LEARNING FEATURE INTERFACES TO STEER PRESCRIBED CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many online decision systems are modular: an upstream component controls the features exposed to a downstream learner but cannot change the learner's reward or learning rule. While the learner optimizes its prescribed reward, the upstream component may optimize a different system-level objective. We study feature-interface optimization in repeated linear contextual-bandit deployments. Before each deployment, a leader applies a positive coordinate-wise rescaling to the raw features exposed to a prescribed contextual-bandit learner, which then learns from bandit feedback. Even this simple intervention leaves the follower's reward task unchanged while altering its finite-sample estimates, exploration, and collected feedback. As a result, the leader objective can depend nontrivially on the feature interface. We propose CILOT, which optimizes the interface using an unbiased trajectory-gradient estimator that differentiates through the follower's feedback-dependent online updates using only selected-action feedback. A single follower calibration guarantees $\widetilde{O}(d\sqrt{H})$ regret uniformly over all admissible interfaces, while the leader optimization admits an $\widetilde{O}(M^{-1/2})$ projected-stationarity guarantee. Experiments on synthetic and real-world datasets demonstrate the benefit of optimizing the feature interface.

1 INTRODUCTION

Modern online decision systems are often modular. A recommender, model router, or tool selector may learn from bandit feedback, while an upstream platform controls the features exposed to the learner and is responsible for system-level considerations such as serving cost, latency, or resource usage (Sculley et al., 2015; Baylor et al., 2017). In many deployments, the downstream contextual-bandit learner is prescribed in advance: the upstream component cannot modify its reward, feedback protocol, or update rule. The two components may nevertheless optimize different objectives. While the downstream learner seeks to maximize its bandit reward, the upstream component may care about a trajectory-level utility, such as task quality net of serving cost. This raises a natural question: can the feature interface be optimized for the upstream objective without changing the downstream learner's reward task or algorithm, while retaining its learning guarantees?

We study this question in repeated, resettable linear contextual-bandit deployments. We call the upstream component the *leader* and the downstream learner the *follower*. Before each deployment, the leader applies a positive coordinate-wise rescaling $v \in [\alpha, 1]^d$, with $\alpha > 0$, to the raw features exposed to the follower. The follower is prescribed in advance and uses ridge-based optimistic scores from linear contextual bandits (Li et al., 2010; Abbasi-Yadkori et al., 2011), together with a fixed softmax action-selection rule, learning online from selected-action bandit feedback. The rescaling is invertible and therefore preserves the conditional rewards and reward-optimal action set, but it can still change the follower's finite-sample behavior. Along any realized history, ridge regression on the exposed features is equivalent to ridge regression on the raw features with diagonal regularization $\lambda \operatorname{Diag}(v_1^{-2}, \ldots, v_d^{-2})$. Even this simple, invertible intervention can therefore alter the follower's estimation and exploration behavior, giving the leader a nontrivial way to influence its trajectory-level objective. Unlike feature or representation learning for linear bandits (Yang et al., 2020; Papini et al., 2021; Tirinzoni et al., 2022; 2023), we keep the raw feature map, reward task, and follower algorithm fixed and optimize only the feature interface.

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.

Although the leader modifies only the feature interface, its effect propagates through the follower's online learning dynamics. Changes in early action probabilities alter which feedback is collected, which in turn changes the regression state governing subsequent decisions. Consequently, differentiating the current action probabilities while holding the state induced by earlier feedback fixed can miss part of the gradient of the leader objective.

In this work, we develop *Competence-preserving Interface Learning from Online Trajectories* (CILOT) to optimize the feature interface across repeated deployments. CILOT differentiates the follower's conditional action probabilities through the regression updates induced by earlier feedback and uses the resulting score terms to construct a likelihood-ratio trajectory gradient, building on score-function and hypergradient methods (Schulman et al., 2015; Maclaurin et al., 2015; Franceschi et al., 2018; Xu et al., 2018). The resulting estimator is unbiased and requires only selected-action feedback. Across deployments, CILOT uses averaged trajectory gradients in projected stochastic updates of the interface.

We further establish theoretical guarantees for both levels of the system. A single follower calibration fixed before interface optimization yields $\widetilde{O}(d\sqrt{H})$ expected regret uniformly over the admissible interface class. For the leader, the outer optimization achieves an $\widetilde{O}(M^{-1/2})$ bound on the expected squared norm of the projected gradient mapping. Experiments on synthetic and real-world data show that optimizing the feature interface can improve the leader objective while retaining sublinear follower regret.

Contributions. The major contributions of this work are highlighed as follows.

- We formulate feature-interface optimization for a prescribed contextual-bandit learner, where the leader can modify only the feature interface while the follower's reward task and learning algorithm remain fixed.
- We develop CILOT, an unbiased trajectory-gradient method that differentiates through the follower's feedback-dependent online updates using only selected-action feedback.
- We establish compatible theoretical guarantees for both levels of the system: the follower achieves $\widetilde{O}(d\sqrt{H})$ expected regret uniformly over the admissible interface class, while the leader optimization achieves an $\widetilde{O}(M^{-1/2})$ rate for the expected squared norm of the projected gradient mapping.
- Experiments on synthetic and real-world data show that optimizing the feature interface improves the leader objective while maintaining strong follower performance.

2 Related Work

Representation learning and adaptation in contextual bandits. Our follower builds on optimistic linear contextual bandits (Li et al., 2010; Abbasi-Yadkori et al., 2011). Prior work has studied feature selection, learned representations, shared structure, and adaptive experimental design as ways to improve statistical efficiency in contextual bandits (Yang et al., 2020; Papini et al., 2021; Tirinzoni et al., 2022; 2023; Fiez et al., 2019; Wagenmaker et al., 2021; Cella et al., 2020). Meta-bandit methods use experience across tasks to adapt priors, hypothesis classes, or learning strategies (Sharaf & Daumé III, 2019; Kveton et al., 2021; Kassraie et al., 2022). In our setting, the raw feature map, reward task, and follower algorithm remain fixed. The leader modifies only an invertible feature interface, which changes the follower's finite-sample estimation and exploration behavior without changing the reward task.

Differentiating through adaptive learning processes. Our method draws on stochastic computation graphs, differentiable learning dynamics, and bilevel optimization (Schulman et al., 2015; Maclaurin et al., 2015; Franceschi et al., 2018; Xu et al., 2018). Closely related work has optimized differentiable bandit policies across task instances (Boutilier et al., 2020), while contextual bilevel reinforcement learning optimizes an upper-level objective through the policies learned by followers in lower-level decision problems (Thoma et al., 2024). Our problem has a different intervention: the leader does not modify the follower's reward problem or learning algorithm, but only the feature interface presented to a prescribed online learner. Within a deployment, the follower's actions determine the bandit feedback used in subsequent regression updates, so the interface affects later decisions

through the feedback it induces. This creates a closed-loop dependence through the follower's evolving statistical state that is not captured by formulations that optimize a fixed policy or modify the lower-level objective. Computing the leader gradient therefore requires differentiating through this within-deployment state.

System objectives and leader–follower bandits. Related work has considered bandit settings in which a system-level objective differs from the reward optimized by an individual learner. Constrained and utility-aware formulations incorporate additional objectives or constraints directly into the learner's decision problem (Agrawal et al., 2016). Stackelberg and principal–agent bandits instead study strategic interactions in which a leader influences an agent through incentives, commitments, or transfers (Bai et al., 2021; Harris et al., 2024; Scheid et al., 2024; Balcan et al., 2025; Lin & Chen, 2025). In our setting, the follower remains a prescribed online learner with a fixed reward and learning rule. The leader influences its behavior only through the feature interface selected before each deployment.

3 Problem Formulation

Repeated deployments. We consider repeated, resettable contextual-bandit deployments with horizon H. Let $\mathcal{X}$ be the context space and $\mathcal{A} := [K] := \{1, \dots, K\}$ the finite action set. Each $a \in \mathcal{A}$ indexes one available action. The raw feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and its coordinate system are fixed in advance. We refer to the upstream component that controls the features exposed to the learner as the *leader*, and the downstream contextual-bandit learner as the *follower*.

Before each deployment, the leader chooses $v \in \mathcal{V} := [\alpha, 1]^d$, where $0 < \alpha \leq 1$. We call v the *feature interface*: it is a coordinate-wise scaling layer placed between the raw feature map and the follower. In particular, coordinate i is multiplied by v_i, so the follower observes

$$z_v(x, a) := v \odot \phi(x, a), \qquad (x, a) \in \mathcal{X} \times \mathcal{A}.$$

Thus the leader can attenuate individual feature coordinates but cannot replace the raw feature map or introduce new features. The lower bound $v_i \geq \alpha > 0$ ensures that every admissible interface is invertible. We use this restricted interface class to separate the effect of feature exposure from representation learning or model redesign. Even this simple intervention provides a nontrivial degree of control over the follower's finite-sample behavior, allowing the leader to optimize its trajectory-level objective while leaving the underlying reward task unchanged.

The interface is chosen before the deployment-specific reward parameter and contexts are realized. After v is fixed, a fresh reward parameter $\theta^\star \sim \mathcal{P}_\Theta$ and an exogenous context sequence $(x_1, \dots, x_H) \sim \mathcal{D}_{\theta^\star}^{(H)}$ are drawn. The follower is initialized from the same state at the beginning of every deployment. At round t, the follower observes x_t and the exposed features $\{z_v(x_t, a) : a \in \mathcal{A}\}$, selects an action $a_t \in \mathcal{A}$ according to its prescribed decision rule, and observes the corresponding noisy bandit reward

$$Y_t = \theta^{\star\top} \phi(x_t, a_t) + \varepsilon_t, \tag{1}$$

where ε_t denotes the reward noise. We write $h_t := (x_1, a_1, Y_1, \dots, x_t, a_t, Y_t), h_0 := \varnothing$, for the interaction history. Let $\mathcal{I}_t$ denote the sigma-field generated by (h_{t-1}, x_t), i.e., the information available immediately before a_t is sampled.

Across deployments, the leader may choose v adaptively using information from completed deployments. The follower, however, is reset before each new deployment.

Assumption 1 (Statistical regularity)**.** *For all $(x, a) \in \mathcal{X} \times \mathcal{A}$, $\|\phi(x, a)\|_2 \leq 1$, and $\|\theta^\star\|_2 \leq S$ almost surely. Conditional on $\mathcal{I}_t$, a_t, and $\theta^\star$, the noise ε_t is zero-mean and σ_Y-sub-Gaussian:*

$$\mathbb{E}[\varepsilon_t \mid \mathcal{I}_t, a_t, \theta^\star] = 0, \qquad \mathbb{E}[\exp(s\varepsilon_t) \mid \mathcal{I}_t, a_t, \theta^\star] \leq \exp\left(\frac{s^2\sigma_Y^2}{2}\right)$$

for every $s \in \mathbb{R}$. For the sensitivity and outer-optimization results, we additionally assume $|Y_t| \leq R_Y$ almost surely.

The follower is trained on Y_t, whereas the leader evaluates the deployment using a possibly different utility $U_t := u(x_t, a_t, Y_t) \in [0, 1]$. For example, U_t may combine task quality with serving cost or resource usage.

Assumption 2 (Deployment exogeneity). *For every admissible interface $v \in \mathcal{V}$:*

(i) Conditional on $\theta^\star$, the context sequence $(x_1, \dots, x_H)$ is drawn from $\mathcal{D}_{\theta^\star}^{(H)}$, whose distribution is independent of v and of the within-deployment interaction history.

(ii) For every round t, conditional on $\mathcal{I}_t$, a_t, and $\theta^\star$, the conditional distribution of Y_t does not depend on v.

The leader utility is evaluated by the fixed measurable function $u : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \to [0, 1]$ introduced above, and therefore depends on the interface only through the induced deployment trajectory.

Assumption 2 ensures that the interface affects the deployment distribution only through the follower's action-selection behavior.

Access model and reward invariance. The follower algorithm is prescribed before interface optimization. Its reward, feedback protocol, regularization, confidence and temperature schedules, action-selection rule, and online updates are fixed. The leader controls only the interface v and observes the feedback generated by the selected actions.

Despite changing the features seen by the follower, every admissible interface preserves the underlying reward task. For any admissible reward parameter θ and $v \in \mathcal{V}$, define $(\psi^\star_{\theta,v})_i := \frac{\theta_i}{v_i}, i \in [d]$. Then, for every $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$\langle \psi^\star_{\theta,v}, z_v(x, a) \rangle = \langle \theta, \phi(x, a) \rangle,$$

and $\|\psi^\star_{\theta,v}\|_2 \le S/\alpha$ whenever $\|\theta\|_2 \le S$. Hence the conditional mean reward of every action, and therefore the reward-optimal action set, is unchanged by the interface. The interface can nevertheless alter the follower's finite-sample estimation and exploration behavior.

Follower guarantee. For a fixed interface v and reward parameter θ, let P_v^θ denote the distribution of the resulting fresh deployment, including its contexts, actions, and observed rewards. Following standard linear contextual-bandit regret formulations (Li et al., 2010; Abbasi-Yadkori et al., 2011), we measure follower performance by the expected pseudo-regret

$$\mathrm{Reg}_H^L(v; \theta) = \mathbb{E}_{h_H \sim P_v^\theta}\left[\sum_{t=1}^{H}\left(\max_{a \in \mathcal{A}} \langle \theta, \phi(x_t, a) \rangle - \langle \theta, \phi(x_t, a_t) \rangle\right)\right].$$

Our requirement is that a single calibration of the prescribed follower, chosen independently of the interface and fixed before interface optimization, yields a sublinear regret bound uniformly over $v \in \mathcal{V}$. In particular, $\sup_{v \in \mathcal{V}} \mathrm{Reg}_H^L(v; \theta)$ should be sublinear in H for every admissible θ. This uniformity is what allows the leader to adapt v across completed deployments without recalibrating the follower.

Leader objective. The leader evaluates performance over the entire deployment rather than at a single round. To allow greater emphasis on earlier decisions while keeping the objective on the same scale across horizons, we use normalized geometric discounting. For $\gamma \in (0, 1)$, define

$$\omega_t^{\gamma,H} := \frac{(1-\gamma)\gamma^{t-1}}{1-\gamma^H}, \qquad R_{\gamma,H} := \sum_{t=1}^{H} \omega_t^{\gamma,H} U_t.$$

The weights satisfy $\sum_{t=1}^{H} \omega_t^{\gamma,H} = 1$, so $R_{\gamma,H} \in [0, 1]$ whenever $U_t \in [0, 1]$.

The leader chooses the interface to maximize its expected deployment utility:

$$\max_{v \in \mathcal{V}} J_{\gamma,H}(v), \qquad J_{\gamma,H}(v) := \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} \mathbb{E}_{h_H \sim P_v^\theta}\left[R_{\gamma,H}\right].$$

Here the inner expectation is over the randomness of a fresh deployment under interface v, including the contexts, follower actions, and observed rewards.

4 The CILOT Method

We develop CILOT to optimize the feature interface through the prescribed follower's online learning process. The central challenge is that the interface affects both the follower's decisions and the feedback used in subsequent updates. We first characterize this dependence through an equivalent regression problem in the raw feature coordinates. We then construct an unbiased trajectory gradient that accounts for the evolving learning state and the distribution of collected feedback. Finally, we use these gradients to optimize the interface across repeated deployments.

4.1 Interface Dependence of the Follower

For a fixed interface v, the follower runs a prescribed optimistic linear contextual-bandit learner (Abbasi-Yadkori et al., 2011) with softmax action selection. It initializes $V_1 = \lambda I_d, b_1 = 0$, and sets $\widehat{\psi}_t := V_t^{-1} b_t$. At round t, the follower receives the exposed features

$$z_t(a) := z_v(x_t, a) = v \odot \phi(x_t, a), \qquad a \in \mathcal{A}.$$

For each candidate action $a \in \mathcal{A}$, it computes the confidence radius $r_t(a) := \sqrt{z_t(a)^\top V_t^{-1} z_t(a)}$. Using this radius, the follower forms an optimistic score and samples an action from the corresponding softmax policy:

$$q_t(a) := \widehat{\psi}_t^\top z_t(a) + \beta_t r_t(a), \tag{2a}$$

$$\pi_t^v(a \mid x_t, h_{t-1}) := \frac{\exp(q_t(a)/\tau_t)}{\sum_{a' \in \mathcal{A}} \exp(q_t(a')/\tau_t)}. \tag{2b}$$

After sampling a_t and observing Y_t, the follower updates

$$V_{t+1} := V_t + z_t(a_t) z_t(a_t)^\top, \quad b_{t+1} := b_t + z_t(a_t) Y_t, \tag{3a}$$

and hence $\widehat{\psi}_{t+1} = V_{t+1}^{-1} b_{t+1}$. The schedules $\beta_t \geq 0$ and $\tau_t > 0$ are deterministic and fixed before interface optimization.

Raw-coordinate representation. Although the follower performs regression in the exposed coordinates, the same state can be represented in the raw feature space. Let $W(v) := \mathrm{Diag}(v)$. For a realized history, define

$$\phi_s := \phi(x_s, a_s), \qquad \Sigma_t := \sum_{s<t} \phi_s \phi_s^\top, \qquad c_t := \sum_{s<t} \phi_s Y_s,$$

and

$$\Lambda(v) := \lambda W(v)^{-2} = \lambda\, \mathrm{Diag}(v_1^{-2}, \ldots, v_d^{-2}), \qquad A_t(v) := \Sigma_t + \Lambda(v), \qquad \widehat{\theta}_t(v) := A_t(v)^{-1} c_t.$$

The following equivalence shows that these raw-coordinate quantities exactly recover the follower's prediction and confidence terms in the exposed coordinates.

Proposition 1 (Raw-coordinate equivalence)**.** *For every $v \in \mathcal{V}$, every realized history, and every candidate action $a \in \mathcal{A}$,*

$$\widehat{\psi}_t^\top z_t(a) = \widehat{\theta}_t(v)^\top \phi(x_t, a), \qquad z_t(a)^\top V_t^{-1} z_t(a) = \phi(x_t, a)^\top A_t(v)^{-1} \phi(x_t, a).$$

Hence the visible-coordinate and raw-coordinate representations produce the same optimistic scores and action probabilities.

Conditional on a realized history, the interface therefore enters the regression problem through the diagonal ridge matrix $\Lambda(v)$, whose entries satisfy $\lambda \leq \Lambda_{ii}(v) \leq \frac{\lambda}{\alpha^2}$. The reward task remains unchanged, while the estimates and confidence radius used by the follower depend on v.

For a candidate action a, define $p_t(a) := A_t(v)^{-1} \phi(x_t, a)$. When taking derivatives with respect to v, we condition on the realized history, so Σ_t and c_t are fixed. Differentiating the optimistic score gives

$$\frac{\partial q_t(a)}{\partial v_i} = 2\lambda v_i^{-3} p_{t,i}(a) \widehat{\theta}_{t,i}(v) + \frac{\beta_t \lambda v_i^{-3}}{r_t(a)} p_{t,i}(a)^2. \tag{4}$$

The first term comes from the prediction component and the second from the confidence bonus. When $r_t(a) = 0$, the second term is defined by continuity as zero. Equation (4) gives the derivative conditional on a fixed history. The dependence of the realized trajectory itself on v is captured by the likelihood-ratio terms introduced next.

4.2 Trajectory Gradients from Bandit Feedback

For the sampled action a_t, define the conditional policy score

$$\ell_t(v) := \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1}) = \frac{1}{\tau_t} \left(\nabla_v q_t(a_t) - \sum_{a \in \mathcal{A}} \pi_t^v(a \mid x_t, h_{t-1}) \nabla_v q_t(a) \right). \tag{5}$$

Here $\nabla_v q_t(a)$ differentiates through the regression state obtained from the preceding online updates.

