state occupancy induced by (π_L, q). For an exactly converged entropy-regularized best response, these quantities satisfy

$$\epsilon_{\mathrm{test}}(q) = \frac{\lambda}{1-\gamma_F} D_{\mathrm{KL}}^{\mathrm{test}}(q\|p),$$

which we also use as an evaluation consistency check.



(a) Epsilon-random follower (b) Near-best adversarial follower (c) Temperature-scaled follower

Figure 7: KL-budget view of the α ablation on the response-sensitive benchmark. Each panel reports the worst leader return among the tested follower responses whose discounted test-occupancy divergence from the nominal response does not exceed the displayed $D_{\mathrm{KL}}^{\mathrm{test}}(q\|p)$ budget. Because $\epsilon_{\mathrm{test}} = \lambda D_{\mathrm{KL}}^{\mathrm{test}}/(1-\gamma_F)$ for the converged entropy-regularized best response, this is an alternative KL-based parameterization of the follower-suboptimality curves in the main text. Shading denotes variation over 10 matched random seeds.

Accordingly, Figure 7 reports the same budgeted robustness evaluation using $D_{\mathrm{KL}}^{\mathrm{test}}(q\|p)$ rather than ϵ_{test} on the horizontal axis. Because the nominal follower is recomputed to convergence, the two budget variables are linked by the identity above; the KL plot should therefore be interpreted as an alternative, KL-aligned parameterization of the same robustness result rather than as an independent experiment.

Exact leader evaluation. Because the benchmark is finite and its transition model is available, no Monte Carlo rollouts are required for the robustness curves. For each frozen leader and test follower, the state distribution is propagated exactly through the induced transition matrix. We report both the finite-horizon leader return and its discounted counterpart,

$$J_{L,H}(q) = \mathbb{E}\left[\sum_{t=0}^{H-1} r_L(s_t, a_t, b_t)\right], \qquad J_{L,H}^{\gamma}(q) = \mathbb{E}\left[\sum_{t=0}^{H-1} \gamma_L^t r_L(s_t, a_t, b_t)\right],$$

with $H = 150$.

Budgeted robustness curves. For a perturbation family $\mathcal{Q}_{\mathrm{test}}$ and a follower suboptimality budget ϵ, the plotted quantity is the finite-test-set envelope

$$\widehat{R}_{\mathrm{test}}(\epsilon) = \min_{q \in \mathcal{Q}_{\mathrm{test}}: \epsilon_{\mathrm{test}}(q) \le \epsilon} J_{L,H}^{\gamma}(q).$$

The budget axis is restricted to the common range covered by every (α, seed) run; we do not interpolate between untested follower policies or extrapolate beyond the tested range. Paired-difference plots subtract the corresponding value obtained by the matched $\alpha = 0$ run with the same random seed.

Response induced during training. For diagnostic purposes, we also evaluate the response

$$\pi_F^{\alpha}(b \mid s, a) \propto p(b \mid s, a) \exp(-\alpha Q_L^{\pi_L, p}(s, a, b))$$

using the exact tabular nominal leader value, and record both its test-occupancy KL divergence and its follower optimality gap. A second diagnostic constructs the same response from the saved nominal follower and learned nominal critic. Their agreement provides a check that the observed response displacement is not solely an artifact of critic approximation.

C.4 Nominal Performance

Effect of the training pessimism. Figure 8 summarizes the trade-off induced by the training parameter α. Increasing α produces only a modest reduction in nominal leader performance: the mean

