MINIMAX OPTIMAL REGRET FOR CAUSAL LOGISTIC BANDITS WITH COUNTERFACTUAL FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study causal logistic bandits with counterfactual fairness constraints. The causal structure is given through known factual and counterfactual feature maps that share an unknown logistic reward parameter, but the learner observes only factual rewards. Consequently, the directions determining counterfactual feasibility need not be identifiable from the available feedback. The closest prior analyses either omit a coverage condition or impose a comparatively strong one, and do not establish matching lower bounds. We first show that some coverage condition is necessary: without a coverage-type restriction, factually indistinguishable environments with different optimal fair actions force $\Omega(T)$ expected joint loss. Under a weaker full-rank condition on the factual covariance pooled across actions, we identify a target-specific information scale $V_\star$ that measures the difficulty of estimating rewards and counterfactual effects from factual feedback. We construct worst-case families satisfying this condition on which every policy incurs expected joint loss $\Omega\big([V_\star \min\{\log K, d\}]^{1/3} T^{2/3}\big)$. We also give an explore–then–exploit procedure tuned using $V_\star$ and an adaptive algorithm that does not require its value. Both algorithms achieve $\max\{R_T, V_T\} = \widetilde{O}\big([V_\star \min\{\log K, d\}]^{1/3} T^{2/3} + \kappa d/\sigma_0^2\big)$, where R_T is regret relative to the best fair action and V_T denotes the cumulative stage-wise positive violations. Thus the upper and lower bounds match in their leading dependence on T, $V_\star$, and $\min\{\log K, d\}$, up to logarithmic factors.

1 INTRODUCTION

Contextual bandits model sequential decisions in which a learner observes a context, chooses an action, and receives feedback only for the chosen action. Causal bandits use known causal structure to share information across interventions (Lattimore et al., 2016; Sen et al., 2017; Lu et al., 2020). We study a *causal logistic* bandit in which the causal structure is given through known factual and counterfactual feature maps. These maps encode a specified intervention on a protected attribute, while the logistic reward parameter shared by the two maps is unknown. Our focus is the resulting constrained bandit learning problem: the learner seeks high rewards while keeping the factual–counterfactual mean reward difference within a prescribed tolerance.

The difficulty is that the learner observes only the factual reward generated by the chosen action. The counterfactual reward after intervention, and hence the statistic used to determine fairness, is never observed (Chen et al., 2025). The learner must use factual feedback both to identify rewarding actions and to determine which actions are fair. We ask whether factual observations provide enough information about the parameter directions that determine the counterfactual effect.

The closest predecessors of our work are Chen et al. (2025) and Chen et al. (2026). The former introduces causal logistic bandits with a counterfactual fairness constraint. Its assumptions, however, do not ensure that factual observations cover the parameter directions needed to estimate the counterfactual feedback. This turns out to be a fundamental problem: if counterfactual feedback depends on a parameter component that factual feedback cannot identify, then no policy can guarantee both sublinear regret and sublinear violation. The latter makes this issue explicit by imposing a covariate diversity assumption and adding forced exploration, obtaining sublinear guarantees. Its assumption, however, requires every action separately to cover every parameter direction, whereas we only re-

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.

Table 1: Upper and lower bounds for the joint loss, which is the sum of regret and violation.

Result	Knowledge of $V_\star$	Guarantee
Fixed ETE (Alg. 1)	Used to choose m	$\widetilde{O}\big([V_\star \min\{\log K, d\}]^{1/3} T^{2/3} + \kappa d/\sigma_0^2\big)$
Adaptive ETE (Alg. 2)	Not assumed	$\widetilde{O}\big([V_\star \min\{\log K, d\}]^{1/3} T^{2/3} + \kappa d/\sigma_0^2\big)$
Lower bound (Theorem 4.2)	—	$\Omega\big((V_\star \min\{\log K, d\})^{1/3} T^{2/3}\big)$
Proposition 3.2	—	$\Omega(T)$ without a coverage assumption

quire the covariance pooled across actions to be full rank. Moreover, Chen et al. (2026) does not provide a corresponding lower bound establishing whether its upper rate is unavoidable.

Under the pooled coverage condition, the learning difficulty depends on how accurately factual observations can estimate rewards and counterfactual effects. We measure this through a target scale $V_\star$, the largest local variance for estimating these quantities from uniform factual exploration.

Results. We write R_T for regret relative to the highest-reward fair oracle and V_T for the cumulative stage-wise positive violations, which cannot cancel across rounds. Their sum $R_T + V_T$ is the joint loss. These quantities are defined precisely in Section 2. Our contributions are threefold.