Let $G_t^{\gamma,H} := \sum_{s=t}^{H} \omega_s^{\gamma,H} U_s$ denote the normalized discounted return-to-go. For a sampled deployment trajectory ζ, define $\widehat{g}(v;\zeta) := \sum_{t=1}^{H} G_t^{\gamma,H} \ell_t(v)$. The following result connects this trajectory quantity to the gradient of the leader objective.

Theorem 1 (Unbiased trajectory hypergradient)**.** *Suppose Assumptions 1 and 2 hold and $|Y_t| \le R_Y$ almost surely. For any deterministic follower schedules $\{\beta_t, \tau_t\}_{t=1}^{H}$ with $\tau_t > 0$,*

$$\nabla J_{\gamma,H}(v) = \mathbb{E}\left[\widehat{g}(v;\zeta)\right] = \mathbb{E}\left[\sum_{t=1}^{H} G_t^{\gamma,H} \ell_t(v)\right]. \tag{6}$$

Thus $\widehat{g}(v;\zeta)$ is an unbiased estimator of the leader gradient.

The conditional policy score must differentiate through the follower state produced by earlier regression updates. Stopping the derivative through this state omits part of $\nabla_v \log \pi_t^v$ and can therefore bias the trajectory-gradient estimator. Proposition 5 provides a two-round construction in which the stopped-state derivative misses a nonzero component of the population gradient.

Online implementation. The trajectory estimator can be accumulated as feedback arrives. Define the cumulative policy score $e_0 := 0$ and $e_t := e_{t-1} + \ell_t(v) = \sum_{s=1}^{t} \ell_s(v)$. Exchanging the order of summation gives

$$\widehat{g}(v;\zeta) = \sum_{t=1}^{H} \omega_t^{\gamma,H} U_t e_t. \tag{7}$$

Thus each observed utility contributes to the trajectory-gradient estimator through the policy scores of all preceding decisions, including the current one. This online form requires only one auxiliary d-dimensional vector e_t and avoids storing the returns-to-go. A predictable action-independent baseline can also reduce variance without changing the estimator's expectation.

Together, the raw-coordinate representation and the trajectory-gradient construction capture two channels through which the interface affects the leader objective. Along a realized history, the interface changes the regression estimates and confidence bonuses entering the conditional policy score. Across possible trajectories, it changes action probabilities and therefore the feedback available for subsequent learning; the likelihood-ratio terms account for this distributional dependence. This construction allows the leader to optimize its own objective using only selected-action feedback while keeping the follower's reward and learning rule fixed.

4.3 Interface Optimization across Deployments

We now use the resulting trajectory gradients to update the interface across repeated deployments. At outer iteration m, the leader runs B independent fresh deployments under the current interface v_m and forms $\widehat{g}_m := B^{-1} \sum_{b=1}^{B} \widehat{g}(v_m;\zeta_{m,b})$, where $\zeta_{m,b}$ denotes the b-th deployment trajectory. To smooth the stochastic variation of the trajectory gradients, CILOT maintains an averaged gradient estimate across outer iterations and updates the interface through a damped projected ascent step. Starting from $s_0 = 0$,

$$s_m := (1-\vartheta_m) s_{m-1} + \vartheta_m \widehat{g}_m, \tag{8a}$$

$$\widetilde{v}_{m+1} := \Pi_{\mathcal{V}}\left(v_m + \eta_m s_m\right), \tag{8b}$$

$$v_{m+1} := (1-\mu_m) v_m + \mu_m \widetilde{v}_{m+1}. \tag{8c}$$

where $\Pi_{\mathcal{V}}$ denotes Euclidean projection onto $\mathcal{V}$, $\vartheta_m \in (0,1]$ controls gradient averaging, $\eta_m > 0$ is the projected ascent step size, and $\mu_m \in (0,1]$ controls how far the interface moves toward the projected point. The projection and the final convex combination keep every interface in $\mathcal{V}$. Algorithm 1 summarizes the procedure. Section 5 establishes a follower regret guarantee uniform over this interface class and analyzes convergence of the leader updates to projected stationarity.

5 Theoretical Guarantees

We provide guarantees for both levels of the system. We first show that a single follower calibration gives sublinear regret uniformly over the admissible interface class. We then bound the sensitivity of the closed-loop follower dynamics and use these bounds to analyze the leader optimization.

Algorithm 1 CILOT with an averaged–damped projected outer update

Input: $v_1 \in \mathcal{V}$, budgets M, B, H, follower schedules $\{\beta_t, \tau_t\}_{t=1}^{H}$, and outer parameters $\{\eta_m, \vartheta_m, \mu_m\}_{m=1}^{M}$.
Initialize $s_0 \leftarrow 0$.
for $m = 1, \dots, M$ **do**
  for $b = 1, \dots, B$ **do**
    Reset the prescribed follower and run one H-round deployment under interface v_m using (2)–(3).
    Accumulate the trajectory gradient $\widehat{g}(v_m; \zeta_{m,b})$ using (7).
  end for
  Set $\widehat{g}_m \leftarrow B^{-1} \sum_{b=1}^{B} \widehat{g}(v_m; \zeta_{m,b})$.
  Update s_m, $\widetilde{v}_{m+1}$, and v_{m+1} according to (8).
end for
Draw $\widehat{m} \sim \mathrm{Unif}\{1, \dots, M\}$ independently and return $\widehat{v} = v_{\widehat{m}}$.

Class-uniform follower guarantee. Although the interface changes the coordinates observed by the follower, the reward model remains realizable with comparator norm at most S/α. Hence the confidence schedule can be fixed before interface optimization and used for every admissible interface.

Theorem 2 (Class-uniform expected follower regret)**.** *Fix θ with $\|\theta\|_2 \le S$ and $\delta_0 \in (0,1)$. Let $\delta_t = \delta_0/[t(t+1)]$ and $\beta_t := \sqrt{\lambda}\frac{S}{\alpha} + \sigma_Y \sqrt{2\log\frac{1}{\delta_t} + d\log\left(1 + \frac{t-1}{d\lambda}\right)}$. Then, for any deterministic temperature schedule $\tau_t > 0$,*

$$\sup_{v \in \mathcal{V}} \mathrm{Reg}_H^L(v;\theta) \le \widetilde{O}(d\sqrt{H}) + (\log K) \sum_{t=1}^{H} \tau_t.$$

Consequently, any schedule satisfying $\sum_{t=1}^{H} \tau_t = o(H)$ gives sublinear regret uniformly over $v \in \mathcal{V}$. In particular, $\tau_t = \tau_0 t^{-1/2}$ yields $\sup_{v\in\mathcal{V}} \mathrm{Reg}_H^L(v;\theta) = \widetilde{O}(d\sqrt{H})$.

The uniformity is what allows the leader to select interfaces adaptively from completed deployments while retaining the same follower guarantee on each fresh deployment.

Sensitivity and objective regularity. The raw-coordinate representation in Proposition 1 makes the dependence of the follower policy on the interface explicit.

Theorem 3 (Closed-loop sensitivity and objective regularity)**.** *Under Assumptions 1 and 2, with $|Y_t| \le R_Y$ almost surely, the follower policy satisfies uniformly over admissible interfaces, realized histories, contexts, and actions, we have*

$$\|\nabla_v \log \pi_t^v(a)\|_2 = \widetilde{O}\left(\frac{\sqrt{d}\,(\sqrt{t} + \beta_t)}{\tau_t}\right), \quad \text{and} \quad \|\nabla_v^2 \log \pi_t^v(a)\|_{\mathrm{op}} = \widetilde{O}\left(d\frac{\sqrt{t} + \beta_t}{\tau_t} + d\frac{t + \beta_t^2}{\tau_t^2}\right). \tag{9}$$

Consequently, $J_{\gamma,H}$ is $L_{\gamma,H}$-smooth on $\mathcal{V}$, and the trajectory estimator satisfies $\mathbb{E}\|\widehat{g}(v;\zeta)\|_2^2 \le \sigma_{\gamma,H}^2$. Explicit expressions for $L_{\gamma,H}$ and $\sigma_{\gamma,H}^2$ are given in Appendices E and F.

The second-order bound includes the softmax covariance term, which is quadratic in the first-order score sensitivity.

Leader optimization guarantee. Let $\mathcal{G}_\eta(v) := \eta^{-1}\left(\Pi_{\mathcal{V}}(v + \eta \nabla J_{\gamma,H}(v)) - v\right)$ denote the projected gradient mapping, and let $\Delta_1 := \sup_{v\in\mathcal{V}} J_{\gamma,H}(v) - J_{\gamma,H}(v_1)$.

Theorem 4 (Fixed-batch projected stationarity)**.** *Run Algorithm 1 with $s_0 = 0$, $\vartheta_m = \mu_m = M^{-1/2}$, and $\eta_m = \eta = c_\eta / L_{\gamma,H}$ for some $0 < c_\eta \le 1/8$. If each $\widehat{g}_m$ averages B conditionally independent fresh trajectory estimators, then the returned iterate $\widehat{v}$ satisfies*

$$\mathbb{E}\|\mathcal{G}_\eta(\widehat{v})\|_2^2 \le \widetilde{O}\left(\frac{L_{\gamma,H}\Delta_1 + G_{\gamma,H}^2}{\sqrt{M}} + \frac{\sigma_{\gamma,H}^2}{B\sqrt{M}}\right),$$

where $G_{\gamma,H} := \sup_{v\in\mathcal{V}} \|\nabla J_{\gamma,H}(v)\|_2 < \infty$. Hence, for fixed problem parameters and batch size,

$$\mathbb{E}\|\mathcal{G}_\eta(\widehat{v})\|_2^2 = \widetilde{O}(M^{-1/2}).$$

6 EXPERIMENTAL EVALUATION

We evaluate whether CILOT improves the leader objective through the admissible feature interface. We consider one synthetic environment and two real-world recommendation datasets, Last.FM and MovieLens. The main text reports leader utility only. Full implementation details, additional results, and analysis are deferred to the Appendix L.

6.1 EXPERIMENTAL SETUP

Synthetic environment. Following the experiments of (Wang et al., 2023), we create $N = |\mathcal{A}| = 5{,}000$ arms, $N_u = 200$ follower types, and use feature dimension $d = 50$. For each arm $a \in \mathcal{A}$ and follower type $u \in \{1, \dots, N_u\}$, we independently draw $\phi_a, \theta_u^\star \sim \mathcal{N}(0, I_d)$ and then normalize each vector to unit ℓ_2 norm, i.e., $\phi_a \leftarrow \phi_a/\|\phi_a\|_2$,$\theta_u^\star \leftarrow \theta_u^\star/\|\theta_u^\star\|_2$. To construct a controlled hidden platform objective, we sample an environment-level latent salience vector $w \in [0,1]^d$, with $w_j \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$, and define $s_a = \sum_{j=1}^d w_j \phi_{a,j}$. Thus, w_j controls the contribution of coordinate j to the platform-side value and provides a known latent salience signal against which the learned interface can be compared. We standardize s_a across arms and transform it into the platform value

$$c_a = \text{sigmoid}\left(\frac{s_a - \text{mean}_{b\in\mathcal{A}}\, s_b}{\text{std}_{b\in\mathcal{A}}\, s_b}\right), \qquad \text{sigmoid}(z) = \frac{1}{1+e^{-z}}.$$

Last.FM and MovieLens. We use the HetRec Last.FM and MovieLens datasets (Cantador et al., 2011). Following prior bandit evaluations (Li et al., 2019; Zhang et al., 2020; Wu et al., 2021), we treat each artist or movie as an arm and retain the $N = 2{,}000$ items and $N_u = 500$ users with the most tag assignments. Define the binary matrix $F \in \{0,1\}^{N_u \times N}$ by $F_{u,a} = \mathbb{I}\{\text{user } u \text{ assigned at least one tag to item } a\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function. From the singular value decomposition $F = U\Sigma V^\top$, we retain the largest 50 singular values and form $\theta_u = (U_{50}\Sigma_{50}^{1/2})_{u,:}$,$\phi_a = (V_{50}\Sigma_{50}^{1/2})_{a,:}$, followed by row normalization. Because these datasets do not contain ground-truth business utility, we use dataset-native aggregate engagement as an item-level platform-value proxy. For Last.FM, let $q_{u,a}$ be the interaction weight recorded in Last.FM and set $p_a = \sum_{u=1}^{N_u} F_{u,a} \log(1+q_{u,a})$. For MovieLens, let $p_a = \sum_{u=1}^{N_u} \mathbb{I}\{(u,a) \text{ appears in the rating data}\}$, the number of distinct users who rated movie a. In both datasets, we set

$$r_a = \log(1+p_a), \qquad c_a = \text{sigmoid}\left(\frac{r_a - \text{mean}_{b\in\mathcal{A}}\, r_b}{\text{std}_{b\in\mathcal{A}}\, r_b}\right).$$

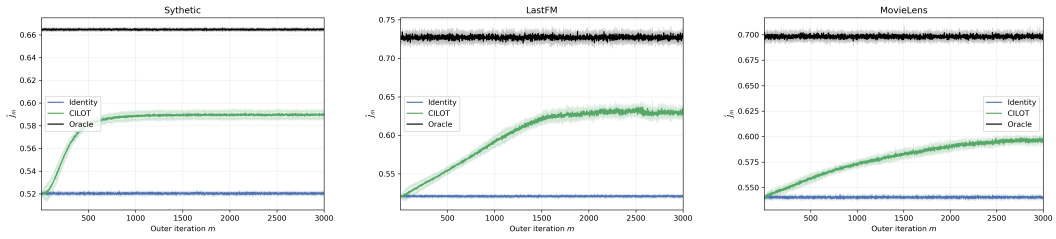
Online protocol and evaluation. Each deployment samples a follower type uniformly with replacement. At round t, an available set $\mathcal{A}_t \subseteq \mathcal{A}$ with $|\mathcal{A}_t| = 50$ is sampled uniformly without replacement from the full arm pool, independently across rounds. After selecting $a_t \in \mathcal{A}_t$, the follower observes $Y_t = \theta_u^{\star\top}\phi_{a_t} + \sigma_y \epsilon_t$, where ϵ_t is standard Gaussian noise truncated to $[-2,2]$, $\sigma_y = 0.05$ for Synthetic, and $\sigma_y = 0$ for Last.FM and MovieLens. The platform value c_{a_t} is not revealed to the follower. We use equal weight on follower reward and platform value and normalize the leader utility as

$$U_t = \frac{Y_t + c_{a_t} + 1 + 2\sigma_y}{3 + 4\sigma_y} \in [0,1]. \tag{10}$$

We compare CILOT with the Identity interface $v = \mathbf{1}$. The follower implementation is identical in both cases. We additionally report a full-information Oracle reference. Given follower type u and candidate set $\mathcal{A}_t$, the Oracle knows the true mean follower reward and leader objective and directly selects $a_t^{\text{oracle}} \in \arg\max_{a\in\mathcal{A}_t} \left(\theta_u^{\star\top}\phi_a + c_a\right)$. The Oracle does not observe future reward noise and is used only for evaluation.

All environments use trajectory horizon $H = 200$, discount factor $\gamma = 0.99$, batch size $B = 16$, and $M = 3000$ outer iterations. At each iteration m, we report

$$\widehat{J}_m = \frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{H} \omega_t^{\gamma,H} U_{m,b,t}, \qquad \omega_t^{\gamma,H} = \frac{(1-\gamma)\gamma^{t-1}}{1-\gamma^H}, \tag{11}$$



(a) Synthetic (b) Last.FM (c) MovieLens

Figure 1: Leader utility during interface optimization on the Synthetic, Last.FM, and MovieLens environments. Curves show the mean over 10 random seeds, and shaded regions denote one standard deviation.

where $U_{m,b,t}$ is the round-t utility of deployment b at outer iteration m. We use 10 independent random seeds and show the mean with one standard deviation. Final leader performance is the last-iteration value $\widehat{J}_M$.

6.2 Main Results

Figure 1 reports leader utility during interface optimization for all three environments. Panel (a) shows the controlled Synthetic environment, while panels (b) and (c) show results on the Last.FM and MovieLens datasets, respectively. All curves report the mean over 10 random seeds, and shaded regions indicate one standard deviation.

Synthetic environment. We first evaluate CILOT in the controlled Synthetic environment, where the platform objective is generated from a known latent coordinate salience vector w. At the final iteration, Identity achieves $\widehat{J}_M = 0.520338 \pm 0.001633$, while CILOT improves the leader utility to 0.589723 ± 0.003678, corresponding to a 13.33% relative improvement.

Beyond the utility improvement, the learned interface closely reflects the latent feature salience used to construct the platform objective. Across random seeds, the final interface v_M has a mean within-seed Pearson correlation of 0.9579 ± 0.0057 with w, indicating a strong positive linear association between latent coordinate salience and the learned interface. We provide a more detailed analysis of this relationship in Appendix L.

Real-world datasets. We further evaluate CILOT on Last.FM and MovieLens datasets. On Last.FM, Identity achieves $\widehat{J}_M = 0.520586 \pm 0.001536$, while CILOT improves the utility to 0.628342 ± 0.005982, corresponding to a 20.70% relative improvement. On MovieLens, Identity obtains $\widehat{J}_M = 0.540315 \pm 0.001373$, whereas CILOT achieves 0.597347 ± 0.003471, yielding a 10.56% relative improvement. The consistent improvements across both recommendation datasets demonstrate that CILOT can learn useful interfaces from naturally occurring user-item interactions. In addition to the final utility improvement, we also observe lower regret for CILOT compared with Identity on both real-world datasets during deployment. Detailed regret evaluations and comparisons are provided in Appendix L.

7 Conclusion

We study feature-interface optimization for a prescribed contextual-bandit learner when the upstream and downstream components have different objectives. Even an invertible coordinate-wise rescaling can change the follower's finite-sample estimation and exploration behavior without changing its reward task. CILOT exploits this dependence by differentiating through the feedback-dependent online updates of the follower, yielding an unbiased trajectory-gradient estimator from selected-action feedback. We establish sublinear follower regret uniformly over the admissible interface class and convergence of the leader optimization to projected stationarity. Experiments on synthetic and real-world data further demonstrate the benefit of optimizing the feature interface.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, manuscript editing, critique of mathematical derivations, and experimental design. The authors reviewed the AI-assisted material and take responsibility for all claims, proofs, citations, code, experiments, and conclusions in the final submission.

REPRODUCIBILITY STATEMENT

The main text and appendix specify the information order, utility normalization, exogeneity and statistical assumptions, common-contract comparison, follower and leader schedules, pathwise derivative convention, exact constants, complete proofs, and a result-ready experimental protocol. The algorithm uses only selected-action feedback. Anonymous code, exact data processing, random seeds, and final configurations will accompany the submission.

REFERENCES

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

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

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

Maria-Florina Balcan, Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Keegan Harris, and Zhiwei Steven Wu. Nearly-optimal bandit learning in stackelberg games with side information. *arXiv preprint arXiv:2502.00204*, 2025.

Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, et al. Tfx: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1387–1395, 2017.

Craig Boutilier, Chih-Wei Hsu, Branislav Kveton, Martin Mladenov, Csaba Szepesvari, and Manzil Zaheer. Differentiable bandit exploration. *arXiv preprint arXiv:2002.06772*, 2020.

Iván Cantador, Peter Brusilovsky, Tsvi Kuflik, et al. Second workshop on information heterogeneity and fusion in recommender systems (hetrec2011). *RecSys*, 11:387–388, 2011.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*, pp. 1360–1370. PMLR, 2020.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pp. 1568–1577. PMLR, 2018.