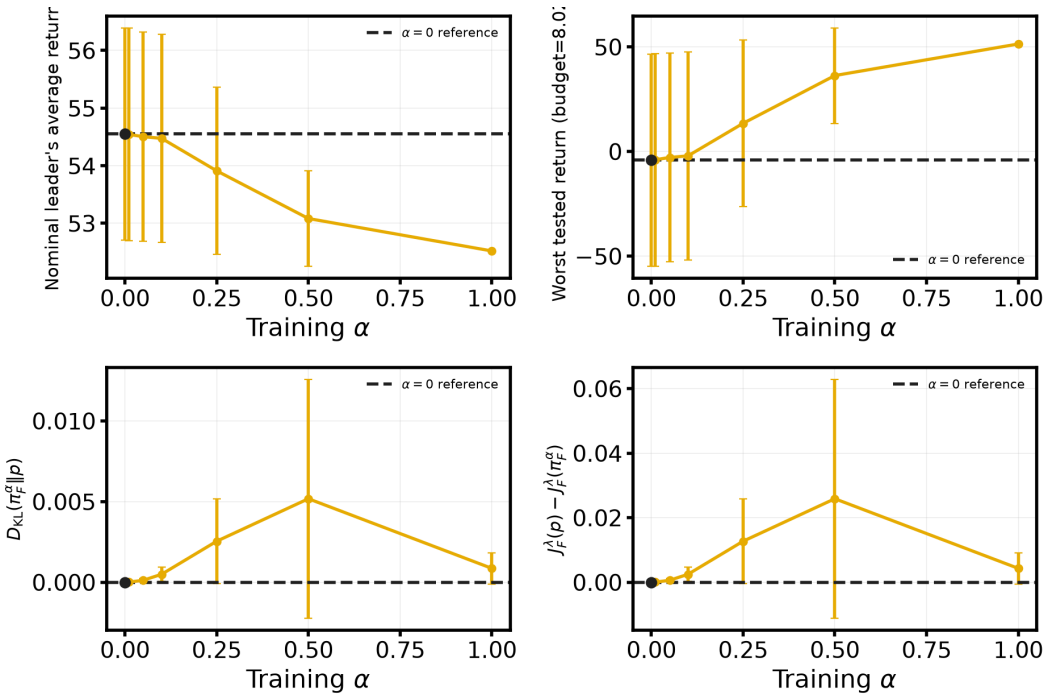


Figure 8: Effect of the training pessimism parameter α on the response-sensitive benchmark.
Top left: nominal leader return under the recomputed entropy-regularized follower best response.
Top right: worst tested leader return at the reported follower-suboptimality budget.
Bottom left: test-occupancy KL divergence between the exact pessimistic response π_F^α and the nominal response p.
Bottom right: entropy-regularized follower optimality gap induced by π_F^α. Increasing α yields a modest nominal-performance cost but a substantial robustness gain, while the realized response displacement is non-monotone after retraining the leader.
Points show means over 10 matched seeds and error bars indicate variation across seeds.

nominal return decreases gradually from approximately 54.5 at $\alpha = 0$ to 52.5 at $\alpha = 1$. In contrast, robustness improves substantially. At the largest reported follower-suboptimality budget, the worst tested leader return increases sharply once α reaches 0.25, with the strongest performance obtained for $\alpha = 1$.

The realized displacement of the pessimistic response is notably non-monotone after re-optimizing the leader. Both $D_{\mathrm{KL}}(\pi_F^\alpha \| p)$ and the follower optimality gap increase up to $\alpha = 0.5$ and then decrease at $\alpha = 1$. Hence, the largest robustness gain does not coincide with the largest realized perturbation of the nominal follower. This indicates that the effect of α cannot be understood solely as increasing the distance between π_F^α and p: changing α also changes the leader policy obtained after training.

Aggregate stress-test summary. To summarize performance across the complete perturbation suite, we also compute, for each trained leader and random seed,

$$\widehat{R}_{\mathrm{all}}(\pi_L^\alpha) = \min_{q \in \mathcal{Q}_{\mathrm{eval}}} J_{L,H}(\pi_L^\alpha, q),$$

where $\mathcal{Q}_{\mathrm{eval}}$ is the union of all tested epsilon-random, temperature-scaled, and near-best adversarial responses across the perturbation strengths reported above. Figure 6 reports the mean of this minimum over the 10 matched seeds. The statistic is intended only as a compact summary of the finite stress-test suite; it does not solve the global worst-response problem and does not constrain the different perturbation families to a common follower-suboptimality budget.

The minimum tested return changes little for $\alpha \leq 0.1$, begins to improve at $\alpha = 0.25$, and increases substantially for $\alpha \in \{0.5, 1\}$. The large variation for intermediate values also shows that the

robustness improvement is initially seed-dependent, whereas the $\alpha = 1$ policies are considerably more consistent across the tested perturbations.