1. *Necessity of a coverage condition.* Without a coverage-type restriction, we construct two environments that generate identical observable history distributions under every policy but have different optimal fair actions. Every policy therefore incurs $\mathbb{E}[R_T + V_T] = \Omega(T)$ in at least one of them, ruling out uniform guarantees of both sublinear regret and sublinear violation over the unrestricted class. For the next two results, we adopt a sufficient condition: the factual covariance averaged over uniform action sampling is full rank (Assumption 3.1), while individual action covariances may be singular.
2. *Lower bounds on worst-case families.* Even under this coverage condition, we construct a two-environment, three-action family on which every policy incurs $\Omega(V_\star^{1/3} T^{2/3})$ expected joint loss in at least one environment. We then extend the construction to M environments and K actions, obtaining an additional $(\log M)^{1/3}$ factor, with $\log M = \Theta(\min\{\log K, d\})$.
3. *Matching upper bounds with adaptation to $V_\star$.* Our explore–then–exploit algorithm with a fixed exploration length $m \geq n_0$ guarantees with high probability that

$$\max\{R_T, V_T\} \leq m + 1024T\sqrt{\frac{V_\star \ell}{m}}$$

where n_0 is an initial estimation threshold and $\ell = O(\log(T/\delta) + \min\{\log K, d\})$ is a confidence factor. If $V_\star$ is known, balancing the two terms gives $O\big(n_0 + (V_\star \ell)^{1/3} T^{2/3}\big)$ joint loss. When $V_\star$ is unknown, an adaptive procedure uses observable confidence widths to choose when to collect more exploration data, retaining the same $V_\star$-dependent rate without taking $V_\star$ as an input. These upper bounds match the lower bound on worst-case families in their leading dependence on T, $V_\star$, and $\min\{\log K, d\}$, up to logarithms.

Table 1 summarizes our results, and notation $\widetilde{O}$ hides logarithmic factors in T, $1/\delta$, and fixed constants. The two upper bounds hold with probability at least $1 - \delta$ and control $\max\{R_T, V_T\}$. The lower bounds concern expected $R_T + V_T$ on worst-case families. A discussion of related work is provided in Appendix B.

2 Problem formulation

Fix a horizon $T \geq 2$, a finite action set $\mathcal{A} = [K]$, a context distribution P_X, and a parameter set $\Theta \subset \mathbb{R}^d$. Let $\mu(u) = 1/(1 + e^{-u})$ and note that $\dot{\mu}(u) = \mu(u)(1 - \mu(u))$. For every context–action pair (x, a) and $\theta \in \Theta$, define

$$r_\theta(x, a) = \mu\big(\phi(x, a)^\top \theta\big), \quad r_\theta^G(x, a) = \mu\big(\phi^G(x, a)^\top \theta\big), \quad \Delta_\theta(x, a) = r_\theta(x, a) - r_\theta^G(x, a).$$

The feature maps ϕ and ϕ^G, sending a context-action pair to a feature vector in $\mathbb{R}^d$, are known to the learner. Here $\phi^G(x,a)$ is an alternative notation for $\phi(x',a)$, where x' is a context x after counterfactual intervention. We call $\phi(x,a)^\top\theta$ the factual logit and $\phi^G(x,a)^\top\theta$ the counterfactual logit. We use the term *effect* for the factual–counterfactual statistic $\Delta_\theta(x,a)$.

In each round $t \in [T]$, a context $X_t \sim P_X$ is drawn independently. The learner observes $\phi(X_t,a)$ and $\phi^G(X_t,a)$ for every $a \in \mathcal{A}$ and chooses an action A_t. A reward is then drawn as $R_t \sim \text{Bernoulli}\big(r_{\theta^\star}(X_t,A_t)\big)$. The true parameter $\theta^\star$ is unknown, and neither $r^G_{\theta^\star}(X_t,A_t)$ nor $\Delta_{\theta^\star}(X_t,A_t)$ is observed. The use of the same parameter $\theta^\star$ in the two feature maps is the structural assumption that permits inferring counterfactual information from factual observations.

Assumption 2.1. Θ is convex and compact, $\theta^\star \in \Theta$, and $\|\theta\|_2 \le S$ for every $\theta \in \Theta$. Moreover, $\|\phi(x,a)\|_2 \le 1, \big\|\phi^G(x,a)\big\|_2 \le 1$ for every (x,a).

We define the global inverse curvature $\kappa = \sup_{\theta\in\Theta,x,a} \max\{\dot\mu(\phi(x,a)^\top\theta)^{-1}, \dot\mu(\phi^G(x,a)^\top\theta)^{-1}\}$, which is finite by Assumption 2.1. These boundedness and curvature conditions are standard in logistic bandits (Faury et al., 2020). We assume that κ is known to the learner; a known upper bound can replace it in the algorithms and their guarantees.