Keegan Harris, Zhiwei S Wu, and Maria-Florina Balcan. Regret minimization in stackelberg games with side information. *Advances in Neural Information Processing Systems*, 37:12944–12976, 2024.

Parnian Kassraie, Jonas Rothfuss, and Andreas Krause. Meta-learning hypothesis spaces for sequential decision-making. In *International Conference on Machine Learning*, pp. 10802–10824. PMLR, 2022.

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, pp. 5884–5893. PMLR, 2021.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Shuai Li, Wei Chen, and Kwong-Sak Leung. Improved algorithm on online clustering of bandits. *arXiv preprint arXiv:1902.09162*, 2019.

Tao Lin and Yiling Chen. Generalized principal-agent problem with a learning agent. In *International Conference on Learning Representations*, volume 2025, pp. 89757–89788, 2025.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pp. 2113–2122. PMLR, 2015.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pp. 8371–8380. PMLR, 2021.

Antoine Scheid, Daniil Tiapkin, Etienne Boursier, Aymeric Capitaine, El Mahdi El Mhamdi, Éric Moulines, Michael I Jordan, and Alain Durmus. Incentivized learning in principal-agent bandit games. *arXiv preprint arXiv:2403.03811*, 2024.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. *Advances in neural information processing systems*, 28, 2015.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

Amr Sharaf and Hal Daumé III. Meta-learning for contextual bandit exploration. *arXiv preprint arXiv:1901.08159*, 2019.

Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. *Advances in Neural Information Processing Systems*, 37:127369–127435, 2024.

Andrea Tirinzoni, Matteo Papini, Ahmed Touati, Alessandro Lazaric, and Matteo Pirotta. Scalable representation learning in linear contextual bandits with constant regret guarantees. *Advances in Neural Information Processing Systems*, 35:2307–2319, 2022.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. On the complexity of representation learning in contextual linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 7871–7896. PMLR, 2023.

Andrew Wagenmaker, Julian Katz-Samuels, and Kevin Jamieson. Experimental design for regret minimization in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3088–3096. PMLR, 2021.

Zhiyong Wang, Xutong Liu, Shuai Li, and John CS Lui. Efficient explorative key-term selection strategies for conversational contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10288–10295, 2023.

Junda Wu, Canzhe Zhao, Tong Yu, Jingyang Li, and Shuai Li. Clustering of conversational bandits for user preference learning and elicitation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2129–2139, 2021.

Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-gradient reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Jiaqi Yang, Wei Hu, Jason D Lee, and Simon S Du. Impact of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*, 2020.

Xiaoying Zhang, Hong Xie, Hang Li, and John CS Lui. Conversational contextual bandit: Algorithm and application. In *Proceedings of the web conference 2020*, pp. 662–672, 2020.

This appendix collects all auxiliary statements and proofs. The first section provides core notation, complete auxiliary statements, and an explicit proof-dependency map. Explicit sensitivity constants are summarized where they are first used in Appendix E; subsequent sections follow the dependency order stated below.

A Technical Statements, Notation, and Proof Roadmap

This section records the notation, deterministic envelopes, auxiliary statements, and proof dependencies used throughout the appendix. The proofs then proceed from structural properties and follower regret to the trajectory gradient, raw-coordinate sensitivity, Fisher-refined regularity, estimator moments, and outer-loop convergence.

A.1 Core notation

Table 1: Core notation used in the appendix.

Symbol	Meaning
M	Number of outer iterations
B	Number of independent fresh deployments per outer iteration
H	Horizon of each deployment
K	Number of actions
d	Raw-feature dimension
$\theta^\star$	Deployment-specific raw-coordinate reward parameter
$\phi(x,a)$	Raw feature vector
$W(v)$	Diagonal interface matrix $\mathrm{Diag}(v)$
$z_v(x,a)$	Exposed feature $W(v)\phi(x,a)$
V_t, b_t	Visible-coordinate ridge-regression state
$\widehat{\psi}_t$	Visible-coordinate ridge estimate $V_t^{-1}b_t$
Σ_t, c_t	Raw-coordinate sufficient statistics before round t
$\Lambda(v)$	Raw-coordinate ridge matrix $\lambda W(v)^{-2}$
$A_t(v)$	Regularized raw-coordinate Gram matrix $\Sigma_t + \Lambda(v)$
$\widehat{\theta}_t(v)$	Raw-coordinate ridge estimate $A_t(v)^{-1}c_t$
$r_t^v(a)$	Confidence radius $\sqrt{\phi(x_t,a)^\top A_t(v)^{-1}\phi(x_t,a)}$
$q_t^v(a)$	Optimistic score $\widehat{\theta}_t(v)^\top\phi(x_t,a) + \beta_t r_t^v(a)$
π_t^v	Soft optimistic action distribution
U_t	Bounded leader utility $u(x_t,a_t,Y_t) \in [0,1]$
$G_t^{\gamma,H}$	Discounted return-to-go $\sum_{s=t}^{H}\omega_s^{\gamma,H}U_s$
ℓ_t	Conditional policy score $\nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1})$
$\widehat{g}_m$	Mini-batch complete-state trajectory-gradient estimator

A.2 Proof roadmap

The proof dependencies are as follows. Proposition 2 provides the common visible-space comparator used by the follower-regret analysis. The structural examples in Appendix B are independent of the asymptotic guarantees. Theorem 6 uses only the trajectory factorization and bounded score envelopes. The raw-coordinate equivalence and sensitivity analysis in Appendix E provide those envelopes. They are then combined with the Fisher-domination argument in Appendix F and the estimator-moment bounds in Appendix G. Finally, Appendix H converts unbiasedness, smoothness, and variance control into projected stationarity, and Appendix I specializes the exact bounds to the square-root temperature schedule. The visible-coordinate recursions in Appendix J are an equivalent implementation check and are not needed for the uniform raw-coordinate bounds.

A.3 Tail-moment identities for normalized geometric weights

The normalized geometric weights are

$$\omega_t^{\gamma,H} = \frac{(1-\gamma)\gamma^{t-1}}{1-\gamma^H}, \qquad t = 1,\dots,H.$$

They sum to one, so $0 \le R_{\gamma,H} \le 1$ whenever $0 \le U_t \le 1$.

Lemma 1 (Exact discounted tail masses)**.** *For every $t \le H$,*

$$\overline{W}_t^{\gamma,H} := \sum_{s=t}^{H} \omega_s^{\gamma,H} = \frac{\gamma^{t-1}(1-\gamma^{H-t+1})}{1-\gamma^H} \le \gamma^{t-1}.$$

Proof. Summing the finite geometric series gives

$$\sum_{s=t}^{H} \frac{(1-\gamma)\gamma^{s-1}}{1-\gamma^H} = \frac{\gamma^{t-1}-\gamma^H}{1-\gamma^H} = \frac{\gamma^{t-1}(1-\gamma^{H-t+1})}{1-\gamma^H}.$$

Since $0 < \gamma < 1$, $1-\gamma^{H-t+1} \le 1-\gamma^H$, proving the inequality. □

Lemma 2 (Exact discounted tail moments)**.** *Define*

$$S_0(\gamma,H) := \sum_{t=1}^{H} \overline{W}_t^{\gamma,H}, \qquad S_1(\gamma,H) := \sum_{t=1}^{H} t\overline{W}_t^{\gamma,H}.$$

Then

$$S_0(\gamma,H) = \sum_{t=1}^{H} t\omega_t^{\gamma,H} \le (1-\gamma)^{-1}, \quad S_1(\gamma,H) = \frac{1}{2}\sum_{t=1}^{H} t(t+1)\omega_t^{\gamma,H} \le (1-\gamma)^{-2}.$$

Proof.

$$S_0 = \sum_{t=1}^{H}\sum_{s=t}^{H} \omega_s^{\gamma,H} = \frac{\sum_{t=1}^{H}(\gamma^{t-1}-\gamma^H)}{1-\gamma^H} = \sum_{t=1}^{H} t\omega_t^{\gamma,H},$$

and

$$S_1 = \sum_{t=1}^{H} t\sum_{s=t}^{H} \omega_s^{\gamma,H} = \sum_{t=1}^{H} \omega_t^{\gamma,H}\sum_{s=1}^{t} s = \frac{1}{2}\sum_{t=1}^{H} t(t+1)\omega_t^{\gamma,H}.$$

Using Lemma 2, the inequality comes from

$$S_0 = \sum_{t=1}^{H} \overline{W}_t^{\gamma,H} \le \sum_{t=1}^{H} \gamma^{t-1} = \frac{1-\gamma^H}{1-\gamma} \le (1-\gamma)^{-1},$$

and

$$S_1 = \sum_{t=1}^{H} t\overline{W}_t^{\gamma,H} \le \sum_{t=1}^{H} t\gamma^{t-1} \le \sum_{t=1}^{\infty} t\gamma^{t-1} = (1-\gamma)^{-2}.$$

□

Remark 1 (Exact finite-horizon quantities)**.** *The factors $(1-\gamma)^{-1}$ and $(1-\gamma)^{-2}$ are convenient upper bounds, not replacements for the exact finite-horizon sums. The smoothness, variance, and stationarity statements retain the exact tail masses $\overline{W}_t^{\gamma,H}$ through $\mathcal{A}_{\gamma,H}$ and $\mathcal{C}_{\gamma,H}$.*

A.4 Realizability and structural statements

Proposition 2 (Realizability under every interface)**.** *For every $v \in \mathcal{V}$ and every $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 \le S$, define*

$$\psi^\star_{\theta,v} := W(v)^{-T}\theta.$$

Then, for every context-action pair (x,a),

$$\left\langle \psi^\star_{\theta,v}, z_v(x,a)\right\rangle = \langle \theta, \phi(x,a)\rangle, \qquad \|\psi^\star_{\theta,v}\|_2 \le \frac{S}{\alpha}.$$

Proposition 3 (Non-vacuity under aligned and conflicting objectives). *Suppose* $0 < \alpha < 1$, $S > 0$, $\beta_1 > 0$, *and* $\tau_1 > 0$. *There exist one-round, two-action realizable instances in which an admissible interface strictly increases the leader objective and either* (i) *strictly decreases follower regret when the objectives are aligned or* (ii) *strictly increases follower regret when the objectives conflict. Both instances use the same admissible interface class.*

Proposition 4 (A robust closed-loop data-acquisition pathway). *Suppose* $0 < \alpha < 1$, $S > 0$, $\beta_1 > 0$, $\beta_2 > 0$, *and* $\tau_1, \tau_2 > 0$. *There exists a two-round, two-action realizable instance with a nonzero reward parameter in which the first-round action probability depends nontrivially on the interface, the selected first-round action changes the second-round covariance, and the resulting data-acquisition-mediated term in the derivative of the expected second-round leader utility is nonzero. These properties persist on an open set of nearby admissible instances.*

Proposition 5 (Bias from stopped-state differentiation). *Suppose* $0 < \alpha < 1$ *and* $S > 0$. *For every finite* $\beta_2 \geq 0$ *and* $\tau_2 > 0$, *there exist a realizable two-round, two-action instance with nonzero action-feature vectors and an admissible interface* v *such that one coordinate of the expected stopped-state trajectory estimator is zero, whereas the corresponding coordinate of* $\nabla J_{\gamma,2}(v)$ *is nonzero. The stopped-state estimator differentiates the current conditional policy while treating* V_t, b_t, *and* $\widehat{\psi}_t$ *as constants. The mismatch persists under all sufficiently small admissible perturbations of the construction.*

A.5 Class-uniform regret of temperature-scheduled Soft-LinUCB

Theorem 5 (Anytime uniform schedule-dependent Soft-LinUCB regret). *Fix a follower type* θ, *let* $B_z := 1$ *and* $B_\psi := S/\alpha$, *and suppose Assumption 1 holds. Choose* $\delta_0 \in (0,1)$, *set*

$$\delta_t := \frac{\delta_0}{t(t+1)},$$

and use any deterministic optimism schedule satisfying

$$\beta_t \geq \sqrt{\lambda} B_\psi + \sigma_Y \sqrt{2\log(1/\delta_t) + d\log\left(1 + \frac{(t-1)B_z^2}{d\lambda}\right)}.$$

For every deterministic positive temperature schedule,

$$\sup_{v \in \mathcal{V}} \mathrm{Reg}_H^L(v;\theta) \leq 2\bar{\beta}_H \left(\frac{B_z}{\sqrt{\lambda}} + 1\right) \sqrt{2dH\log\left(1 + \frac{HB_z^2}{d\lambda}\right)} + (\log K)\sum_{t=1}^{H} \tau_t + 2S\sum_{t=1}^{H} \delta_t,$$

where $\bar{\beta}_H := \max_{t \leq H} \beta_t$. *Since* $\sum_{t \geq 1} \delta_t = \delta_0$, *the last term is at most* $2S\delta_0$. *The same deterministic upper bound holds pointwise for every fixed interface; therefore taking the supremum over* $\mathcal{V}$ *does not require a simultaneous confidence event over the uncountable interface class.*

Corollary 1 (Anytime square-root-temperature no-regret). *Under Theorem 5, choose* $\tau_t = \tau_0 t^{-1/2}$. *Then, for every follower type* θ,

$$\sup_{v \in \mathcal{V}} \mathrm{Reg}_H^L(v;\theta) \leq \widetilde{O}\left(d\sqrt{H}\ \mathrm{poly}(S/\alpha, \sigma_Y, \lambda, \lambda^{-1}) + \tau_0(\log K)\sqrt{H}\right),$$

and hence $\sup_v \mathrm{Reg}_H^L(v;\theta)/H = \widetilde{O}(H^{-1/2})$. *The schedules do not depend on the terminal horizon.*

A.6 Trajectory hypergradient and pathwise convention

Let

$$G_t := G_t^{\gamma,H} = \sum_{s=t}^{H} \omega_s^{\gamma,H} U_s.$$

Theorem 6 (Normalized-weight trajectory hypergradient identity)**.** *Under Assumptions 2 and 1, including $|Y_t| \le R_Y$ almost surely, the population objective is differentiable on an open neighborhood of $\mathcal{V}$ contained in $(0,\infty)^d$, and*

$$\nabla J_{\gamma,H}(v) = \mathbb{E}_{\theta\sim\mathcal{P}_\Theta}\mathbb{E}_{h_H\sim P_v^\theta}\left[\sum_{t=1}^{H} G_t \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1})\right].$$

Remark 2 (Pathwise derivative convention)**.** *All derivatives of $V_t, b_t, \widehat{\psi}_t$, and the conditional policy are computed along a fixed realized history. The realized contexts, actions, and selected-action rewards are held fixed as coordinates of that trajectory. The dependence of the trajectory law on v is handled only through the likelihood-ratio score increments. Differentiating sampled actions inside the state recursion and again through the likelihood ratio would double count the same dependence.*

A.7 Equivalent visible-coordinate derivative recursions

For $i, j \in [d]$, define

$$\Xi_t^{(i)} := \partial_{v_i} V_t, \quad \Omega_t^{(ij)} := \partial_{v_j}\partial_{v_i} V_t, \quad \zeta_t^{(i)} := \partial_{v_i} b_t, \quad \xi_t^{(ij)} := \partial_{v_j}\partial_{v_i} b_t,$$
$$\chi_t^{(i)} := \partial_{v_i}\widehat{\psi}_t, \qquad \Gamma_t^{(ij)} := \partial_{v_j}\partial_{v_i}\widehat{\psi}_t.$$

Lemma 3 (Exact state derivative recursions)**.** *Along any realized trajectory,*

$$\Xi_{t+1}^{(i)} = \Xi_t^{(i)} + (\partial_{v_i} z_t(a_t)) z_t(a_t)^\top + z_t(a_t)(\partial_{v_i} z_t(a_t))^\top,$$
$$\begin{aligned}\Omega_{t+1}^{(ij)} = \Omega_t^{(ij)} &+ (\partial^2_{v_i v_j} z_t(a_t)) z_t(a_t)^\top + (\partial_{v_i} z_t(a_t))(\partial_{v_j} z_t(a_t))^\top \\ &+ (\partial_{v_j} z_t(a_t))(\partial_{v_i} z_t(a_t))^\top + z_t(a_t)(\partial^2_{v_i v_j} z_t(a_t))^\top,\end{aligned}$$
$$\zeta_{t+1}^{(i)} = \zeta_t^{(i)} + (\partial_{v_i} z_t(a_t)) Y_t, \qquad \xi_{t+1}^{(ij)} = \xi_t^{(ij)} + (\partial^2_{v_i v_j} z_t(a_t)) Y_t,$$
$$\chi_t^{(i)} = -V_t^{-1}\Xi_t^{(i)}\widehat{\psi}_t + V_t^{-1}\zeta_t^{(i)},$$

and

$$\begin{aligned}\Gamma_t^{(ij)} = {} & V_t^{-1}\Xi_t^{(j)}V_t^{-1}\Xi_t^{(i)}\widehat{\psi}_t + V_t^{-1}\Xi_t^{(i)}V_t^{-1}\Xi_t^{(j)}\widehat{\psi}_t - V_t^{-1}\Omega_t^{(ij)}\widehat{\psi}_t \\ & - V_t^{-1}\Xi_t^{(i)}V_t^{-1}\zeta_t^{(j)} - V_t^{-1}\Xi_t^{(j)}V_t^{-1}\zeta_t^{(i)} + V_t^{-1}\xi_t^{(ij)}.\end{aligned}$$

Lemma 4 (Exact visible-coordinate score derivatives)**.** *Let*

$$s_t^v(a) := z_t(a)^\top V_t^{-1} z_t(a), \qquad r_t^v(a) := \sqrt{s_t^v(a)}, \qquad q_t^v(a) := \widehat{\psi}_t^\top z_t(a) + \beta_t r_t^v(a).$$

Then, for all $i, j \in [d]$,

$$\partial_{v_i} s_t^v(a) = 2(\partial_{v_i} z_t(a))^\top V_t^{-1} z_t(a) - z_t(a)^\top V_t^{-1}\Xi_t^{(i)} V_t^{-1} z_t(a),$$
$$\begin{aligned}\partial^2_{v_i v_j} s_t^v(a) = {} & 2(\partial^2_{v_i v_j} z_t(a))^\top V_t^{-1} z_t(a) + 2(\partial_{v_i} z_t(a))^\top V_t^{-1}(\partial_{v_j} z_t(a)) \\ & - 2(\partial_{v_i} z_t(a))^\top V_t^{-1}\Xi_t^{(j)} V_t^{-1} z_t(a) - 2(\partial_{v_j} z_t(a))^\top V_t^{-1}\Xi_t^{(i)} V_t^{-1} z_t(a) \\ & + z_t(a)^\top V_t^{-1}\Xi_t^{(j)} V_t^{-1}\Xi_t^{(i)} V_t^{-1} z_t(a) + z_t(a)^\top V_t^{-1}\Xi_t^{(i)} V_t^{-1}\Xi_t^{(j)} V_t^{-1} z_t(a) \\ & - z_t(a)^\top V_t^{-1}\Omega_t^{(ij)} V_t^{-1} z_t(a).\end{aligned}$$

If $\phi(x_t, a) \neq 0$, then $r_t^v(a) > 0$ and

$$\partial_{v_i} r_t^v(a) = \frac{\partial_{v_i} s_t^v(a)}{2 r_t^v(a)},$$
$$\partial^2_{v_i v_j} r_t^v(a) = \frac{\partial^2_{v_i v_j} s_t^v(a)}{2 r_t^v(a)} - \frac{\partial_{v_i} s_t^v(a)\partial_{v_j} s_t^v(a)}{4 r_t^v(a)^3}.$$

If $\phi(x_t, a) = 0$, then $r_t^v(a) \equiv 0$ and both derivatives are zero. Finally,

$$\partial_{v_i} q_t^v(a) = (\chi_t^{(i)})^\top z_t(a) + \widehat{\psi}_t^\top \partial_{v_i} z_t(a) + \beta_t \partial_{v_i} r_t^v(a),$$

and

$$\begin{aligned}\partial^2_{v_i v_j} q_t^v(a) = {} & (\Gamma_t^{(ij)})^\top z_t(a) + (\chi_t^{(i)})^\top \partial_{v_j} z_t(a) + (\chi_t^{(j)})^\top \partial_{v_i} z_t(a) \\ & + \widehat{\psi}_t^\top \partial^2_{v_i v_j} z_t(a) + \beta_t \partial^2_{v_i v_j} r_t^v(a).\end{aligned}$$

Lemma 5 (Softmax score derivatives). *Suppose* $\pi_t^v(a \mid x_t, h_{t-1}) \propto \exp(q_t^v(a)/\tau_t)$ *with deterministic* $\tau_t > 0$. *Then*

$$\nabla_v \log \pi_t^v(a \mid x_t, h_{t-1}) = \frac{1}{\tau_t}\left(\nabla_v q_t^v(a) - \sum_{b=1}^{K} \pi_t^v(b \mid x_t, h_{t-1}) \nabla_v q_t^v(b)\right).$$

Moreover, for all $i, j \in [d]$,

$$\partial^2_{v_i v_j} \log \pi_t^v(a \mid x_t, h_{t-1}) = \frac{1}{\tau_t}\left(\partial^2_{v_i v_j} q_t^v(a) - \sum_{b=1}^{K} \pi_t^v(b) \partial^2_{v_i v_j} q_t^v(b)\right) - \frac{1}{\tau_t^2}\left(\sum_{b=1}^{K} \pi_t^v(b) \partial_{v_i} q_t^v(b) \partial_{v_j} q_t^v(b) - \bar{q}_i(t)\bar{q}_j(t)\right),$$

where $\bar{q}_i(t) := \sum_{b=1}^{K} \pi_t^v(b) \partial_{v_i} q_t^v(b)$.