Algorithm 1 Robust BC-HG (RBC-HG), tabular implementation

Require: Leader policy $\pi_L(\cdot \mid s;\theta_L)$, pessimism strength α, follower temperature λ, discount γ_L, horizon H, importance-weight clip $w_{\max}$, replay buffer $\mathcal{D}$

Initialize nominal leader critic $\widehat{Q}_L^v$, modified-response critic $\widehat{Q}_L^\beta$, and their target networks
for each training iteration **do**
  Solve the follower lower-level problem by Soft Q-Iteration until convergence, obtaining the nominal response π_F^*
  Collect trajectories under $v(a,b \mid s) = \pi_L(a \mid s)\pi_F^*(b \mid s,a)$ and store them in the replay buffer $\mathcal{D}$
  for N_Q nominal-critic updates **do**
    Sample transitions (s,a,b,r_L,s',d) from $\mathcal{D}$
    Form the nominal Bellman target

$$y^v = r_L + \gamma_L(1-d)\mathbb{E}_{\substack{a'\sim\pi_L^-(\cdot|s')\\ b'\sim\pi_F^*(\cdot|s',a')}}\left[\widehat{Q}_L^{v,-}(s',a',b')\right]$$

    Update $\widehat{Q}_L^v$ by minimizing $(\widehat{Q}_L^v(s,a,b) - y^v)^2$
  end for
  for N_π leader updates **do**
    Sample a discounted minibatch $\mathcal{B} \subset \mathcal{D}$ with sampling weights proportional to γ_L^t
    Construct the pessimistic response

$$\pi_F^\alpha(b \mid s,a) \propto \pi_F^*(b \mid s,a)\exp\Big(-\alpha\ \mathrm{sg}\left[\widehat{Q}_L^v(s,a,b)\right]\Big)$$

    Form the modified-response Bellman target

$$y^\beta = r_L + \gamma_L(1-d)\mathbb{E}_{\substack{a'\sim\pi_L^-(\cdot|s')\\ b'\sim\pi_F^\alpha(\cdot|s',a')}}\left[\widehat{Q}_L^{\beta,-}(s',a',b')\right]$$

    Update $\widehat{Q}_L^\beta$ using $(\widehat{Q}_L^\beta(s,a,b) - y^\beta)^2$ and Polyak-update its target network
    Compute the clipped follower-action ratio

$$\bar{w}(s,a,b) = \min\left\{\frac{\pi_F^\alpha(b \mid s,a)}{\pi_F^*(b \mid s,a)}, w_{\max}\right\}$$

    Compute finite-horizon occupancies d_H^v, d_H^β and

$$\zeta_H(s) = \frac{d_H^\beta(s)}{d_H^v(s)}$$

    Compute the exact tabular value sensitivities

$$G_F(s,a,b) = \nabla_L Q_F^*(s,a,b), \qquad G_L(s,a,b) = \nabla_L Q_L^v(s,a,b)$$

    by differentiating the corresponding Bellman fixed points
    Compute the modified-response Benefit

$$B_L^\beta(s,a,b) = \widehat{Q}_L^\beta(s,a,b) - \mathbb{E}_{b'\sim\pi_F^\alpha(\cdot|s,a)}\left[\widehat{Q}_L^\beta(s,a,b')\right]$$

    Construct S_α such that

$$\nabla_L S_\alpha(s,a,b) = \frac{1}{\lambda}G_F(s,a,b) - \alpha G_L(s,a,b)$$

    Form the leader surrogate

$$\widehat{\mathcal{L}}_L = -\frac{1}{|\mathcal{B}|}\sum_{(s,a,b)\in\mathcal{B}} \mathrm{sg}[\zeta_H(s)\bar{w}(s,a,b)]\left[\mathrm{sg}\Big[\widehat{Q}_L^\beta(s,a,b)\Big]\log\pi_L(a \mid s) + \mathrm{sg}\Big[B_L^\beta(s,a,b)\Big]\,S_\alpha(s,a,b)\right]$$

    Update θ_L by descending $\nabla_{\theta_L}\widehat{\mathcal{L}}_L$
  end for
  Update the nominal target networks
end for