Fix a threshold $\tau \in [0,1]$ and define a true feasible set $\mathcal{A}_\tau(x) = \{a \in \mathcal{A} : |\Delta_{\theta^\star}(x,a)| \le \tau\}$.

Assumption 2.2 (Fair sets are nonempty)**.** $\mathcal{A}_\tau(x) \neq \varnothing$ for P_X-almost every x.

Using any fixed deterministic tie-breaking, let $a^\star(x) \in \arg\max_{a\in\mathcal{A}_\tau(x)} r_{\theta^\star}(x,a)$. The cumulative regret relative to this fair oracle and the cumulative stage-wise positive violations are

$$R_T = \sum_{t=1}^T \left[r_{\theta^\star}(X_t,a^\star(X_t)) - r_{\theta^\star}(X_t,A_t)\right], \quad V_T = \sum_{t=1}^T [|\Delta_{\theta^\star}(X_t,A_t)| - \tau]_+.$$

We call the sum $R_T + V_T$ the *joint loss*.

3 Coverage and the target scale $V_\star$

3.1 Coverage and learnability

Let $X \sim P_X$ and $A \sim \text{Unif}(\mathcal{A})$ independently. Define the *pooled factual covariance* Σ and Fisher information $I(\theta)$ as

$$\Sigma = \mathbb{E}\big[\phi(X,A)\phi(X,A)^\top\big], \quad I(\theta) = \mathbb{E}\left[\dot\mu\big(\phi(X,A)^\top\theta\big)\phi(X,A)\phi(X,A)^\top\right].$$

Assumption 3.1 (Coverage)**.** $\lambda_{\min}(\Sigma) = \sigma_0^2 > 0$, and σ_0 is known to the learner.

The pooled full-rank condition in Assumption 3.1 is also used in finite-armed contextual generalized linear bandits (Li et al., 2017; Jun et al., 2021). We will show that the same assumption on factual features suffices for sublinear regret and violation under counterfactual fairness constraints.

Without a restriction on parameter directions that factual observations cannot identify, no policy can guarantee both sublinear regret and sublinear violation uniformly over the model class.

Proposition 3.2 (Linear loss without counterfactual learnability)**.** *There exists a two-dimensional, two-action causal logistic bandit with θ_0,θ_1 satisfying Assumptions 2.1 and 2.2 such that, for every policy π,*

$$\max\{\mathbb{E}_{0,\pi}[R_T], \mathbb{E}_{1,\pi}[V_T]\} \ge \frac{T}{6}, \quad \max_{j\in\{0,1\}} \mathbb{E}_{j,\pi}[R_T + V_T] \ge \frac{T}{8}$$

where $\mathbb{E}_{j,\pi}$ denotes expectation under $\theta^\star = \theta_j$ and policy π for $j \in \{0,1\}$.

The construction for Proposition 3.2 uses a single context x_0, two actions, and two parameters θ_0,θ_1. The factual reward laws are identical under both parameters, but one action's counterfactual reward differs, changing the optimal fair action. The full proof is in Appendix C.

The issue in the construction is that factual features are orthogonal to a parameter direction that changes the counterfactual effect. Assumption 3.1 rules out this possibility by requiring Σ to be positive definite, so $\ker(\Sigma) = \{0\}$. Individual action covariances may still be singular. Thus, it is weaker than the covariate-diversity condition in Assumption 4 of Chen et al. (2026).

3.2 THE TARGET SCALE $V_\star$

We now use Assumption 3.1 to quantify the factual information available for the reward and effect directions. Write $I_\star = I(\theta^\star)$. By the definition of κ, we have $I_\star \succeq \frac{1}{\kappa}\Sigma \succeq \frac{\sigma_0^2}{\kappa} I_d$ and thus $I_\star^{-1} \preceq \frac{\kappa}{\sigma_0^2} I_d$. With the true parameter $\theta^\star$, define $q_r(x,a) = \dot{\mu}\big(\phi(x,a)^\top \theta^\star\big)\phi(x,a)$, $q_G(x,a) = \dot{\mu}\big(\phi^G(x,a)^\top \theta^\star\big)\phi^G(x,a)$, and $q_\Delta(x,a) = q_r(x,a) - q_G(x,a)$. These are the gradients of factual reward, counterfactual reward, and the effect at $\theta^\star$, respectively.