A.8 Estimator moments, smoothness, and projected ascent

Let D_m denote the B population episodes sampled at outer iteration m. The filtration $\mathcal{F}_{m-1}$ contains all information revealed before the current fresh batch, including v_m and s_{m-1}. Conditional on $\mathcal{F}_{m-1}$, the episodes in D_m are independent and use fresh type, context, action, and reward randomness. For episode b, define

$$Z_{m,b} := \sum_{t=1}^{H} G_{m,b,t} \nabla_v \log \pi_t^{v_m}(a_{m,b,t} \mid x_{m,b,t}, h_{m,b,t-1}), \qquad \widehat{g}_m := \frac{1}{B}\sum_{b=1}^{B} Z_{m,b}.$$

Theorem 7 (Conditional unbiasedness). *Under Assumptions 2 and 1, including bounded feedback, and under the fresh-batch conditions above,*

$$\mathbb{E}[\widehat{g}_m \mid \mathcal{F}_{m-1}] = \nabla J_{\gamma,H}(v_m).$$

Theorem 8 (Fisher-refined per-episode second moment). *Suppose that, for each* $t \le H$, *a deterministic constant* $G_{\pi,t}$ *satisfies*

$$\|\nabla_v \log \pi_t^v(a \mid x_t, h_{t-1})\|_2 \le G_{\pi,t}$$

uniformly over admissible interfaces, histories, contexts, and actions. Define

$$C_{\mathrm{est}} := \sum_{t=1}^{H} \overline{W}_t^{\gamma,H} G_{\pi,t}, \qquad V_{\mathrm{est}} := \sum_{t=1}^{H} \overline{W}_t^{\gamma,H} G_{\pi,t}^2.$$

Then $\|Z_{m,b}\|_2 \le C_{\mathrm{est}}$ *almost surely and*

$$\mathbb{E}[\|Z_{m,b}\|_2^2 \mid \mathcal{F}_{m-1}] \le V_{\mathrm{est}}.$$

Consequently,

$$\mathbb{E}\left[\|Z_{m,b} - \mathbb{E}[Z_{m,b} \mid \mathcal{F}_{m-1}]\|_2^2 \middle| \mathcal{F}_{m-1}\right] \le V_{\mathrm{est}}.$$

Lemma 6 (Conditional mini-batch variance reduction). *Suppose* $Z_{m,1}, \dots, Z_{m,B}$ *are conditionally independent given* $\mathcal{F}_{m-1}$ *and each has conditional variance at most* $\sigma^2_{\gamma,H}$. *Then*

$$\mathbb{E}\left[\|\widehat{g}_m - \nabla J_{\gamma,H}(v_m)\|_2^2 \middle| \mathcal{F}_{m-1}\right] \le \frac{\sigma^2_{\gamma,H}}{B}.$$

In particular, Theorem 8 permits the deterministic choice $\sigma^2_{\gamma,H} = V_{\mathrm{est}}$.

Lemma 7 (Bounded-multiplier Fisher domination). *Let* $\{\ell_t\}_{t=1}^{s}$ *be the trajectory-score increments up to time* s, *and let* $\mathcal{I}_t$ *denote the pre-action information at time* t. *Suppose*

$$\mathbb{E}[\ell_t \mid \mathcal{I}_t] = 0, \qquad \|\ell_t\|_2 \le G_{\pi,t} \quad \textit{almost surely}.$$

Then, for every random variable R *satisfying* $0 \le R \le 1$,

$$\left\| \mathbb{E}\left[R \left(\sum_{t=1}^{s} \ell_t\right)\left(\sum_{t=1}^{s} \ell_t\right)^{\top}\right] \right\|_{\mathrm{op}} \le \sum_{t=1}^{s} G_{\pi,t}^2.$$

Theorem 9 (Discounted Fisher-refined smoothness)**.** *Under Assumption 2, suppose the conditional policies are twice continuously differentiable on an open neighborhood of $\mathcal{V}$ and, for all $t \le H$ and all admissible histories,*

$$\|\nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1})\|_2 \le G_{\pi,t}, \qquad \|\nabla_v^2 \log \pi_t^v(a_t \mid x_t, h_{t-1})\|_{\mathrm{op}} \le H_{\pi,t}.$$

Then $J_{\gamma,H}$ is L_J-smooth on $\mathcal{V}$, where

$$L_J := \sum_{t=1}^{H} \overline{W}_t^{\gamma,H} \left(H_{\pi,t} + G_{\pi,t}^2\right).$$

Lemma 8 (Exact projected-ascent step)**.** *Let F be L-smooth on a closed convex set $\mathcal{V}$. For*

$$y := \Pi_{\mathcal{V}}(v + \eta \nabla F(v)), \qquad \mathcal{G}_\eta(v) := \eta^{-1}(y - v),$$

if $0 < \eta \le 1/L$, then

$$F(y) - F(v) \ge \frac{\eta}{2} \|\mathcal{G}_\eta(v)\|_2^2.$$

Proposition 6 (Direct constant-step projected stationarity)**.** *Let $F = J_{\gamma,H}$ be $L_{\gamma,H}$-smooth on $\mathcal{V}$, and define $F^\star := \sup_{v \in \mathcal{V}} F(v)$. Suppose the mini-batch estimator is conditionally unbiased and satisfies*

$$\mathbb{E}[\|\widehat{g}_m - \nabla F(v_m)\|_2^2 \mid \mathcal{F}_{m-1}] \le \frac{\sigma_{\gamma,H}^2}{B_m}.$$

If

$$v_{m+1} = \Pi_{\mathcal{V}}(v_m + \eta \widehat{g}_m), \qquad 0 < \eta \le (2L_{\gamma,H})^{-1},$$

then

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\mathcal{G}_\eta(v_m)\|_2^2 \le O\left(\frac{F^\star - F(v_1)}{\eta M} + \frac{\sigma_{\gamma,H}^2}{M} \sum_{m=1}^{M} \frac{1}{B_m}\right).$$

Consequently, fixed $B_m \equiv B$ gives a variance-neighborhood upper bound, whereas $B_m = \lceil B_0 m \rceil$ gives

$$O\left(\frac{F^\star - F(v_1)}{\eta M} + \frac{\sigma_{\gamma,H}^2 \log(eM)}{B_0 M}\right).$$

B Structural Properties and Non-Vacuity

This appendix proves the realizability and structural statements collected in Appendix A. The examples are deliberately small: their purpose is to isolate the finite-sample and data-acquisition pathways created by the interface, rather than to model a particular application.

B.1 Proof of Proposition 2

Proof. For every coordinate, $v_i \in [\alpha, 1]$, so $(\psi^\star_{\theta,v})_i = \theta_i / v_i$ is well defined. Therefore,

$$\left\langle \psi^\star_{\theta,v}, z_v(x,a) \right\rangle = \sum_{i=1}^{d} \frac{\theta_i}{v_i} v_i \phi_i(x,a) = \langle \theta, \phi(x,a) \rangle.$$

Moreover,

$$\|\psi^\star_{\theta,v}\|_2^2 = \sum_{i=1}^{d} \frac{\theta_i^2}{v_i^2} \le \alpha^{-2} \|\theta\|_2^2 \le \frac{S^2}{\alpha^2}.$$

Taking square roots proves the claim. □

B.2 Proof of Proposition 3

Proof. Fix any $\alpha < c < 1$, let $H = 1$ and $d = K = 2$, and consider a deterministic context with raw action features

$$\phi(a_1) = (c, 0)^\top, \qquad \phi(a_2) = (0, 1)^\top.$$

Let

$$U_1 = \mathbf{1}\{a_1 \text{ is selected}\},$$

and take arbitrary $\lambda, \beta_1, \tau_1 > 0$. Since the follower has no observations before the first round, $\widehat{\psi}_1 = 0$, and its optimistic scores depend only on the exposed feature norms.

Under the identity interface,

$$\pi_1^{\mathbf{1}}(a_1) = \frac{\exp\left(\beta_1 c/(\tau_1\sqrt{\lambda})\right)}{\exp\left(\beta_1 c/(\tau_1\sqrt{\lambda})\right) + \exp\left(\beta_1/(\tau_1\sqrt{\lambda})\right)} < \frac{1}{2}.$$

Under the admissible interface $v = (1, \alpha)$,

$$\pi_1^{v}(a_1) = \frac{\exp\left(\beta_1 c/(\tau_1\sqrt{\lambda})\right)}{\exp\left(\beta_1 c/(\tau_1\sqrt{\lambda})\right) + \exp\left(\beta_1 \alpha/(\tau_1\sqrt{\lambda})\right)} > \frac{1}{2},$$

where the strict inequality follows from $\alpha < c$. Hence $J_{\gamma,1}(v) > J_{\gamma,1}(\mathbf{1})$.

Choose any $\rho \in (0, S]$. For the aligned instance, set $\theta = (\rho, 0)^\top$. Action a_1 is uniquely follower optimal, with mean reward ρc, whereas a_2 has mean reward zero. Therefore,

$$\begin{aligned}\mathrm{Reg}_1^L(v; \theta) &= \rho c\big(1 - \pi_1^v(a_1)\big) \\ &< \rho c\big(1 - \pi_1^{\mathbf{1}}(a_1)\big) = \mathrm{Reg}_1^L(\mathbf{1}; \theta).\end{aligned}$$

For the conflicting instance, set $\theta = (0, \rho)^\top$. Action a_2 is uniquely follower optimal, with mean reward ρ, whereas a_1 has mean reward zero. The same interface still increases the leader's probability of selecting a_1, but

$$\mathrm{Reg}_1^L(v; \theta) = \rho\pi_1^v(a_1) > \rho\pi_1^{\mathbf{1}}(a_1) = \mathrm{Reg}_1^L(\mathbf{1}; \theta).$$

Thus, the interface is behaviorally non-vacuous under both aligned and conflicting objectives, without changing linear realizability or the admissible interface class. □

B.3 Proof of Proposition 4

Proof. We first establish a strict data-acquisition effect at a convenient anchor instance and then use continuity to obtain a nonzero reward parameter and an open set of nearby instances.

Let $d = 1$, $K = 2$, $H = 2$, and $\lambda = 1$. At both rounds, take the raw action features

$$\phi(1) = 1, \qquad \phi(2) = c, \qquad 0 < c < 1.$$

Fix finite positive optimism widths β_1, β_2 and temperatures τ_1, τ_2. Let $U_1 = 0$ and

$$U_2 = \mathbf{1}\{a_2 = 1\}.$$

For a scalar interface $v \in (\alpha, 1)$, first consider the anchor reward parameter $\theta = 0$ with deterministic zero noise.

The first-round probability of selecting action 1 is

$$p_1(v, 0) = \frac{\exp(\beta_1 v/\tau_1)}{\exp(\beta_1 v/\tau_1) + \exp(\beta_1 vc/\tau_1)} = \sigma\left(\frac{\beta_1 v(1-c)}{\tau_1}\right),$$

where $\sigma(s) = (1 + e^{-s})^{-1}$. Hence $\partial_v p_1(v, 0) = \frac{\beta_1 v(1-c)}{\tau_1} p_1(v, 0)(1 - p_1(v, 0)) > 0$.

Conditional on the first-round action $j \in \{1, 2\}$, the second-round covariance is

$$V_2(j) = 1 + v^2\phi(j)^2.$$

At the anchor $\theta = 0$ instance, we have $Y_1 = 0, \hat{\psi}_2 = 0, b_2 = 0$. The conditional probability of selecting action 1 in the second round is therefore

$$g_j(v, 0) := \mathbb{P}_v(a_2 = 1 \mid a_1 = j) = \sigma\left(\frac{\beta_2 v(1-c)}{\tau_2\sqrt{1+v^2\phi(j)^2}}\right).$$

Because $\phi(1) = 1 > c = \phi(2)$, selecting action 1 creates the larger second-round covariance and therefore the smaller optimistic logit gap. Thus,

$$g_1(v, 0) < g_2(v, 0) \qquad \text{for every } v > 0.$$

Write the expected second-round utility as

$$J_2(v, \theta) = p_1(v, \theta)g_1(v, \theta) + \big(1 - p_1(v, \theta)\big)g_2(v, \theta).$$

Its derivative contains the data-acquisition-mediated term

$$T(v, \theta) := \partial_v p_1(v, \theta)\big(g_1(v, \theta) - g_2(v, \theta)\big).$$

At the anchor instance, $T(v, 0) < 0$.

Now use deterministic feedback $Y_t = \theta\phi(a_t)$. All logits, action probabilities, and their derivatives are smooth functions of (v, θ, c) on the interior domain $v > 0$ and $0 < c < 1$. Since $\partial_v p_1(v, 0) > 0$ and $g_1(v, 0) < g_2(v, 0)$ are strict, continuity implies that there exist $\varepsilon_\theta > 0$ and an open neighborhood of the chosen (v, c) such that

$$T(v, \theta) \neq 0$$

for every $0 < |\theta| < \varepsilon_\theta$ throughout that neighborhood. Choose $0 < |\theta| < \min\{\varepsilon_\theta, S\}$. The strict inequalities also persist after an arbitrarily small inward scaling of the two raw features, so the feature-norm constraints can be made strict. Hence the construction can be chosen in the relative interior of the admissible parameter set, and the nonzero contribution persists on an open set of nearby admissible instances. The term arises because the interface changes the early action distribution and the selected action changes the future covariance; it is absent from a derivative that considers only the current-round logits. □

B.4 Proof of Proposition 5

The stopped-state estimator omits the dependence of the numerical regression state on v along a fixed realized history. Likelihood-ratio differentiation accounts for how v changes the probability of realizing that history, but it does not recover a derivative that was omitted from the conditional policy score itself.

Proof. Fix any finite $\beta_2 \geq 0$ and $\tau_2 > 0$. Take $H = 2$, $d = K = 2$, and $\lambda = 1$. Choose $u, w \in (\alpha, 1)$ and evaluate the interface at $v = (u, w)$. Choose any $\rho \in (0, S)$ and let

$$\theta = -\frac{\rho}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix},$$

with deterministic zero noise.

At round one, assign both actions the same nonzero feature:

$$\phi_1(1) = \phi_1(2) = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}, \qquad Y_1 = -\rho.$$

The two first-round logits are identical for every interface, so $\ell_1(v) = 0$. Since the two actions also produce the same feature and feedback, the regression state after round one is independent of the sampled action label.

At round two, let

$$\phi_2(1) = e_2, \qquad \phi_2(2) = \varepsilon e_2, \qquad U_1 = 0, \qquad U_2 = \mathbf{1}\{a_2 = 1\},$$

where $\varepsilon > 0$ will be chosen sufficiently small. Both second-round action features are nonzero and have support only in the second coordinate. Consequently, if V_2, b_2, and $\widehat{\psi}_2$ are held fixed,

both second-round logits have zero derivative with respect to v_1. Together with $\ell_1(v) = 0$, the stopped-state trajectory estimator has zero first coordinate on every realized trajectory.

It remains to show that the true derivative has a nonzero first coordinate. First consider the continuous limit $\varepsilon = 0$. Along the fixed first-round history,

$$\Sigma_2 = \phi_1 \phi_1^T = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad c_2 = \phi_1 Y_1 = -\frac{\rho}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

At the interior interface $v = (u, w)$,

$$A_2 = \Sigma_2 + \lambda W(v)^{-2} = \begin{pmatrix} 1/2 + u^{-2} & 1/2 \\ 1/2 & 1/2 + w^{-2} \end{pmatrix}, \; A_2^{-1} = \frac{u^2 + w^2 + 2}{2u^2 w^2}\begin{pmatrix} 1/2 + w^{-2} & -1/2 \\ -1/2 & 1/2 + u^{-2} \end{pmatrix},$$

and direct inversion gives

$$\widehat{\theta}_{2,1} = (A_2^{-1} c_2)_1 = -\frac{\sqrt{2}\rho u^2}{u^2 + w^2 + 2} < 0,$$
$$p_{2,1} := (A_2^{-1} e_2)_1 = -\frac{u^2 w^2}{u^2 + w^2 + 2} < 0,$$
$$r_2^2 = e_2^\top A_2^{-1} e_2 = \frac{w^2(u^2 + 2)}{u^2 + w^2 + 2} > 0.$$

Applying (4) therefore gives

$$\partial_{v_1} q_2(1) = 2u^{-3} p_{2,1} \widehat{\theta}_{2,1} + \frac{\beta_2 u^{-3}}{r_2} p_{2,1}^2 > 0,$$

because the first term is strictly positive and the second is nonnegative.

At $\varepsilon = 0$, action 2 has the zero feature and hence the constant logit $q_2(2) = 0$. Writing $\pi_2 := \pi_2^v(1 \mid x_2, h_1)$, we obtain

$$\partial_{v_1} J_{\gamma,2}^{(0)}(v) = \omega_2^{\gamma,2} \frac{\pi_2(1 - \pi_2)}{\tau_2} \partial_{v_1} q_2(1) > 0,$$

where $J_{\gamma,2}^{(\varepsilon)}$ denotes the objective of the instance whose second action feature is εe_2.

The second-round logit difference, its derivative with respect to v_1, and the induced softmax probability are continuous in ε at zero. Hence there exists $\bar{\varepsilon} > 0$ such that

$$\partial_{v_1} J_{\gamma,2}^{(\varepsilon)}(v) > 0$$

for every $\varepsilon \in (0, \bar{\varepsilon})$. Choosing any such positive ε gives an instance in which all action-feature vectors are nonzero, the stopped-state estimator has zero first coordinate, and the true leader derivative has a strictly positive first coordinate. The strict mismatch persists under sufficiently small admissible perturbations of the features, reward parameter, and interface. If desired, an arbitrarily small inward scaling of the raw features makes all feature-norm constraints strict while preserving the mismatch. □

B.5 Pathwise derivatives at interior and boundary interfaces

All interface-dependent score maps extend smoothly to an open neighborhood of $\mathcal{V}$ contained in $(0, \infty)^d$. Derivatives at the boundary of the closed feasible set are ordinary derivatives of this smooth extension; the closed set matters only for the projected outer update.

When differentiating $V_t(h_{t-1}; v)$ or $b_t(h_{t-1}; v)$, the realized sequence $(x_s, a_s, Y_s)_{s<t}$ is held fixed. This changes the numerical follower state along the same realized history. The probability of observing that history is differentiated separately through

$$\sum_{t=1}^{H} \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1}).$$

C Proof of Theorem 5

Proof. Fix an interface $v \in \mathcal{V}$ and a follower type θ. Write $B_z := 1$, $B_\psi := S/\alpha$, and $\psi^\star := \psi^\star_{\theta,v}$. For each round t, the standard self-normalized linear-regression inequality, applied with failure probability δ_t, gives

$$\mathbb{P}(\mathcal{E}_t^c) \le \delta_t, \qquad \mathcal{E}_t := \left\{ \|\widehat{\psi}_t - \psi^\star\|_{V_t} \le \beta_t \right\}.$$

The displayed confidence width follows from

$$\frac{\det(V_t)}{\det(\lambda I)} \le \left(1 + \frac{(t-1)B_z^2}{d\lambda}\right)^d.$$

No simultaneous event over all horizons is needed: expected regret is bounded round by round, and the failure probabilities are summable.

On $\mathcal{E}_t$, for every action a,

$$\left| \widehat{\psi}_t^\top z_t(a) - (\psi^\star)^\top z_t(a) \right| \le \beta_t \|z_t(a)\|_{V_t^{-1}}.$$

Hence the optimistic score

$$q_t(a) = \widehat{\psi}_t^\top z_t(a) + \beta_t \|z_t(a)\|_{V_t^{-1}}$$

upper bounds the conditional mean reward of every action. Let $p_t(a) \propto \exp(q_t(a)/\tau_t)$. The log-sum-exp identity gives

$$\tau_t \log \sum_{a=1}^K \exp(q_t(a)/\tau_t) = \sum_{a=1}^K p_t(a) q_t(a) + \tau_t \mathcal{H}(p_t),$$

where $\mathcal{H}(p_t) \le \log K$. Since the log-sum-exp is at least $\max_a q_t(a)$,

$$\max_a q_t(a) - \sum_{a=1}^K p_t(a) q_t(a) \le \tau_t \log K.$$

Consequently, conditional on the pre-action information and on $\mathcal{E}_t$,

$$\begin{aligned} &\mathbb{E}_t \left[\max_a \langle \psi^\star, z_t(a) \rangle - \langle \psi^\star, z_t(a_t) \rangle \right] \\ &\quad \le \max_a q_t(a) - \mathbb{E}_t[q_t(a_t)] + 2\beta_t \mathbb{E}_t \|z_t(a_t)\|_{V_t^{-1}} \\ &\quad \le \tau_t \log K + 2\beta_t \mathbb{E}_t \|z_t(a_t)\|_{V_t^{-1}}. \end{aligned}$$

On $\mathcal{E}_t^c$, the instantaneous regret is at most $2S$, because $|\langle \theta, \phi(x,a) \rangle| \le S$. Therefore the total failure contribution is at most

$$2S \sum_{t=1}^H \delta_t \le 2S\delta_0.$$

For every realized action sequence, the elliptical-potential lemma yields

$$\sum_{t=1}^H \left(1 \wedge \|z_t(a_t)\|^2_{V_t^{-1}}\right) \le 2d \log\left(1 + \frac{HB_z^2}{d\lambda}\right).$$

Also, $\|z_t(a_t)\|_{V_t^{-1}} \le B_z/\sqrt{\lambda}$. For every $x \in [0, B_z/\sqrt{\lambda}]$,

$$x \le \left(\frac{B_z}{\sqrt{\lambda}} + 1\right) \sqrt{1 \wedge x^2}.$$

Applying Cauchy–Schwarz therefore gives

$$\begin{aligned} \sum_{t=1}^H \|z_t(a_t)\|_{V_t^{-1}} &\le \left(\frac{B_z}{\sqrt{\lambda}} + 1\right) \sqrt{H \sum_{t=1}^H \left(1 \wedge \|z_t(a_t)\|^2_{V_t^{-1}}\right)} \\ &\le \left(\frac{B_z}{\sqrt{\lambda}} + 1\right) \sqrt{2dH \log\left(1 + \frac{HB_z^2}{d\lambda}\right)}. \end{aligned}$$

Using $\beta_t \le \bar{\beta}_H$, summing the preceding bounds, and taking expectations proves the stated bound for the fixed interface v.

All quantities in the final upper bound depend on the interface class only through the common feature bound B_z and comparator bound B_ψ. Hence the same deterministic bound holds for every fixed $v \in \mathcal{V}$, and taking the supremum does not require a joint confidence event over $\mathcal{V}$. Furthermore, if an interface is selected using completed deployments, then conditioning on the sigma-field generated by those deployments makes the selected interface fixed before the fresh deployment. The same pointwise bound therefore applies conditionally and hence unconditionally.

Finally, under $\tau_t = \tau_0 t^{-1/2}$, $\sum_{t=1}^{H} \tau_t \le 2\tau_0\sqrt{H}$. Combining this with the displayed confidence radius proves Corollary 1. For the specific schedule stated in Theorem 2, both terms in β_t are nondecreasing, so $\bar{\beta}_H = \beta_H$; this gives the main theorem exactly. □

D Proof of Theorem 6

We first prove the identity for a fixed follower type and then average over the population. The proof separates two distinct derivatives: the numerical follower state inside each conditional policy is differentiated pathwise along a fixed history, while the probability of observing that history is differentiated through likelihood-ratio scores.

Proof. Fix a follower type θ. Let

$$h_H = (x_1, a_1, Y_1, \dots, x_H, a_H, Y_H)$$

be a complete deployment and let h_{t-1} denote its prefix before round t. Define

$$\overline{U}_\gamma(h_H) := \sum_{s=1}^{H} \omega_s^{\gamma,H} U_s, \qquad G_t(h_H) := \sum_{s=t}^{H} \omega_s^{\gamma,H} U_s.$$

The type-conditioned objective is

$$J_{\gamma,H}(v;\theta) := \mathbb{E}_{h_H \sim P_v^\theta}[\overline{U}_\gamma(h_H)].$$

Step 1: a common environment measure for each action sequence. For every deterministic action sequence $a_{1:H} \in [K]^H$, Assumption 2 induces a probability measure $\nu^\theta_{a_{1:H}}$ on the context and feedback variables that does not depend on v. The action set and horizon are finite, so

$$J_{\gamma,H}(v;\theta) = \sum_{a_{1:H}\in[K]^H} \int \overline{U}_\gamma(h_H) \prod_{t=1}^{H} \pi_t^v(a_t \mid x_t, h_{t-1})\, d\nu^\theta_{a_{1:H}}.$$

The realized contexts, actions, and rewards are fixed inside each integrand. Thus all pathwise dependence of V_t, b_t, and $\widehat{\psi}_t$ is contained inside the scalar conditional policy factors.

Step 2: differentiating the full-return expectation. For a fixed trajectory,

$$\nabla_v \prod_{t=1}^{H} \pi_t^v(a_t \mid x_t, h_{t-1}) = \left(\prod_{t=1}^{H} \pi_t^v(a_t \mid x_t, h_{t-1})\right) \sum_{t=1}^{H} \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1}).$$

Because $0 \le \overline{U}_\gamma \le 1$ and bounded feedback implies the deterministic score envelopes summarized in Appendix A,

$$\left\| \overline{U}_\gamma(h_H) \sum_{t=1}^{H} \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1}) \right\|_2 \le \sum_{t=1}^{H} G^{\mathrm{raw}}_{\pi,t}.$$

The right-hand side is finite and deterministic for fixed H. Dominated convergence therefore permits differentiation under every integral and the finite sum over action sequences, yielding

$$\nabla_v J_{\gamma,H}(v;\theta) = \mathbb{E}_{P_v^\theta}\left[\overline{U}_\gamma(h_H) \sum_{t=1}^{H} \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1})\right]. \tag{12}$$

Step 3: replacing the full return by return-to-go. For each t, decompose

$$\overline{U}_\gamma(h_H) = P_t(h_{t-1}) + G_t(h_H), \qquad P_t(h_{t-1}) := \sum_{s=1}^{t-1} \omega_s^{\gamma,H} U_s.$$

The prefix return $P_t(h_{t-1})$ is measurable with respect to the pre-action information $\mathcal{I}_t = \sigma(h_{t-1}, x_t)$. Conditional on $\mathcal{I}_t$, $a_t \sim \pi_t^v(\cdot \mid x_t, h_{t-1})$, and therefore

$$\begin{aligned}\mathbb{E}[\ell_t \mid \mathcal{I}_t] &= \sum_{a=1}^{K} \pi_t^v(a \mid x_t, h_{t-1}) \nabla_v \log \pi_t^v(a \mid x_t, h_{t-1}) \\ &= \sum_{a=1}^{K} \nabla_v \pi_t^v(a \mid x_t, h_{t-1}) = 0.\end{aligned}$$

Hence, by the tower property,

$$\mathbb{E}[P_t(h_{t-1})\ell_t] = \mathbb{E}[\mathbb{E}[P_t(h_{t-1})\ell_t]|\mathcal{I}_t] = \mathbb{E}[P_t(h_{t-1})\mathbb{E}[\ell_t \mid \mathcal{I}_t]] = 0.$$

Applying this cancellation to every term in (12) gives

$$\nabla_v J_{\gamma,H}(v;\theta) = \mathbb{E}_{P_v^\theta}\left[\sum_{t=1}^{H} G_t(h_H) \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1})\right].$$

Step 4: averaging over deployment types. The type distribution $\mathcal{P}_\Theta$ is independent of v. The same deterministic score envelope is uniform over all types satisfying the stated feature and feedback bounds. A second application of dominated convergence therefore yields

$$\nabla_v J_{\gamma,H}(v) = \mathbb{E}_{\theta\sim\mathcal{P}_\Theta}[\nabla_v J_{\gamma,H}(v;\theta)],$$

which proves the theorem after substituting the fixed-type identity. □

Online eligibility form and baselines. Let $e_t := \sum_{s=1}^{t} \ell_s$. Exchanging the order of summation gives

$$\sum_{t=1}^{H} G_t^{\gamma,H} \ell_t = \sum_{t=1}^{H} \omega_t^{\gamma,H} U_t e_t,$$

which is (7). Moreover, for any $\mathcal{I}_t$-measurable baseline b_t,

$$\mathbb{E}[b_t \ell_t] = \mathbb{E}[b_t \mathbb{E}[\ell_t \mid \mathcal{I}_t]] = 0.$$

Thus b_t may be subtracted from the return multiplier without changing the expectation.

E Refined Raw-Coordinate Analysis

This appendix proves Proposition 1 and Theorem 3. The central benefit of the raw-coordinate representation is that the entire pathwise dependence of the follower's regression state on the interface is represented by the single matrix $A_t(v) = \Sigma_t + \Lambda(v)$.

E.1 EXPLICIT SENSITIVITY CONSTANTS

Table 2: Explicit deterministic envelopes used in the leader-side analysis.

Symbol	Definition
D_1	$2\lambda\alpha^{-3}$
D_2	$6\lambda\alpha^{-4}$
$B_{\theta,t}$	$R_Y\sqrt{t-1}/\sqrt{\lambda}$
$Q_{1,t}^{\mathrm{raw}}$	$D_1 B_{\theta,t}/\lambda + \beta_t D_1/(2\lambda^{3/2})$
R_2^{raw}	$\lambda^{-1/2}\big(5D_1^2/(4\lambda^2) + D_2/(2\lambda)\big)$
$Q_{2,t}^{\mathrm{raw}}$	$B_{\theta,t}\big(2D_1^2/\lambda^2 + D_2/\lambda\big) + \beta_t R_2^{\mathrm{raw}}$
$G_{\pi,t}^{\mathrm{raw}}$	$2\sqrt{d}\,Q_{1,t}^{\mathrm{raw}}/\tau_t$
$H_{\pi,t}^{\mathrm{raw}}$	$d\big(2Q_{2,t}^{\mathrm{raw}}/\tau_t + 2(Q_{1,t}^{\mathrm{raw}})^2/\tau_t^2\big)$
$\overline{W}_t^{\gamma,H}$	$\sum_{s=t}^{H}\omega_s^{\gamma,H}$
$\mathcal{A}_{\gamma,H}$	$\sum_{t=1}^{H}\overline{W}_t^{\gamma,H}\big[(t+\beta_t^2)/\tau_t^2 + (\sqrt{t}+\beta_t)/\tau_t\big]$
$\mathcal{C}_{\gamma,H}$	$\sum_{t=1}^{H}\overline{W}_t^{\gamma,H}(t+\beta_t^2)/\tau_t^2$

Table 3: Dependencies suppressed by the asymptotic notation in the main text. Logarithmic factors and universal numerical constants are omitted.

Quantity	Leading scale shown in the main text	Important suppressed dependencies
$G_{\pi,t}^{\mathrm{raw}}$	$\sqrt{d}(\sqrt{t}+\beta_t)/\tau_t$	Primitive dependence up to α^{-3}; up to α^{-4} after class-wide calibration
$H_{\pi,t}^{\mathrm{raw}}$	$d(t+\beta_t^2)/\tau_t^2 + d(\sqrt{t}+\beta_t)/\tau_t$	Primitive dependence up to α^{-6}; worst-case α^{-8} after class-wide calibration
$L_{\gamma,H}^{\mathrm{raw}}$	$d\mathcal{A}_{\gamma,H}$	Worst-case α^{-8} and the exact discounted tail sums
$\sigma_{\gamma,H}^2$	$d\mathcal{C}_{\gamma,H}$	Worst-case α^{-8} through squared first-order scores
Follower regret	$d\sqrt{H} + (\log K)\sum_{t\le H}\tau_t$	Comparator radius S/α, noise scale σ_Y, ridge λ, and confidence logarithms

The tables summarize explicit valid envelopes, not minimal or sharp constants. Their derivations appear below.

E.2 PROOF OF PROPOSITION 1

Proof. Fix an interface v and a realized history before round t. Write $W := W(v) = \mathrm{Diag}(v)$ and

$$\Sigma_t := \sum_{s<t}\phi_s\phi_s^\top, \qquad c_t := \sum_{s<t}\phi_s Y_s,$$

where $\phi_s := \phi(x_s, a_s)$. Since $z_s(a_s) = W\phi_s$ and $W = W^\top$,

$$V_t = \lambda I + W\Sigma_t W, \qquad b_t = Wc_t.$$

Define

$$\Lambda(v) := \lambda W^{-2}, \qquad A_t(v) := \Sigma_t + \Lambda(v).$$

Because $W\Lambda(v)W = \lambda I$,

$$V_t = WA_t(v)W, \qquad V_t^{-1} = W^{-1}A_t(v)^{-1}W^{-1}.$$

Therefore,

$$\widehat{\psi}_t = V_t^{-1}b_t = W^{-1}A_t(v)^{-1}c_t.$$

Let $\widehat{\theta}_t(v) := A_t(v)^{-1}c_t$. Then $W\widehat{\psi}_t = \widehat{\theta}_t(v)$, and for every candidate action a,

$$\widehat{\psi}_t^\top z_v(x_t, a) = \widehat{\theta}_t(v)^\top \phi(x_t, a),$$
$$z_v(x_t, a)^\top V_t^{-1} z_v(x_t, a) = \phi(x_t, a)^\top A_t(v)^{-1}\phi(x_t, a).$$

Thus the visible and raw representations have identical predictions, confidence radii, optimistic scores, and action probabilities along every realized history. Induction over rounds then gives identical trajectory laws.

Finally,

$$\Lambda(v) = \mathrm{Diag}(\lambda v_1^{-2}, \ldots, \lambda v_d^{-2}).$$

For each coordinate, the map $v_i \mapsto \lambda v_i^{-2}$ is a bijection from $[\alpha, 1]$ to $[\lambda, \lambda/\alpha^2]$. Taking the Cartesian product proves the stated coordinate-wise bijection. □

The proposition is an exact characterization of the accessible intervention. If the leader could directly set the follower's raw-coordinate diagonal ridge, the same trajectory law could be reproduced without an explicit interface. Under the access model of the paper, however, the feature boundary is the leader's admissible control surface, while the follower implementation is fixed.

Remark 3 (General linear visibility maps)**.** *The same algebra extends to a fixed invertible matrix A. With $z_A(x, a) = A\phi(x, a)$,*

$$V_t = \lambda I + A\Sigma_t A^\top = A\left(\Sigma_t + \lambda A^{-1}A^{-\top}\right)A^\top,$$

so the induced raw-coordinate ridge is

$$\Lambda(A) = \lambda A^{-1}A^{-\top}.$$

Uniform spectral constraints are required to retain bounded comparator norms and derivative envelopes. The main theory restricts attention to positive coordinate-wise interfaces because they preserve the semantics of fixed feature channels, are directly auditable, and keep the leader parameter dimension equal to d.