Definition 3.3 (Target scale)**.** Given $I_\star = I(\theta^\star)$, the target scale is defined as

$$V_\star = \operatorname*{ess\,sup}_{x \sim P_X} \max_{a \in \mathcal{A}} \max \left\{ \|q_r(x,a)\|^2_{I_\star^{-1}}, \|q_\Delta(x,a)\|^2_{I_\star^{-1}} \right\}.$$

The matrix $I_\star$ measures the information about each parameter direction provided by uniform factual exploration. The vectors q_r and q_Δ record which directions are relevant for reward and fairness. Hence $q^\top I_\star^{-1} q$ gives the local variance for estimating the target in direction q from factual data. The essential supremum bounds this variance for P_X-almost every context.

Definition 3.3 also gives $\|q_r\|_{I_\star^{-1}} \le \sqrt{V_\star}$ and $\|q_\Delta\|_{I_\star^{-1}} \le \sqrt{V_\star}$. Moreover, since $q_G = q_r - q_\Delta$, it follows that $\|q_G\|_{I_\star^{-1}} \le 2\sqrt{V_\star}$ and $\|q_r\|_{I_\star^{-1}} + \|q_G\|_{I_\star^{-1}} \le 3\sqrt{V_\star}$ hold almost surely. Furthermore, $\dot{\mu} \le 1/4$ and the feature norms are at most one, so $V_\star \le \kappa/(4\sigma_0^2)$. In the lower bound, the inequality $(q^\top h)^2 \le \big(q^\top I^{-1} q\big)\big(h^\top I h\big)$ relates the change in a target along a parameter direction h to the information available in that direction.

4 LOWER BOUNDS

We derive our lower bounds using information-theoretic arguments for stochastic bandits (Lattimore & Szepesvári, 2019). To the best of our knowledge, these are the first minimax lower bounds for causal logistic bandits with counterfactual fairness constraints. We construct environments with different optimal fair actions that can be distinguished only by choosing low-reward actions. The factual and counterfactual features are chosen so that the information needed to distinguish these environments is determined by $V_\star$, while the pooled factual covariance Σ remains full rank.

We compare environments with different true parameters: in environment j, the true parameter is $\theta^\star = \theta_j$. Write $V_{\star,j}$ for the target scale in Definition 3.3 evaluated at θ_j. The resulting lower bounds on the worst-case expected joint loss match the upper bounds in Section 5 in their leading dependence on T and $V_\star$, up to logarithmic factors.

4.1 TWO-ENVIRONMENT THREE-ACTION CONSTRUCTION

The first construction has dimension $d = 2$, three actions, and two possible parameters θ_0 and θ_1. Contexts are drawn independently and uniformly from $\{-1, 1\}$. The three actions, H (high-reward), S (safe), and E (informative), have the following roles.

Table 2: Roles of the three actions in the two-dimensional family.

Action	Reward role	Information role	Fairness role
H	high reward	uninformative	fair only in environment 0
S	slightly lower reward	uninformative	always fair
E	low reward	informative	always fair

Set $v_0 = \dot{\mu}(3/4)$ and $\Delta_E = 1/2 - \mu(-3/4) - 1/128 > 0$. Fix $B \ge 1$ and $\rho \in (0, 1]$, write $v_B = \dot{\mu}(B)$, and define $C = (1+\rho^2)/(4e^{1/4}\rho^2 v_B)$. The full construction is given in the appendix.

For a parameter α to be chosen later, the family has the following properties: (1) in environment 0, H is the optimal fair action, and every other action incurs a joint cost of at least a constant multiple of α; (2) in environment 1, S is the optimal fair action, and choosing H incurs a joint cost of at least a constant multiple of α; (3) each selection of E incurs a joint cost of at least Δ_E in both

environments; (4) H and S have the same factual reward laws in the two environments; (5) the KL divergence between the two environments for a single observation from E is at most $\alpha^2/(8C)$; (6) both environments have target scale $V_{\star,j} \asymp C$.

Theorem 4.1 (Binary lower bound)**.** *There exist universal constants $c_1, C_0 > 0$ such that, for every $B \geq 1$, every $\rho \in (0, 1]$, and every horizon $T \geq C_0 C$, there exists a two-environment family with $d = 2$ and $K = 3$, satisfying Assumptions 2.1–3.1 and $V_{\star,j} \asymp C$, for which*

$$\inf_{\pi} \max_{j \in \{0,1\}} \mathbb{E}_{j,\pi}[R_T + V_T] \geq c_1 C^{1/3} T^{2/3}. \tag{1}$$

Only observations from the costly action E distinguish the two environments. Any policy must therefore incur worst-case expected joint loss $\Omega(C/\alpha^2)$ from exploration or $\Omega(\alpha T)$ from uncertainty about whether H is fair. Balancing these costs with $\alpha \asymp (C/T)^{1/3}$ yields the lower bound.

4.2 The general lower bound

We extend Section 4.1 to M environments by replacing the single high-reward action with candidates $H_1, \ldots, H_M$. All candidates have factual reward $1/2$, but separated parameter directions and suitable counterfactual features make H_j the only fair candidate in environment j. Every other action incurs joint cost $\Omega(\alpha)$. A safe action remains available, and only costly informative actions distinguish the environments. The action budget and a packing in $d - 1$ dimensions allow $\log M \asymp \min\{\log K, d\}$. The exact parameter choices and construction are given in Appendix E.1.

Let D denote the information scale defined in the theorem below, with $V_{\star,j} \asymp D$. Each informative selection incurs constant joint cost, while the KL divergence between any two environments for one such observation is $O(\alpha^2/D)$. Reliably identifying H_j requires information of order $\log M$. Fano's inequality (Cover et al., 1991) then forces worst-case expected joint loss $\Omega(D \log M/\alpha^2)$ from exploration or $\Omega(\alpha T)$ from failing to select H_j. Balancing these costs yields the rate $(D \log M)^{1/3} T^{2/3}$ below.

Theorem 4.2 (General lower bound)**.** *There exist constants $c_2, C_2 > 0$ with the following property. Fix $B \geq 1$, $\rho \in (0, 1]$, $d \geq 46$, and $K \geq 8$, and set $D = (d-1)(1+\rho^2)/(\rho^2 \dot{\mu}(B))$. For every horizon $T \geq C_2 D \min\{\log K, d\}$, there exists a finite family of d-dimensional, K-action environments indexed by $j \in \mathcal{J}$, each satisfying Assumptions 2.1–3.1 and $V_{\star,j} \asymp D$, such that*

$$\inf_{\pi} \max_{j \in \mathcal{J}} \mathbb{E}_{j,\pi}[R_T + V_T] \geq c_2 \big(D \min\{\log K, d\}\big)^{1/3} T^{2/3}. \tag{2}$$

Remark 4.3 (Connection to partial monitoring). The $T^{2/3}$ rate has a similar origin to that in hard partial-monitoring problems: distinguishing competing actions requires observations from actions with a constant loss gap (Kirschner et al., 2020, Appendix G). Here, the relevant hidden information is counterfactual feasibility, and learning it requires costly factual observations. Balancing exploration costs against the loss from choosing the wrong action yields the same horizon dependence.

5 Algorithms and upper bounds

We give two algorithms that use uniform exploration to estimate rewards and counterfactual effects, then select actions using confidence intervals. More exploration incurs regret and violation but narrows these intervals. The algorithms balance this cost against the regret and violation during exploitation, which we bound by the interval widths.

5.1 Confidence intervals and optimistic screening

Both algorithms use the following estimates and action-selection rule. We describe them for n exploration observations; Sections 5.2 and 5.3 specify when to collect these observations and update the estimates.

Let $\mathcal{D} = ((X_s^E, A_s^E, R_s^E))_{s=1}^n$ denote the observations retained for estimation, listed in order of collection, with $\phi_s^E = \phi(X_s^E, A_s^E)$. The algorithms collect these observations by scheduling exploration before seeing its contexts and choosing actions uniformly. Using all n observations, define

the unconstrained logistic MLE by

$$\widehat{\theta}_n \in \arg\max_{\theta\in\mathbb{R}^d} \sum_{s=1}^{n} \left[R_s^E (\phi_s^E)^\top \theta - \log\bigl(1 + e^{(\phi_s^E)^\top \theta}\bigr) \right]. \tag{3}$$

The same estimate gives the factual prediction $\mu(\phi(x,a)^\top \widehat{\theta}_n)$ and the counterfactual prediction $\mu(\phi^G(x,a)^\top \widehat{\theta}_n)$. To quantify their estimation errors, we use the estimated Fisher matrix

$$\widehat{H}_n = \sum_{s=1}^{n} \dot{\mu}\bigl((\phi_s^E)^\top \widehat{\theta}_n\bigr) \phi_s^E (\phi_s^E)^\top. \tag{4}$$