E.3 Derivatives of the diagonal ridge matrix

For

$$\Lambda(v) = \mathrm{Diag}(\lambda v_1^{-2}, \ldots, \lambda v_d^{-2}),$$

write $\lambda_i(v_i) := \lambda v_i^{-2}$. Then

$$\lambda_i'(v_i) = -2\lambda v_i^{-3}, \qquad \lambda_i''(v_i) = 6\lambda v_i^{-4}.$$

Since $v_i \geq \alpha$,

$$|\lambda_i'(v_i)| \leq 2\lambda\alpha^{-3} =: D_1, \qquad |\lambda_i''(v_i)| \leq 6\lambda\alpha^{-4} =: D_2.$$

Define

$$D_i(v) := \partial_{v_i}\Lambda(v) = \lambda_i'(v_i)e_i e_i^\top,$$

and

$$E_{ij}(v) := \partial^2_{v_i v_j}\Lambda(v).$$

Then

$$E_{ii}(v) = \lambda_i''(v_i)e_i e_i^\top, \qquad E_{ij}(v) = 0 \quad \text{for } i \neq j,$$

and uniformly on $\mathcal{V}$,

$$\|D_i(v)\|_{\mathrm{op}} \leq D_1, \qquad \|E_{ij}(v)\|_{\mathrm{op}} \leq D_2.$$

E.4 A DETERMINISTIC BOUND ON THE RAW-COORDINATE ESTIMATE

Let X_t be the matrix whose rows are $\phi_s^\top$ for $s < t$, and let Y_t^{vec} collect the corresponding realized rewards. Then

$$\Sigma_t = X_t^\top X_t, \qquad c_t = X_t^\top Y_t^{\mathrm{vec}}, \qquad A_t = X_t^\top X_t + \Lambda(v).$$

Since $\Lambda(v) \succeq \lambda I$, we have $A_t \succeq \lambda I$ and $A_t \succeq X_t^\top X_t$. The nonzero eigenvalues of $X_t A_t^{-1} X_t^\top$ equal those of $A_t^{-1/2} X_t^\top X_t A_t^{-1/2} \preceq I$, so

$$\|X_t A_t^{-1} X_t^\top\|_{\mathrm{op}} \le 1.$$

Therefore,

$$\begin{aligned}\|\widehat{\theta}_t\|_{A_t}^2 &= (Y_t^{\mathrm{vec}})^\top X_t A_t^{-1} X_t^\top Y_t^{\mathrm{vec}} \\ &\le \|Y_t^{\mathrm{vec}}\|_2^2 \le (t-1) R_Y^2.\end{aligned}$$

Using $A_t \succeq \lambda I$ gives

$$\|\widehat{\theta}_t\|_2 \le \frac{1}{\sqrt{\lambda}} \|\widehat{\theta}_t\|_{A_t} \le B_{\theta,t} := \frac{R_Y \sqrt{t-1}}{\sqrt{\lambda}}.$$

E.5 FIRST-ORDER RAW-COORDINATE DERIVATIVES

Fix a candidate raw feature $\phi := \phi(x_t, a)$ and write

$$p_t(a) := A_t(v)^{-1}\phi, \qquad r_t(a) := \sqrt{\phi^\top A_t(v)^{-1}\phi}.$$

Pathwise, Σ_t and c_t are fixed with respect to v. Hence

$$\partial_{v_i} \widehat{\theta}_t = -A_t^{-1} D_i \widehat{\theta}_t.$$

Consequently,

$$\left|\partial_{v_i} (\widehat{\theta}_t^\top \phi)\right| = \left|\phi^\top A_t^{-1} D_i \widehat{\theta}_t\right| \le \frac{D_1 B_{\theta,t}}{\lambda}.$$

For $s_t(a) := r_t(a)^2 = \phi^\top A_t^{-1}\phi$,

$$\partial_{v_i} s_t(a) = -\phi^\top A_t^{-1} D_i A_t^{-1} \phi.$$

Since D_i is supported on coordinate i,

$$|\partial_{v_i} s_t(a)| \le D_1 (e_i^\top A_t^{-1}\phi)^2.$$

Cauchy–Schwarz in the A_t^{-1} inner product gives

$$(e_i^\top A_t^{-1}\phi)^2 \le (e_i^\top A_t^{-1} e_i)(\phi^\top A_t^{-1}\phi) \le \frac{1}{\lambda} r_t(a)^2.$$

Thus, whenever $r_t(a) > 0$,

$$|\partial_{v_i} r_t(a)| = \frac{|\partial_{v_i} s_t(a)|}{2 r_t(a)} \le \frac{D_1}{2\lambda} r_t(a) \le \frac{D_1}{2\lambda^{3/2}}.$$

Because $A_t(v)^{-1} \succ 0$, $r_t(a) = 0$ if and only if $\phi = 0$. In that case $r_t(a)$ is identically zero as a function of v, so the same bound holds with derivative zero.

Combining the prediction and radius derivatives,

$$|\partial_{v_i} q_t^v(a)| \le Q_{1,t}^{\mathrm{raw}} := \frac{D_1 B_{\theta,t}}{\lambda} + \frac{\beta_t D_1}{2\lambda^{3/2}}.$$

Substituting the exact rank-one form $D_i = -2\lambda v_i^{-3} e_i e_i^\top$ gives

$$\partial_{v_i} q_t^v(a) = 2\lambda v_i^{-3} p_{t,i}(a) \widehat{\theta}_{t,i}(v) + \frac{\beta_t \lambda v_i^{-3}}{r_t(a)} p_{t,i}(a)^2,$$

with the second term defined as zero when $r_t(a) = 0$. This is (4).

The softmax score identity then gives

$$\|\nabla_v \log \pi_t^v(a \mid x_t, h_{t-1})\|_2 \le G_{\pi,t}^{\mathrm{raw}} := \frac{2\sqrt{d}\, Q_{1,t}^{\mathrm{raw}}}{\tau_t}.$$

E.6 SECOND-ORDER RAW-COORDINATE DERIVATIVES

Differentiating $\partial_{v_i}\widehat{\theta}_t = -A_t^{-1}D_i\widehat{\theta}_t$ gives

$$\partial^2_{v_iv_j}\widehat{\theta}_t = A_t^{-1}D_jA_t^{-1}D_i\widehat{\theta}_t + A_t^{-1}D_iA_t^{-1}D_j\widehat{\theta}_t - A_t^{-1}E_{ij}\widehat{\theta}_t.$$

Using $\|A_t^{-1}\|_{\mathrm{op}} \le 1/\lambda$ and $\|\phi\|_2 \le 1$,

$$\left|\partial^2_{v_iv_j}(\widehat{\theta}_t^\top\phi)\right| \le B_{\theta,t}\left(\frac{2D_1^2}{\lambda^2} + \frac{D_2}{\lambda}\right).$$

For $s_t(a) = \phi^\top A_t^{-1}\phi$,

$$\begin{aligned}\partial^2_{v_iv_j}s_t(a) = {}& \phi^\top A_t^{-1}D_jA_t^{-1}D_iA_t^{-1}\phi + \phi^\top A_t^{-1}D_iA_t^{-1}D_jA_t^{-1}\phi \\ & - \phi^\top A_t^{-1}E_{ij}A_t^{-1}\phi.\end{aligned}$$

To bound the first term, let $x := A_t^{-1/2}\phi$ and $B_i := A_t^{-1/2}D_iA_t^{-1/2}$. Then

$$\begin{aligned}\left|\phi^\top A_t^{-1}D_jA_t^{-1}D_iA_t^{-1}\phi\right| &= |x^\top B_jB_ix| \\ &\le \|B_j\|_{\mathrm{op}}\|B_i\|_{\mathrm{op}}\|x\|_2^2 \\ &\le \frac{D_1^2}{\lambda^2}r_t(a)^2.\end{aligned}$$

The second product term obeys the same bound, while

$$\left|\phi^\top A_t^{-1}E_{ij}A_t^{-1}\phi\right| \le \frac{D_2}{\lambda}r_t(a)^2.$$

Therefore,

$$|\partial^2_{v_iv_j}s_t(a)| \le \left(\frac{2D_1^2}{\lambda^2} + \frac{D_2}{\lambda}\right)r_t(a)^2.$$

If $r_t(a) > 0$, the chain rule and $|\partial_{v_i}s_t(a)| \le (D_1/\lambda)r_t(a)^2$ yield

$$\begin{aligned}|\partial^2_{v_iv_j}r_t(a)| &\le \frac{|\partial^2_{v_iv_j}s_t(a)|}{2r_t(a)} + \frac{|\partial_{v_i}s_t(a)\partial_{v_j}s_t(a)|}{4r_t(a)^3} \\ &\le r_t(a)\left(\frac{5D_1^2}{4\lambda^2} + \frac{D_2}{2\lambda}\right) \\ &\le R_2^{\mathrm{raw}} := \frac{1}{\sqrt{\lambda}}\left(\frac{5D_1^2}{4\lambda^2} + \frac{D_2}{2\lambda}\right).\end{aligned}$$

If $\phi = 0$, the radius and both derivatives are zero, so the same bound applies.

Hence

$$|\partial^2_{v_iv_j}q_t^v(a)| \le Q_{2,t}^{\mathrm{raw}} := B_{\theta,t}\left(\frac{2D_1^2}{\lambda^2} + \frac{D_2}{\lambda}\right) + \beta_tR_2^{\mathrm{raw}}.$$

For each Hessian entry, the exact softmax formula and $|\partial_{v_i}q_t^v(a)| \le Q_{1,t}^{\mathrm{raw}}$ give

$$\left|\partial^2_{v_iv_j}\log\pi_t^v(a \mid x_t, h_{t-1})\right| \le \frac{2Q_{2,t}^{\mathrm{raw}}}{\tau_t} + \frac{2(Q_{1,t}^{\mathrm{raw}})^2}{\tau_t^2}.$$

A $d \times d$ matrix has operator norm at most d times its largest absolute entry, so

$$\|\nabla_v^2\log\pi_t^v(a \mid x_t, h_{t-1})\|_{\mathrm{op}} \le H_{\pi,t}^{\mathrm{raw}} := d\left(\frac{2Q_{2,t}^{\mathrm{raw}}}{\tau_t} + \frac{2(Q_{1,t}^{\mathrm{raw}})^2}{\tau_t^2}\right).$$

Since $B_{\theta,t} = O(\sqrt{t})$,

$$Q_{1,t}^{\mathrm{raw}} = O(\sqrt{t} + \beta_t), \qquad Q_{2,t}^{\mathrm{raw}} = O(\sqrt{t} + \beta_t),$$

with constants depending only on α, λ, and R_Y. Consequently,

$$G^{\mathrm{raw}}_{\pi,t} = O\left(\frac{\sqrt{d}(\sqrt{t}+\beta_t)}{\tau_t}\right),$$

and

$$H^{\mathrm{raw}}_{\pi,t} = O\left(\frac{d(t+\beta_t^2)}{\tau_t^2} + \frac{d(\sqrt{t}+\beta_t)}{\tau_t}\right).$$

This proves Theorem 3.

Dependence on the interface floor. Holding β_t fixed, the primitive first-order dependence reaches α^{-3} and the second-order terms reach α^{-6}. Under the class-wide confidence calibration, β_t contains the additional factor S/α. Thus $Q^{\mathrm{raw}}_{1,t}$ has worst-case dependence α^{-4}, $Q^{\mathrm{raw}}_{2,t}$ has worst-case dependence α^{-7}, and squared-score, policy-Hessian, smoothness, and variance bounds have worst-case dependence α^{-8}.

F Discounted Fisher-Refined Smoothness

This appendix proves Lemma 7 and Theorem 9. The key subtlety is that the utility multiplier is generally correlated with the trajectory-score increments. Therefore the weighted cross-time products cannot be removed by an independence argument. The proof first dominates the weighted score product by the unweighted Fisher information and only then applies martingale cancellation.

F.1 Proof of Lemma 7

Proof. Let

$$S_s := \sum_{t=1}^{s} \ell_t.$$

Because $0 \le R \le 1$, for every deterministic vector $u \in \mathbb{R}^d$,

$$u^\top \mathbb{E}[RS_sS_s^\top]u = \mathbb{E}[R(u^\top S_s)^2] \le \mathbb{E}[(u^\top S_s)^2] = u^\top \mathbb{E}[S_sS_s^\top]u.$$

Thus, in the Loewner order,

$$0 \preceq \mathbb{E}[RS_sS_s^\top] \preceq \mathbb{E}[S_sS_s^\top].$$

Expanding the unweighted matrix gives

$$\mathbb{E}[S_sS_s^\top] = \sum_{t=1}^{s} \mathbb{E}[\ell_t\ell_t^\top] + \sum_{1\le i<j\le s} \mathbb{E}[\ell_i\ell_j^\top] + \sum_{1\le j<i\le s} \mathbb{E}[\ell_i\ell_j^\top].$$

For $i < j$, ℓ_i is measurable with respect to the pre-action information $\mathcal{I}_j$, and $\mathbb{E}[\ell_j \mid \mathcal{I}_j] = 0$. Hence

$$\mathbb{E}[\ell_i\ell_j^\top] = \mathbb{E}\left[\ell_i \mathbb{E}[\ell_j^\top \mid \mathcal{I}_j]\right] = 0.$$

The terms with $j < i$ vanish by transposition. Therefore,

$$\mathbb{E}[S_sS_s^\top] = \sum_{t=1}^{s} \mathbb{E}[\ell_t\ell_t^\top].$$

Consequently,

$$\begin{aligned}\|\mathbb{E}[RS_sS_s^\top]\|_{\mathrm{op}} &\le \|\mathbb{E}[S_sS_s^\top]\|_{\mathrm{op}} \\ &\le \sum_{t=1}^{s} \|\mathbb{E}[\ell_t\ell_t^\top]\|_{\mathrm{op}} \\ &\le \sum_{t=1}^{s} \mathbb{E}\|\ell_t\|_2^2 \le \sum_{t=1}^{s} G^2_{\pi,t}.\end{aligned}$$

The proof does not claim that $\mathbb{E}[R\ell_i\ell_j^\top] = 0$; this is precisely why it remains valid when R depends on the trajectory. □

F.2 Proof of Theorem 9

Proof. We first fix a follower type θ and then average over $\mathcal{P}_\Theta$. Write

$$F_{\gamma,H}(v;\theta) := \sum_{s=1}^{H} \omega_s^{\gamma,H} \mathbb{E}[U_s].$$

Fix $s \le H$ and let

$$S_s(v) := \sum_{t=1}^{s} \nabla_v \log \pi_t^v(a_t \mid x_t, h_{t-1}).$$

The deterministic score and score-Hessian envelopes justify differentiating twice under the expectation. Since U_s has no explicit dependence on v and the environment kernels are interface independent, the second-order score identity gives

$$\nabla_v^2 \mathbb{E}[U_s] = \mathbb{E}\left[U_s \sum_{t=1}^{s} \nabla_v^2 \log \pi_t^v(a_t \mid x_t, h_{t-1})\right] + \mathbb{E}[U_s S_s(v) S_s(v)^\top].$$

For the first term,

$$\left\| \mathbb{E}\left[U_s \sum_{t=1}^{s} \nabla_v^2 \log \pi_t^v(a_t \mid x_t, h_{t-1})\right] \right\|_{\mathrm{op}} \le \sum_{t=1}^{s} H_{\pi,t},$$

because $0 \le U_s \le 1$. For the score-product term, applying Lemma 7 with $R = U_s$ yields

$$\|\mathbb{E}[U_s S_s(v) S_s(v)^\top]\|_{\mathrm{op}} \le \sum_{t=1}^{s} G_{\pi,t}^2.$$

Therefore,

$$\|\nabla_v^2 \mathbb{E}[U_s]\|_{\mathrm{op}} \le \sum_{t=1}^{s} \left(H_{\pi,t} + G_{\pi,t}^2\right).$$

Multiplying by $\omega_s^{\gamma,H}$, summing over s, and exchanging the two finite sums gives

$$\begin{aligned}\|\nabla_v^2 F_{\gamma,H}(v;\theta)\|_{\mathrm{op}} &\le \sum_{s=1}^{H} \omega_s^{\gamma,H} \sum_{t=1}^{s} \left(H_{\pi,t} + G_{\pi,t}^2\right) \\ &= \sum_{t=1}^{H} \overline{W}_t^{\gamma,H} \left(H_{\pi,t} + G_{\pi,t}^2\right).\end{aligned}$$

The same deterministic bound holds for every admissible type, so averaging over $\mathcal{P}_\Theta$ preserves it. The resulting uniform Hessian bound implies L_J-smoothness on $\mathcal{V}$. □

Why the utility-time decomposition is useful. Differentiating the tail-return gradient directly produces products between a weighted tail score and the full trajectory score. Decomposing the objective as $\sum_s \omega_s^{\gamma,H} \mathbb{E}[U_s]$ applies Fisher domination before the sum over utility times and yields the exact tail mass $\overline{W}_t^{\gamma,H}$ rather than a square-of-sums bound.

G Estimator Unbiasedness and Moment Bounds

This appendix proves Theorems 7 and 8, Lemma 6, and then combines the raw-coordinate envelopes with Theorem 9 to prove Theorem 3.

G.1 Proof of Theorem 7

Proof. Fix m and condition on $\mathcal{F}_{m-1}$. The iterate v_m is then fixed, while each episode in the current mini-batch has the population trajectory law at v_m. By Theorem 6, for every b,

$$\mathbb{E}[Z_{m,b} \mid \mathcal{F}_{m-1}] = \nabla J_{\gamma,H}(v_m).$$

Averaging over the B episodes preserves this conditional expectation, so

$$\mathbb{E}[\widehat{g}_m \mid \mathcal{F}_{m-1}] = \nabla J_{\gamma,H}(v_m).$$

□

G.2 Proof of Theorem 8

Proof. Suppress the outer indices and write

$$e_s := \sum_{t=1}^{s} \ell_t.$$

By the eligibility identity (7),

$$Z = \sum_{s=1}^{H} \omega_s^{\gamma,H} U_s e_s.$$

The nonnegative coefficients $a_s := \omega_s^{\gamma,H} U_s$ satisfy $\sum_s a_s \le 1$. Assigning the remaining mass to the zero vector and using convexity of the squared norm gives, pathwise,

$$\begin{aligned} \|Z\|_2^2 &\le \sum_{s=1}^{H} \omega_s^{\gamma,H} U_s \|e_s\|_2^2 \\ &\le \sum_{s=1}^{H} \omega_s^{\gamma,H} \|e_s\|_2^2. \end{aligned}$$

Conditional on $\mathcal{F}_{m-1}$, the score increments form a martingale-difference sequence with respect to the pre-action filtration. Thus, for $i < j$,

$$\mathbb{E}[\ell_i^\top \ell_j \mid \mathcal{F}_{m-1}] = \mathbb{E}\left[\ell_i^\top \mathbb{E}[\ell_j \mid \mathcal{I}_j, \mathcal{F}_{m-1}] \mid \mathcal{F}_{m-1}\right] = 0.$$

Therefore,

$$\begin{aligned} \mathbb{E}[\|e_s\|_2^2 \mid \mathcal{F}_{m-1}] &= \sum_{t=1}^{s} \mathbb{E}[\|\ell_t\|_2^2 \mid \mathcal{F}_{m-1}] \\ &\le \sum_{t=1}^{s} G_{\pi,t}^2. \end{aligned}$$

Substituting this bound and exchanging the finite sums gives

$$\begin{aligned} \mathbb{E}[\|Z\|_2^2 \mid \mathcal{F}_{m-1}] &\le \sum_{s=1}^{H} \omega_s^{\gamma,H} \sum_{t=1}^{s} G_{\pi,t}^2 \\ &= \sum_{t=1}^{H} \overline{W}_t^{\gamma,H} G_{\pi,t}^2 = V_{\mathrm{est}}. \end{aligned}$$

The conditional variance is at most the conditional second moment.

For the almost-sure bound, use the return-to-go form and $0 \le G_t^{\gamma,H} \le \overline{W}_t^{\gamma,H}$:

$$\|Z\|_2 \le \sum_{t=1}^{H} G_t^{\gamma,H} \|\ell_t\|_2 \le \sum_{t=1}^{H} \overline{W}_t^{\gamma,H} G_{\pi,t} = C_{\mathrm{est}}.$$

□

G.3 Proof of Lemma 6

Proof. Let

$$\overline{Z}_{m,b} := Z_{m,b} - \mathbb{E}[Z_{m,b} \mid \mathcal{F}_{m-1}].$$

Conditional independence and conditional centering imply that all cross terms vanish. Hence

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_m - \nabla J_{\gamma,H}(v_m)\|_2^2 \middle| \mathcal{F}_{m-1}\right] &= \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}[\|\overline{Z}_{m,b}\|_2^2 \mid \mathcal{F}_{m-1}] \\ &\le \frac{\sigma_{\gamma,H}^2}{B}.\end{aligned}$$

□

G.4 Smoothness and estimator-moment bounds

Proof of the remaining claims in Theorem 3. Apply Theorem 9 with the raw-coordinate envelopes established above. This gives

$$L_{\gamma,H}^{\mathrm{raw}} = \sum_{t=1}^{H} \overline{W}_t^{\gamma,H}\left(H_{\pi,t}^{\mathrm{raw}} + (G_{\pi,t}^{\mathrm{raw}})^2\right).$$

Using

$$Q_{1,t}^{\mathrm{raw}} = O(\sqrt{t} + \beta_t), \qquad Q_{2,t}^{\mathrm{raw}} = O(\sqrt{t} + \beta_t),$$

we obtain

$$H_{\pi,t}^{\mathrm{raw}} + (G_{\pi,t}^{\mathrm{raw}})^2 = \widetilde{O}\left(d\frac{t+\beta_t^2}{\tau_t^2} + d\frac{\sqrt{t}+\beta_t}{\tau_t}\right),$$

where the suppressed constants are independent of t, H, and the realized trajectory. Therefore,

$$L_{\gamma,H}^{\mathrm{raw}} = \widetilde{O}\left(d\mathcal{A}_{\gamma,H}\right).$$

Theorem 8, with $G_{\pi,t} = G_{\pi,t}^{\mathrm{raw}}$, gives

$$\begin{aligned}\mathbb{E}\|\widehat{g}(v)\|_2^2 &\le \sum_{t=1}^{H}\overline{W}_t^{\gamma,H}(G_{\pi,t}^{\mathrm{raw}})^2 \\ &= \widetilde{O}\left(d\mathcal{C}_{\gamma,H}\right).\end{aligned}$$

Since $\mathbb{E}\|\widehat{g}(v) - \nabla J_{\gamma,H}(v)\|_2^2 \le \mathbb{E}\|\widehat{g}(v)\|_2^2$, the stated variance bound follows. □

The variance bound above is a tail-weighted sum of squared score envelopes, not the square of their sum. This distinction is essential for the fixed discount specialization in Appendix I.