Let $\ell > 0$ be a confidence factor, whose value we specify below. When the MLE is finite and $\widehat{H}_n$ is invertible, define the logit radius $b_n(z) = 11\sqrt{\ell}\,\|z\|_{\widehat{H}_n^{-1}}$. The radius $b_n(z)$ is larger when z points in a direction with little information in $\widehat{H}_n$. Applying the increasing function μ to the two endpoints of the logit interval gives the mean-reward interval $[\mu(z^\top \widehat{\theta}_n - b_n(z)), \mu(z^\top \widehat{\theta}_n + b_n(z))]$. For a pair (x,a), denote this interval by $[\underline{r}_n(x,a), \overline{r}_n(x,a)]$ when $z = \phi(x,a)$ and by $[\underline{r}_n^G(x,a), \overline{r}_n^G(x,a)]$ when $z = \phi^G(x,a)$. Since the effect is the factual mean minus the counterfactual mean, its interval is $[\underline{\Delta}_n(x,a), \overline{\Delta}_n(x,a)] = [\underline{r}_n(x,a) - \overline{r}_n^G(x,a), \overline{r}_n(x,a) - \underline{r}_n^G(x,a)]$. The largest reward or effect interval width at context x

$$W_n(x) = \max_{a\in\mathcal{A}} \max\left\{ \overline{r}_n(x,a) - \underline{r}_n(x,a), \overline{\Delta}_n(x,a) - \underline{\Delta}_n(x,a) \right\}$$

can be computed from the data without knowing $V_\star$.

Given these intervals, we retain every action whose effect interval intersects $[-\tau, \tau]$ and choose the retained action with the largest reward upper endpoint. The intersection means that the data have not ruled out the action being fair. We call an interval *valid* when it contains the corresponding true reward or effect. If all intervals are valid, the optimal fair action is retained. Optimistic reward selection then bounds regret by the selected action's reward interval width. A retained action can still violate the constraint, but its true effect and a value in $[-\tau, \tau]$ lie in the same interval, so its violation is at most the effect interval width. In the following lemma, we suppress the sample-size subscript n.

Lemma 5.1. *Fix a context x with $\mathcal{A}_\tau(x) \neq \varnothing$ and suppose that every reward and effect interval is valid. Retain the actions whose effect intervals intersect $[-\tau, \tau]$, and let A maximize the reward upper endpoint among the retained actions. Then $a^\star(x)$ is retained, and we have $r_{\theta^\star}(x, a^\star(x)) - r_{\theta^\star}(x, A) \le \overline{r}(x, A) - \underline{r}(x, A)$ and $[|\Delta_{\theta^\star}(x, A)| - \tau]_+ \le \overline{\Delta}(x, A) - \underline{\Delta}(x, A)$.*

Thus, valid intervals bound each of the one-round regret and violation by $W_n(x)$. We now choose the parameters that make these intervals valid with high probability. The minimum sample size n_0 ensures that the fitted MLE is finite and its Fisher matrix is invertible with high probability. The confidence factor ℓ accounts for all model fits and interval queries during the run. For $\delta \in (0,1)$, set $J = \lceil \log_2 T \rceil$, $M_{\mathrm{dir}} = \min\{KT, 5^d\}$, and $\ell = \log(32(J+1)(T+2)M_{\mathrm{dir}}/\delta)$. Here $J+1$ bounds the number of model fits in either algorithm. Set $n_0 = \lceil \kappa \Gamma_{\mathrm{corr}}/\sigma_0^2 \rceil$, where $\Gamma_{\mathrm{corr}} = \widetilde{O}(d)$ is the explicit threshold given in the appendix. In particular, $n_0 = \widetilde{O}(\kappa d/\sigma_0^2)$.

Lemma 5.2. *Suppose Assumptions 2.1 and 3.1 hold and $n_0 < T$. Run either algorithm in Sections 5.2–5.3. With probability at least $1 - \delta$, every MLE computed by the algorithm is finite and its estimated Fisher matrix is invertible. Whenever the algorithm uses an estimate based on n exploration observations to compute intervals at X_t, all factual reward and effect intervals are valid and*

$$W_n(X_t) \le 1024\sqrt{\frac{V_\star \ell}{n}}. \tag{5}$$

Both conclusions hold on the same event, simultaneously for all such fits and rounds.

The proof establishes finiteness of the MLE and combines Fisher concentration (Tropp, 2015) with a fixed-design logistic confidence bound (Jun et al., 2021). Applying μ and subtracting endpoints gives the reward and effect intervals. The width bound decreases as $n^{-1/2}$, with dependence on the reward and effect directions measured by $V_\star$. Lemma 5.1 then bounds per-round regret and violation during exploitation.

5.2 FIXED-LENGTH EXPLORE–THEN–EXPLOIT

Fix $m \in \{n_0, \dots, T\}$ in advance; this does not require $V_\star$. Algorithm 1 explores uniformly for m rounds, fits $(\widehat{\theta}_m, \widehat{H}_m)$, and uses this model with the optimistic screening rule from Section 5.1 in all remaining rounds.

The exploration rounds contribute at most m to each of regret and violation. Lemma 5.2 bounds the exploitation cost by $1024T\sqrt{V_\star \ell / m}$. Increasing m raises the first cost and lowers the second; balancing them gives the $T^{2/3}$ rate. This tuning uses $V_\star$ and respects $n_0 \le m \le T$.

Algorithm 1 Fixed-length ETE

Require: $T, m, \delta, \tau, \phi, \phi^G, \kappa, \sigma_0$, where $m \in \{n_0, \dots, T\}$
1: **for** $t = 1, \dots, m$ **do**
2: Observe X_t, choose $A_t \sim \mathrm{Unif}(\mathcal{A})$ independently, and observe R_t
3: **end for**
4: Compute $(\widehat{\theta}_m, \widehat{H}_m)$ using Eqs. (3)–(4)
5: **for** $t = m+1, \dots, T$ **do**
6: Observe X_t and compute the reward and effect intervals
7: $\widehat{\mathcal{A}}_t \leftarrow \{a \in \mathcal{A} : [\underline{\Delta}_m(X_t, a), \overline{\Delta}_m(X_t, a)] \cap [-\tau, \tau] \neq \varnothing\}$
8: $A_t \in \arg\max_{a \in \widehat{\mathcal{A}}_t} \overline{r}_m(X_t, a)$; observe R_t
9: **end for**

Theorem 5.3 (Fixed-length ETE)**.** *Suppose Assumptions 2.1–3.1 hold and $n_0 < T$. For every $m \in \{n_0, \dots, T\}$, fixed-length ETE satisfies, with probability at least $1 - \delta$,*

$$\max\{R_T, V_T\} \le m + 1024T\sqrt{\frac{V_\star \ell}{m}}. \tag{6}$$

Choosing $m = \min\{T, n_0 + \lceil (1024^2 V_\star \ell T^2)^{1/3} \rceil\}$ gives $\max\{R_T, V_T\} = O(n_0 + (V_\star \ell)^{1/3} T^{2/3})$.

Proof Sketch. The first m rounds contribute at most m to each of R_T and V_T. On the event of Lemma 5.2, Lemma 5.1 bounds the regret and violation in each subsequent round by the largest reward or effect interval width at X_t. Summing (5) over the remaining rounds and using $T - m \le T$ gives (6). Balancing the two terms yields the stated choice of m and the $T^{2/3}$ rate. □

5.3 ADAPTIVE EXPLORE–THEN–EXPLOIT

The choice of m in Theorem 5.3 requires $V_\star$, which is generally unknown. Without knowing $V_\star$, we can choose $m = \min\{T, n_0 + \lceil T^{2/3} \rceil\}$ to obtain $\max\{R_T, V_T\} = O\big(n_0 + (1 + \sqrt{V_\star \ell}) T^{2/3}\big)$ with probability at least $1 - \delta$. Algorithm 2 attains the $V_\star^{1/3}$ dependence without knowing $V_\star$ by checking the current interval widths. Narrow intervals allow exploitation; wide intervals trigger another exploration batch and a new fit, so exploration can resume after exploitation has begun.

Suppose $n_0 < T$. We call each model update a *checkpoint* and prescribe the cumulative exploration sample sizes $n_j = \min\{2^j n_0, T\}$, ending the sequence when it first reaches T. At checkpoint j, compute $(\widehat{\theta}_{n_j}, \widehat{H}_{n_j})$ from $\mathcal{D}$, with $|\mathcal{D}| = n_j$, using Eqs. (3)–(4). The algorithm keeps this estimate until another exploration batch is completed. For quantities computed from these observations, we abbreviate the subscript n_j by j; in particular, $\widehat{\theta}_j = \widehat{\theta}_{n_j}$, $\widehat{H}_j = \widehat{H}_{n_j}$, and $W_j = W_{n_j}$. The sample size n_j counts estimation observations, not elapsed rounds. If $n_0 \ge T$, the algorithm instead explores uniformly for all T rounds.

At checkpoint j, the algorithm exploits whenever $W_j(X_t) \le n_j / T$. The threshold compares the bound on regret and violation from the current intervals to the exploration cost n_j. When the intervals are valid, every round that passes this test contributes at most n_j/T to each of regret and violation. Even T such rounds at this checkpoint would thus contribute at most n_j to each. If the test fails, we collect more exploration data up to the next checkpoint and refit, or explore until the horizon ends.