H Outer-Loop Convergence

This appendix proves the averaged–damped stationarity guarantee and the direct constant-step comparison. Let $F := J_{\gamma,H}$, let $L := L_{\gamma,H}^{\mathrm{raw}}$, and define $F^\star := \sup_{v\in\mathcal{V}} F(v)$. The initial filtration $\mathcal{F}_0$ contains v_1 and the deterministic initial state $s_0 = 0$, but no randomness from subsequent optimization batches. At outer iteration m, $\mathcal{F}_{m-1}$ contains all information revealed before the current fresh batch, including v_m and s_{m-1}.

H.1 Proof of Lemma 8

Proof. The projection optimality condition gives

$$\langle v + \eta \nabla F(v) - y, z - y \rangle \le 0, \qquad \forall z \in \mathcal{V}.$$

Taking $z = v$ yields

$$\langle \nabla F(v), y - v \rangle \ge \eta^{-1} \|y - v\|_2^2.$$

By L-smoothness,

$$\begin{aligned} F(y) - F(v) &\ge \langle \nabla F(v), y - v \rangle - \frac{L}{2} \|y - v\|_2^2 \\ &\ge \left(\eta^{-1} - \frac{L}{2} \right) \|y - v\|_2^2. \end{aligned}$$

If $\eta \le 1/L$, the last expression is at least $(2\eta)^{-1} \|y - v\|_2^2 = (\eta/2) \|\mathcal{G}_\eta(v)\|_2^2$. □

H.2 Proof of Theorem 4

Proof. Define

$$\xi_m := \widehat{g}_m - \nabla F(v_m), \qquad e_m := s_m - \nabla F(v_m).$$

By Theorems 7 and 8 and Lemma 6,

$$\mathbb{E}[\xi_m \mid \mathcal{F}_{m-1}] = 0, \qquad \mathbb{E}[\|\xi_m\|_2^2 \mid \mathcal{F}_{m-1}] \le \frac{\sigma_{\gamma,H}^2}{B}.$$

Let

$$\widetilde{v}_{m+1} := \Pi_{\mathcal{V}}(v_m + \eta s_m), \qquad u_m := \eta^{-1}(\widetilde{v}_{m+1} - v_m),$$

so that

$$v_{m+1} = v_m + \mu \eta u_m.$$

Step 1: ascent controlled by the tracking error. Projection optimality with comparison point v_m gives

$$\langle s_m, u_m \rangle \ge \|u_m\|_2^2.$$

Since $\nabla F(v_m) = s_m - e_m$, L-smoothness gives

$$\begin{aligned} F(v_{m+1}) - F(v_m) &\ge \mu\eta \langle \nabla F(v_m), u_m \rangle - \frac{L\mu^2\eta^2}{2} \|u_m\|_2^2 \\ &\ge \mu\eta \left(1 - \frac{L\mu\eta}{2} \right) \|u_m\|_2^2 - \mu\eta \langle e_m, u_m \rangle. \end{aligned}$$

Using $-\langle e_m, u_m \rangle \ge -\|u_m\|_2^2/4 - \|e_m\|_2^2$ and $L\mu\eta \le 1/8$,

$$F(v_{m+1}) - F(v_m) \ge \frac{\mu\eta}{2} \|u_m\|_2^2 - \mu\eta \|e_m\|_2^2.$$

Summing from $m = 1$ to M, taking expectations, and using $F(v_{M+1}) \le F^\star$ gives

$$\overline{U}_M := \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|u_m\|_2^2 \le \frac{2\Delta_1}{\mu\eta M} + 2\overline{E}_M, \qquad \overline{E}_M := \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|e_m\|_2^2. \tag{13}$$

Projection nonexpansiveness also gives

$$\|\mathcal{G}_\eta(v_m) - u_m\|_2 \le \|\nabla F(v_m) - s_m\|_2 = \|e_m\|_2,$$

and hence

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\mathcal{G}_\eta(v_m)\|_2^2 \le 2\overline{U}_M + 2\overline{E}_M. \tag{14}$$

Step 2: gradient-tracking recursion. Set $v_0 := v_1$ and $e_0 := s_0 - \nabla F(v_1)$. For $m \geq 1$,

$$\begin{aligned} e_m &= (1-\vartheta)s_{m-1} + \vartheta\widehat{g}_m - \nabla F(v_m) \\ &= (1-\vartheta)(e_{m-1} + \delta_m) + \vartheta\xi_m, \end{aligned}$$

where

$$\delta_m := \nabla F(v_{m-1}) - \nabla F(v_m), \qquad \delta_1 = 0.$$

Conditional on $\mathcal{F}_{m-1}$, the vector $e_{m-1} + \delta_m$ is measurable and ξ_m is centered. Therefore,

$$\begin{aligned} \mathbb{E}[\|e_m\|_2^2 \mid \mathcal{F}_{m-1}] = (1-\vartheta)^2\|e_{m-1} + \delta_m\|_2^2 \\ + \vartheta^2\mathbb{E}[\|\xi_m\|_2^2 \mid \mathcal{F}_{m-1}]. \end{aligned}$$

For $0 < \vartheta \leq 1/2$, Young's inequality gives

$$(1-\vartheta)^2\|a+b\|_2^2 \leq \left(1 - \frac{\vartheta}{2}\right)\|a\|_2^2 + \frac{3}{\vartheta}\|b\|_2^2.$$

Consequently,

$$\mathbb{E}\|e_m\|_2^2 \leq \left(1 - \frac{\vartheta}{2}\right)\mathbb{E}\|e_{m-1}\|_2^2 + \vartheta^2\frac{\sigma_{\gamma,H}^2}{B} + \frac{3}{\vartheta}\mathbb{E}\|\delta_m\|_2^2. \tag{15}$$

Let $r := 1 - \vartheta/2$. Summing the recursion and using

$$\sum_{m=1}^{M} E_m \leq r\left(E_0 + \sum_{m=1}^{M-1} E_m\right) + \sum_{m=1}^{M} b_m$$

with $E_m := \mathbb{E}\|e_m\|_2^2$ gives

$$\overline{E}_M \leq \frac{2\mathbb{E}\|e_0\|_2^2}{\vartheta M} + 2\vartheta\frac{\sigma_{\gamma,H}^2}{B} + \frac{6}{\vartheta^2 M}\sum_{m=1}^{M}\mathbb{E}\|\delta_m\|_2^2. \tag{16}$$

By smoothness and the damped update,

$$\|\delta_m\|_2^2 \leq L^2\|v_m - v_{m-1}\|_2^2 = L^2\mu^2\eta^2\|u_{m-1}\|_2^2.$$

With $u_0 := 0$,

$$\overline{E}_M \leq A_M + \kappa\overline{U}_M, \tag{17}$$

where

$$A_M := \frac{2\mathbb{E}\|e_0\|_2^2}{\vartheta M} + 2\vartheta\frac{\sigma_{\gamma,H}^2}{B}, \qquad \kappa := \frac{6L^2\mu^2\eta^2}{\vartheta^2}.$$

Since $\mu \leq \vartheta$ and $\eta \leq 1/(8L)$,

$$\kappa \leq \frac{6}{64} < \frac{1}{8}.$$

Substituting (13) into (17) gives

$$\overline{E}_M \leq A_M + \frac{2\kappa\Delta_1}{\mu\eta M} + 2\kappa\overline{E}_M.$$

The final term is absorbed because $2\kappa < 1/4$, yielding

$$\overline{E}_M \leq C\left(\frac{\mathbb{E}\|e_0\|_2^2}{\vartheta M} + \vartheta\frac{\sigma_{\gamma,H}^2}{B} + \frac{\Delta_1}{\mu\eta M}\right) \tag{18}$$

for a universal numerical constant C.

Step 3: projected stationarity. Combining (13), (14), and (18) yields

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\mathcal{G}_\eta(v_m)\|_2^2 \le C\left(\frac{\Delta_1}{\mu\eta M}+\frac{\mathbb{E}\|e_0\|_2^2}{\vartheta M}+\vartheta\frac{\sigma_{\gamma,H}^2}{B}\right), \tag{19}$$

for a universal numerical constant C. If $\widehat{m}\sim\mathrm{Unif}\{1,\dots,M\}$ independently of the optimization trajectory, then

$$\mathbb{E}\|\mathcal{G}_\eta(v_{\widehat{m}})\|_2^2=\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\mathcal{G}_\eta(v_m)\|_2^2.$$

For $M\ge 4$, choosing $\vartheta=\mu=M^{-1/2}$ satisfies $\vartheta\le 1/2$. Since $\eta=c_\eta/L$ with $0<c_\eta\le 1/8$,

$$\frac{\Delta_1}{\mu\eta M}=O\left(\frac{L\Delta_1}{\sqrt{M}}\right).$$

Moreover, because $s_0=0$,

$$e_0=-\nabla F(v_1),\qquad \mathbb{E}\|e_0\|_2^2=\|\nabla F(v_1)\|_2^2\le G_{\gamma,H}^2,$$

where

$$G_{\gamma,H}:=\sup_{v\in\mathcal{V}}\|\nabla J_{\gamma,H}(v)\|_2.$$

Therefore, (19) gives

$$\mathbb{E}\|\mathcal{G}_\eta(\widehat{v})\|_2^2\le\widetilde{O}\left(\frac{L\Delta_1+G_{\gamma,H}^2}{\sqrt{M}}+\frac{\sigma_{\gamma,H}^2}{B\sqrt{M}}\right).$$

Taking $L=L_{\gamma,H}$ proves Theorem 4. □

H.3 Proof of Proposition 6

Proof. Let

$$d_m:=v_{m+1}-v_m,\qquad \xi_m:=\widehat{g}_m-\nabla F(v_m).$$

Projection optimality for $v_{m+1}=\Pi_{\mathcal{V}}(v_m+\eta\widehat{g}_m)$, with comparison point v_m, gives

$$\langle\widehat{g}_m,d_m\rangle\ge\eta^{-1}\|d_m\|_2^2.$$

Using smoothness and $\widehat{g}_m=\nabla F(v_m)+\xi_m$,

$$\begin{aligned}F(v_{m+1})-F(v_m)&\ge\langle\nabla F(v_m),d_m\rangle-\frac{L_{\gamma,H}}{2}\|d_m\|_2^2\\&\ge\left(\eta^{-1}-\frac{L_{\gamma,H}}{2}\right)\|d_m\|_2^2-\langle\xi_m,d_m\rangle.\end{aligned}$$

Young's inequality gives

$$-\langle\xi_m,d_m\rangle\ge-\frac{1}{4\eta}\|d_m\|_2^2-\eta\|\xi_m\|_2^2.$$

Hence, when $\eta\le(2L_{\gamma,H})^{-1}$,

$$F(v_{m+1})-F(v_m)\ge\frac{\eta}{2}\left\|\frac{d_m}{\eta}\right\|_2^2-\eta\|\xi_m\|_2^2. \tag{20}$$

Taking expectations, summing over m, and using $F(v_{M+1})\le F^\star$ yields

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\frac{d_m}{\eta}\right\|_2^2\le\frac{2(F^\star-F(v_1))}{\eta M}+\frac{2\sigma_{\gamma,H}^2}{M}\sum_{m=1}^{M}\frac{1}{B_m}. \tag{21}$$

Let

$$y_m:=\Pi_{\mathcal{V}}(v_m+\eta\nabla F(v_m)).$$

Projection nonexpansiveness gives

$$\left\|\mathcal{G}_\eta(v_m)-\frac{d_m}{\eta}\right\|_2=\frac{1}{\eta}\|y_m-v_{m+1}\|_2\le\|\xi_m\|_2.$$

Therefore,

$$\|\mathcal{G}_\eta(v_m)\|_2^2 \le 2\left\|\frac{d_m}{\eta}\right\|_2^2 + 2\|\xi_m\|_2^2.$$

Combining with (21) gives

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\mathcal{G}_\eta(v_m)\|_2^2 \le \frac{4(F^\star - F(v_1))}{\eta M} + \frac{6\sigma_{\gamma,H}^2}{M}\sum_{m=1}^{M}\frac{1}{B_m}.$$

The fixed-batch and increasing-batch displays follow from $B_m \equiv B$ and $\sum_{m=1}^{M} 1/\lceil B_0 m\rceil \le (1+\log M)/B_0$, respectively. □

The direct constant-step result is an upper bound, not a lower bound. It does not exclude diminishing-step projected stochastic methods. The averaged–damped update is used because it retains a fixed reference projection scale while making the fixed-batch stochastic term vanish.

I Square-Root Temperature Schedule

We specialize the preceding bounds to $\tau_t = \tau_0 t^{-1/2}$ with fixed $\gamma \in (0,1)$. The follower guarantee follows from Theorem 2 and

$$\sum_{t=1}^{H} t^{-1/2} \le 2\sqrt{H}.$$

For the leader level, $\tau_t = \tau_0 t^{-1/2}$ gives

$$\frac{t+\beta_t^2}{\tau_t^2} = \tau_0^{-2} t(t+\beta_t^2),$$

and

$$\frac{\sqrt{t}+\beta_t}{\tau_t} = \tau_0^{-1}(t+\beta_t\sqrt{t}).$$

By $(a+b)^2 \le 2a^2+2b^2$, the anytime confidence schedule satisfies

$$\beta_t^2 \le 2\lambda\frac{S^2}{\alpha^2} + 2\sigma_Y^2\left[2\log\frac{t(t+1)}{\delta_0} + d\log\left(1+\frac{t-1}{d\lambda}\right)\right].$$

Using $d\log(1+x/d) \le x$ with $x = (t-1)/\lambda$ gives

$$\beta_t^2 \le C_\beta\left(1+t+\log\frac{t+1}{\delta_0}\right),$$

where C_β depends only on the fixed parameters S, α, σ_Y, and λ, and is independent of d, t, and H.

For fixed $\gamma < 1$, Lemma 1 gives

$$\overline{W}_t^{\gamma,H} \le \gamma^{t-1}.$$

Therefore the sums defining $\mathcal{A}_{\gamma,H}$ and $\mathcal{C}_{\gamma,H}$ are bounded uniformly in H by convergent geometric moments of polynomials and logarithms in t. More explicitly,

$$\sup_{H\ge1}\mathcal{A}_{\gamma,H} < \infty, \qquad \sup_{H\ge1}\mathcal{C}_{\gamma,H} < \infty,$$

where these two sums depend on $\gamma, \tau_0, S, \alpha, \sigma_Y$, and λ, but not on H. The multiplicative constants connecting them to smoothness and variance also depend on R_Y. The dimension dependence remains the explicit leading factor d in Theorem 4.

Combining these bounds with Theorem 4 shows that, for fixed $\gamma < 1$,

$$\mathbb{E}\|\mathcal{G}_\eta(\widehat{v})\|_2^2 \le \widetilde{O}\left(\frac{d\Delta_1}{\sqrt{M}} + \frac{d}{B\sqrt{M}}\right).$$

One may choose a single projection scale η from an infinite-geometric-moment upper bound on $L_{\gamma,H}^{\mathrm{raw}}$, making the same scale valid for every nominal horizon. This horizon-uniform analytical constant is a consequence of the fixed-discount objective; it does not remove the H interaction rounds needed to generate one deployment.

Each outer iteration uses B independent H-round deployments. Therefore, Algorithm 1 uses a total of BMH interaction rounds.

J Equivalent Visible-Coordinate Derivative Recursions

This appendix proves the exact visible-coordinate recursions stated in Appendix A. They are behaviorally equivalent to the raw-coordinate calculus and provide a direct implementation check. The main uniform sensitivity bounds use the more compact raw-coordinate representation in Appendix E. All derivatives below are pathwise: the realized contexts, actions, and rewards are held fixed, while the probability of observing the trajectory is differentiated through Appendix D.

J.1 Preliminary visible-feature derivatives

For $z_t(a;v) = v \odot \phi(x_t, a)$ with $v \in \mathcal{V} = [\alpha, 1]^d$,

$$\partial_{v_i} z_t(a;v) = \phi_i(x_t, a) e_i, \qquad \partial^2_{v_i v_j} z_t(a;v) = 0.$$

Because $\|\phi(x,a)\|_2 \le 1$, we have

$$\|\partial_{v_i} z_t(a)\|_2 \le 1, \qquad \|\partial^2_{v_i v_j} z_t(a)\|_2 = 0.$$

J.2 Proof of Lemma 3: state derivative recursions

The covariance update is $V_{t+1} = V_t + z_t z_t^\top$, where $z_t = z_t(a_t)$. Therefore

$$\Xi^{(i)}_{t+1} = \Xi^{(i)}_t + (\partial_{v_i} z_t) z_t^\top + z_t (\partial_{v_i} z_t)^\top.$$

Differentiating this equation with respect to v_j gives

$$\Omega^{(ij)}_{t+1} = \Omega^{(ij)}_t + (\partial^2_{v_i v_j} z_t) z_t^\top + (\partial_{v_i} z_t)(\partial_{v_j} z_t)^\top + (\partial_{v_j} z_t)(\partial_{v_i} z_t)^\top + z_t (\partial^2_{v_i v_j} z_t)^\top.$$

Similarly, from $b_{t+1} = b_t + z_t Y_t$ and the pathwise convention,

$$\zeta^{(i)}_{t+1} = \zeta^{(i)}_t + (\partial_{v_i} z_t) Y_t, \qquad \xi^{(ij)}_{t+1} = \xi^{(ij)}_t + (\partial^2_{v_i v_j} z_t) Y_t.$$

For $\widehat{\psi}_t = V_t^{-1} b_t$, the inverse derivative identity gives

$$\partial_{v_i} V_t^{-1} = -V_t^{-1} \Xi^{(i)}_t V_t^{-1},$$

and hence

$$\chi^{(i)}_t = -V_t^{-1} \Xi^{(i)}_t \widehat{\psi}_t + V_t^{-1} \zeta^{(i)}_t.$$

A second differentiation uses

$$\partial_{v_j} \partial_{v_i} V_t^{-1} = V_t^{-1} \Xi^{(j)}_t V_t^{-1} \Xi^{(i)}_t V_t^{-1} + V_t^{-1} \Xi^{(i)}_t V_t^{-1} \Xi^{(j)}_t V_t^{-1} - V_t^{-1} \Omega^{(ij)}_t V_t^{-1}.$$

Substituting this expression into $\partial_{v_j} \partial_{v_i} (V_t^{-1} b_t)$ yields

$$\begin{aligned} \Gamma^{(ij)}_t = {} & V_t^{-1} \Xi^{(j)}_t V_t^{-1} \Xi^{(i)}_t \widehat{\psi}_t + V_t^{-1} \Xi^{(i)}_t V_t^{-1} \Xi^{(j)}_t \widehat{\psi}_t - V_t^{-1} \Omega^{(ij)}_t \widehat{\psi}_t \\ & - V_t^{-1} \Xi^{(i)}_t V_t^{-1} \zeta^{(j)}_t - V_t^{-1} \Xi^{(j)}_t V_t^{-1} \zeta^{(i)}_t + V_t^{-1} \xi^{(ij)}_t. \end{aligned}$$

J.3 Proof of Lemma 4: radius and optimistic-score derivatives

Fix t and a and write $z = z_t(a)$ and $A = V_t^{-1}$. With $s = z^\top A z$,

$$\partial_{v_i} s = 2 (\partial_{v_i} z)^\top A z - z^\top A \Xi^{(i)}_t A z.$$

For the second derivative,

$$\begin{aligned} \partial^2_{v_i v_j} s = {} & 2 (\partial^2_{v_i v_j} z)^\top A z + 2 (\partial_{v_i} z)^\top A (\partial_{v_j} z) - 2 (\partial_{v_i} z)^\top A \Xi^{(j)}_t A z - 2 (\partial_{v_j} z)^\top A \Xi^{(i)}_t A z \\ & + z^\top A \Xi^{(j)}_t A \Xi^{(i)}_t A z + z^\top A \Xi^{(i)}_t A \Xi^{(j)}_t A z - z^\top A \Omega^{(ij)}_t A z. \end{aligned}$$

If $\phi(x_t, a) \neq 0$, then $z_t(a) \neq 0$ and positive definiteness of V_t^{-1} gives $r > 0$, so

$$\partial_{v_i} r = \frac{\partial_{v_i} s}{2r}, \qquad \partial^2_{v_i v_j} r = \frac{\partial^2_{v_i v_j} s}{2r} - \frac{\partial_{v_i} s \partial_{v_j} s}{4 r^3}.$$

If $\phi(x_t, a) = 0$, then $z_t(a)$ and all of its interface derivatives are zero. Hence s and r are constant-zero functions of v, and their first and second derivatives are zero. Finally,

$$\partial_{v_i} q_t(a) = (\chi_t^{(i)})^\top z_t(a) + \widehat{\psi}_t^\top \partial_{v_i} z_t(a) + \beta_t \partial_{v_i} r_t(a),$$

and

$$\partial^2_{v_i v_j} q_t(a) = (\Gamma_t^{(ij)})^\top z_t(a) + (\chi_t^{(i)})^\top \partial_{v_j} z_t(a) + (\chi_t^{(j)})^\top \partial_{v_i} z_t(a) + \widehat{\psi}_t^\top \partial^2_{v_i v_j} z_t(a) + \beta_t \partial^2_{v_i v_j} r_t(a).$$

J.4 Proof of Lemma 5: softmax derivatives

At round t, let

$$\pi(a) = \frac{\exp(q(a)/\tau_t)}{\sum_b \exp(q(b)/\tau_t)},$$

where $\tau_t > 0$ is deterministic with respect to v. Then

$$\log \pi(a) = \frac{q(a)}{\tau_t} - \log \sum_b \exp(q(b)/\tau_t).$$

Therefore

$$\nabla \log \pi(a) = \frac{1}{\tau_t} \left(\nabla q(a) - \sum_b \pi(b) \nabla q(b) \right).$$

For coordinates i, j,

$$\partial^2_{v_i v_j} \log \pi(a) = \frac{1}{\tau_t} \left(\partial^2_{v_i v_j} q(a) - \sum_b \pi(b) \partial^2_{v_i v_j} q(b) \right) - \frac{1}{\tau_t^2} \left(\sum_b \pi(b) \partial_{v_i} q(b) \partial_{v_j} q(b) - \bar{q}_i \bar{q}_j \right),$$

where $\bar{q}_i = \sum_b \pi(b) \partial_{v_i} q(b)$.

K Additional Sensitivity Diagnostics

The main results use worst-case Euclidean score and Hessian envelopes. The following one-step example shows that the inverse-temperature-squared second-moment dependence cannot generally be removed without additional structure.

Proposition 7 (One-step CILOT score variance)**.** *Assume $0 < \alpha < 1$, take $d = K = 2$, $\lambda = 1$, and let $c = (1 + \alpha)/2$. At the first round, use*

$$\phi(1) = e_1, \qquad \phi(2) = c e_2,$$

an optimism width $\beta_1 > 0$, and interface $v_0 = (c, 1) \in (\alpha, 1) \times \{1\}$. For the coordinate score with respect to v_1 under temperature $\tau > 0$,

$$\mathbb{E}\left[\left(\partial_{v_1} \log \pi_{v_0}(A) \right)^2 \right] = \frac{\beta_1^2}{4\tau^2}.$$

Proof. At the first round, $\widehat{\psi}_1 = 0$ and $V_1 = I$. Hence

$$q_1(v) = \beta_1 v_1, \qquad q_2(v) = \beta_1 c v_2.$$

At $v_0 = (c, 1)$, the logits are equal, so both actions have probability $1/2$. Moreover,

$$\partial_{v_1} q_1 = \beta_1, \qquad \partial_{v_1} q_2 = 0.$$

The two score values are therefore

$$\partial_{v_1} \log \pi_{v_0}(1) = \frac{\beta_1}{2\tau}, \qquad \partial_{v_1} \log \pi_{v_0}(2) = -\frac{\beta_1}{2\tau}.$$

Averaging their squares gives $\beta_1^2/(4\tau^2)$. □

Table 4: Experimental hyperparameters.

Parameter	Synthetic	Last.FM	MovieLens
Number of items N	5000	2000	2000
Number of follower types N_u	200	500	500
Feature dimension d	50	50	50
Candidate-set size	50	50	50
Outer iterations M	3000	3000	3000
Deployment horizon H	200	200	200
Batch size B	16	16	16
Discount factor γ	0.99	0.99	0.99
Reward noise σ_y	0.05	0	0
Interface range	$[0.1, 1]^d$	$[0.1, 1]^d$	$[0.1, 1]^d$
Interface step size η	20	20	20
Regret horizon H_{eval}	5000	5000	5000
Random seeds	10	10	10

This example does not assert that every instance attains the worst-case bound. It shows that a uniform Euclidean second-moment analysis cannot replace the τ^{-2} dependence by a temperature-independent constant. Margin conditions, Fisher-metric stationarity, or problem-dependent cancellations may produce sharper guarantees under additional assumptions.

Analytical role of the raw-coordinate identity. The visible-coordinate derivative recursion in Appendix J is exact. Used directly for uniform bounds, however, it treats the covariance, response vector, estimate, and radius quotient as separate evolving objects. The raw-coordinate identity collapses the same pathwise dependence into derivatives of the diagonal ridge $\Lambda(v)$. The relative bounds $|\partial_{v_i} r_t| \lesssim r_t$ and $|\partial^2_{v_i v_j} r_t| \lesssim r_t$ then handle the zero-radius case and yield the time-local envelopes used by the smoothness theorem. The two representations remain behaviorally identical.

L Additional Experimental Details

This appendix supplements the experimental setup in Section 6 with implementation and evaluation details used for interface optimization, the zero-baseline ablation, follower-regret evaluation, and analysis of the learned interface in the controlled Synthetic environment.

Online interaction protocol and hyperparameters. We follow the online interaction and leader-utility evaluation protocol described in Section 6. Within each deployment, the follower type is sampled once from the fixed population and remains unchanged for the entire trajectory. At every outer iteration, each method is evaluated using B fresh deployments. The additional hyperparameters are summarized in Table 4. The interface is constrained coordinate-wise to $[0.1, 1]^d$, and all experiments use interface step size $\eta = 20$ in the direct projected-ascent update in (8). The separate follower-regret evaluation uses horizon $H_{\text{eval}} = 5000$.

Leader-utility evaluation and reference policies. At outer iteration m, each method is evaluated on B fresh deployments. Let $U_{m,b,t}$ denote the leader utility at round t of deployment b. We report

$$\widehat{J}_m = \frac{1}{B} \sum_{b=1}^{B} \sum_{t=1}^{H} \omega_t^{\gamma,H} U_{m,b,t},$$

where

$$\omega_t^{\gamma,H} = \frac{(1-\gamma)\gamma^{t-1}}{1-\gamma^H}, \qquad \sum_{t=1}^{H} \omega_t^{\gamma,H} = 1.$$

Final leader performance is defined as the last-iteration value $\widehat{J}_M$, rather than an average over a terminal optimization window.

Identity keeps $v_m = \mathbf{1}, m = 1, \ldots, M$. Although its interface does not change, Identity is evaluated on fresh deployments at every outer iteration. Consequently, its expected utility is constant across m, while its empirical curve exhibits small finite-batch fluctuations.

We additionally evaluate a full-information Oracle under exactly the same deployment protocol. For deployment b at outer iteration m, let $u_{m,b}$ be the sampled follower type and let $\mathcal{A}_{m,b,t}$ be the candidate set at round t. The Oracle selects

$$a^{\text{oracle}}_{m,b,t} \in \arg\max_{a \in \mathcal{A}_{m,b,t}} \left({\theta^\star_{u_{m,b}}}^\top \phi_a + c_a \right).$$

The selected arm is evaluated using the same feedback and utility definition as the other methods. In particular,

$$Y^{\text{oracle}}_{m,b,t} = {\theta^\star_{u_{m,b}}}^\top \phi_{a^{\text{oracle}}_{m,b,t}} + \sigma_y \epsilon_{m,b,t},$$

followed by

$$U^{\text{oracle}}_{m,b,t} = \frac{Y^{\text{oracle}}_{m,b,t} + c_{a^{\text{oracle}}_{m,b,t}} + 1 + 2\sigma_y}{3 + 4\sigma_y}.$$

Hence

$$\widehat{J}^{\text{oracle}}_m = \frac{1}{B} \sum_{b=1}^{B} \sum_{t=1}^{H} \omega_t^{\gamma,H} U^{\text{oracle}}_{m,b,t}.$$

The displayed Oracle curve fluctuates slightly because of deployment sampling.

CILOT baseline. For deployment b at outer iteration m, let

$$G_{m,b,t} = \sum_{s=t}^{H} \omega_s^{\gamma,H} U_{m,b,s}$$

denote the discounted tail credit from round t. Let $z_{m,b,t} = \nabla_v \log \pi_t(a_{m,b,t} \mid x_{m,b,t}, h_{m,b,t-1})$ denote the score of the prescribed follower policy. CILOT uses the score-function estimator

$$\widehat{g}_m = \frac{1}{B} \sum_{b=1}^{B} \sum_{t=1}^{H} (G_{m,b,t} - b_{m,t})\, z_{m,b,t}.$$

We choose the previous outer iteration directly as baseline. $b_{m,t} = 0$ when $m = 1$. For $m \geq 2$, $b_{m,t} = \overline{G}_{m-1,t} = \frac{1}{B} \sum_{b=1}^{B} G_{m-1,b,t}$. Thus the baseline used for iteration m is completely determined before the trajectories of iteration m are sampled.

Zero-baseline ablation. We examine the role of the baseline in the score-function estimator by comparing the default CILOT estimator with a zero-baseline variant, CILOT-ZERO, which sets $b_{m,t} = 0$ for all m, t. All other components of the optimization procedure and all hyperparameters are kept unchanged. Table 5 reports the final leader utility $\widehat{J}_M$ after $M = 3000$ outer iterations.

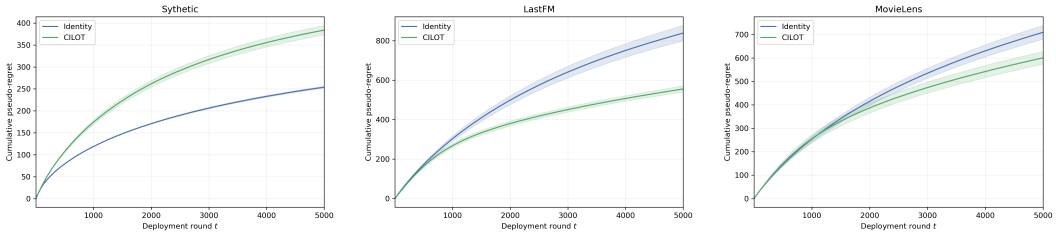
The baseline has a substantial effect on optimization in all three environments. Relative to CILOT-ZERO, the default estimator improves final leader utility by 0.0896 on Synthetic, 0.0932 on Last.FM, and 0.0611 on MovieLens. Moreover, CILOT-ZERO exhibits considerably larger variation across random seeds: its standard deviation is approximately $3.2\times$, $2.4\times$, and $4.2\times$ that of CILOT in the three environments, respectively. These results are consistent with the baseline stabilizing the score-function updates without changing their expectation.

Follower regret after interface optimization. After completing all $M = 3000$ outer iterations, we freeze the final CILOT interface v_M and evaluate the prescribed follower on fresh deployments of horizon $H_{\text{eval}} = 5000$. Identity uses the fixed interface $v = \mathbf{1}$. For a follower type u and candidate set $\mathcal{A}_t$, we measure regret with respect to the follower's own reward objective:

$$\text{Reg}(t) = \sum_{s=1}^{t} \left(\max_{a \in \mathcal{A}_s} {\theta^\star_u}^\top \phi_a - {\theta^\star_u}^\top \phi_{a_s} \right), \qquad t = 1, \ldots, H_{\text{eval}}.$$

Table 5: Zero-baseline ablation. We report the final leader utility $\widehat{J}_M$ at $M = 3000$ as mean $\pm$ standard deviation over 10 random seeds. CILOT uses $b_{m,t} = \overline{G}_{m-1,t}$, whereas CILOT-ZERO sets $b_{m,t} = 0$.

Dataset	CILOT-ZERO	CILOT
Synthetic	0.500077 ± 0.011724	0.589723 ± 0.003678
Last.FM	0.535139 ± 0.014239	0.628342 ± 0.005982
MovieLens	0.536245 ± 0.014654	0.597347 ± 0.003471



(a) Synthetic (b) Last.FM (c) MovieLens

Figure 2: Post-optimization cumulative follower regret. CILOT is first optimized for $M = 3000$ outer iterations, after which the learned interface v_M is frozen and evaluated on fresh deployments for $H_{\text{eval}} = 5000$ rounds. Identity uses $v = \mathbf{1}$. Curves show the mean over 10 random seeds and shaded regions denote one standard deviation.

For each seed, the regret curve is first averaged over the B fresh deployments and then aggregated across the 10 random seeds. We omit the full-information Oracle from this comparison because its action rule optimizes $\theta_u^{\star\top}\phi_a + c_a$, rather than the follower reward used in the regret definition above.

Figure 2 reports the resulting cumulative follower regret in the Synthetic, Last.FM, and MovieLens environments. Across all three environments, the regret curves grow sublinearly over the evaluation horizon and are empirically consistent with the $\widetilde{O}(\sqrt{H})$ dependence predicted by the follower-regret analysis.

At the end of the $H_{\text{eval}} = 5000$ evaluation horizon, the Synthetic environment exhibits higher follower regret under CILOT than under Identity: final cumulative regret is 253.9175 ± 2.1662 for Identity and 383.9217 ± 10.6298 for CILOT. Thus, although the learned interface substantially improves leader utility in the controlled Synthetic environment, it also changes the finite-horizon exploration behavior of the follower and increases cumulative regret relative to the Identity interface.

The ordering reverses on the two real-world datasets. On Last.FM, final regret decreases from 838.2097 ± 38.8744 under Identity to 555.3018 ± 15.5942 under CILOT, while on MovieLens it decreases from 709.6988 ± 29.1516 to 600.6168 ± 26.9783. Hence, in these environments the learned interface improves the leader objective while also reducing the follower's cumulative regret relative to the unmodified Identity interface.

Interpreting the learned interface. The Synthetic environment provides a controlled setting for examining whether the learned interface reflects the latent coordinate salience used to construct the platform objective. Recall that $s_a = \sum_{j=1}^{d} w_j \phi_{a,j}$, where the environment-level vector $w \in [0,1]^d$ directly controls the coordinate-wise contribution to the platform score. Since the arm features and follower parameters are generated symmetrically across coordinates, w provides an explicit coordinate-specific platform-side signal against which the learned interface can be compared.

To aggregate this relationship across random seeds, we rank the coordinates separately within each seed according to w_j. For seed $r \in \{1, \ldots, 10\}$, let $\pi_r(1), \ldots, \pi_r(d)$ denote the permutation satisfying $w^{(r)}_{\pi_r(1)} \leq \cdots \leq w^{(r)}_{\pi_r(d)}$. We then align the latent weights and final learned interface

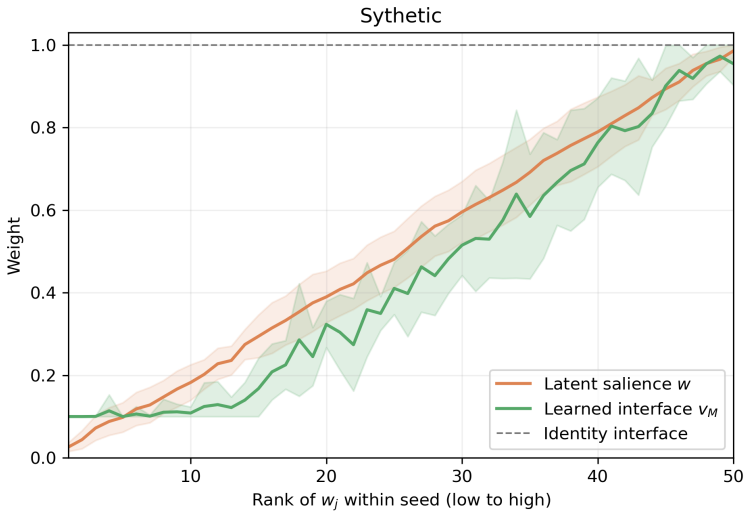


Figure 3: Relationship between latent coordinate salience w and the final learned CILOT interface in the Synthetic environment. Coordinates are ranked within each seed according to w_j and averaged at the same rank across 10 random seeds. Solid curves show across-seed means and shaded regions denote one standard deviation. The dashed horizontal line at $v = 1$ corresponds to the Identity interface.

according to this seed-specific ordering and compute

$$\overline{w}_{(k)} = \frac{1}{10}\sum_{r=1}^{10} w^{(r)}_{\pi_r(k)}, \qquad \overline{v}_{(k)} = \frac{1}{10}\sum_{r=1}^{10} v^{(r)}_{M,\pi_r(k)}, \qquad k = 1, \ldots, d.$$

Figure 3 shows that the learned interface increases steadily with latent coordinate salience. Coordinates with small w_j are more strongly attenuated, whereas coordinates with larger w_j receive progressively larger interface weights. The resulting profile closely follows the increasing latent-salience structure.

We quantify this relationship using the Pearson correlation ρ_r for seed r as $\rho_r = \mathrm{Corr}\left(w^{(r)}, v_M^{(r)}\right)$ between $w^{(r)}$ and the final learned interface $v_M^{(r)}$. Across the 10 random seeds, $\overline{\rho} = 0.9579 \pm 0.0057$. The strong positive correlation shows that, in this controlled environment, the optimized interface is approximately linearly aligned with the latent coordinate salience used to construct the platform objective.