The triggering round is excluded from fitting because the decision to explore depends on its context. Keeping it would favor contexts with wide intervals in the estimation data, even though the action

Algorithm 2 Adaptive ETE

Require: $T, \delta, \tau, \phi, \phi^G, \kappa, \sigma_0$
Compute $\ell, \Gamma_{\mathrm{corr}}, n_0$; define $n_i = \min\{2^i n_0, T\}$ for $i \geq 0$ and set $j \leftarrow 0$
for $t = 1, \ldots, n_0$ **do**
  Observe X_t, choose $A_t \sim \mathrm{Unif}(\mathcal{A})$ independently, and observe R_t
end for
Set $\mathcal{D} \leftarrow ((X_s, A_s, R_s))_{s=1}^{n_0}$; compute $(\widehat{\theta}_j, \widehat{H}_j)$ from $\mathcal{D}$ using Eqs. (3)–(4) with $n = n_j$
Set $t \leftarrow n_0 + 1$
while $t \leq T$ **do**
  Observe X_t and compute $W_j(X_t)$
  if $W_j(X_t) \leq n_j/T$ **then**
    Compute the reward and effect intervals at X_t as in Section 5.1
    $\widehat{\mathcal{A}}_t \leftarrow \{a \in \mathcal{A} : [\underline{\Delta}_j(X_t, a), \overline{\Delta}_j(X_t, a)] \cap [-\tau, \tau] \neq \varnothing\}$
    $A_t \in \arg\max_{a \in \widehat{\mathcal{A}}_t} \overline{r}_j(X_t, a)$; observe R_t and set $t \leftarrow t + 1$
  else
    Choose $A_t \sim \mathrm{Unif}(\mathcal{A})$ and observe R_t, without updating $\mathcal{D}$
    Set $b \leftarrow \min\{n_{j+1} - n_j, T - t\}$
    Before observing their contexts, schedule rounds $t + 1, \ldots, t + b$ for the following loop
    for $s = t + 1, \ldots, t + b$ **do**
      Observe X_s, choose $A_s \sim \mathrm{Unif}(\mathcal{A})$ independently
      Observe R_s and append (X_s, A_s, R_s) to $\mathcal{D}$
    end for
    if $b = n_{j+1} - n_j$ **then**
      Set $j \leftarrow j + 1$; compute $(\widehat{\theta}_j, \widehat{H}_j)$ from $\mathcal{D}$ using Eqs. (3)–(4) with $n = n_j$
    end if
    $t \leftarrow t + b + 1$
  end if
end while

is drawn uniformly. We instead schedule the next exploration batch before seeing its contexts. The fitting data can then be viewed as prefixes of a single iid uniform-exploration sequence, which allows us to use the same confidence analysis at every checkpoint.

All ties are resolved by a fixed deterministic rule. To bound the exploration cost, define $n_{\mathrm{crit}} = \max\{n_0, (1024^2 V_\star \ell T^2)^{1/3}\}$. This quantity is used only in the analysis; the algorithm checks the observable width instead. On the confidence event, once $n_j \geq n_{\mathrm{crit}}$, Lemma 5.2 implies $W_j(X_t) \leq 1024\sqrt{V_\star \ell / n_j} \leq \frac{n_j}{T}$, so no subsequent context can trigger further exploration. Doubling the checkpoint sizes limits any overshoot of this sufficient sample size to a factor of two, giving the same order of exploration cost as the fixed-length choice.

Theorem 5.4 (Adaptive ETE)**.** *Under Assumptions 2.1–3.1, with probability at least* $1 - \delta$*, adaptive ETE satisfies* $\max\{R_T, V_T\} \leq \min\{T, 4n_{\mathrm{crit}} + J + 1\}$*. Consequently,*

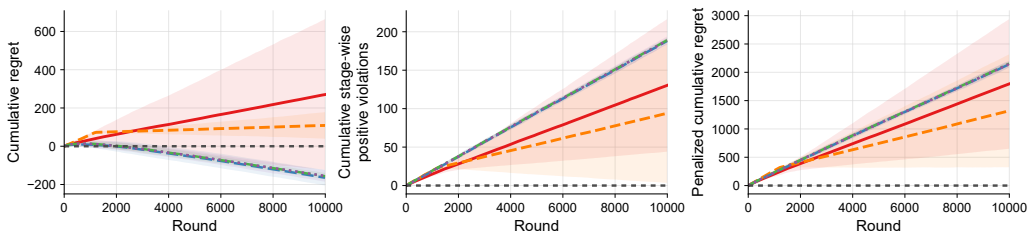
$$\max\{R_T, V_T\} = \widetilde{O}\left(\left[V_\star \min\{\log K, d\}\right]^{1/3} T^{2/3} + \frac{\kappa d}{\sigma_0^2} \right)$$

where $\widetilde{O}$ *suppresses logarithmic factors in* T *and* $1/\delta$*, as well as constants.*

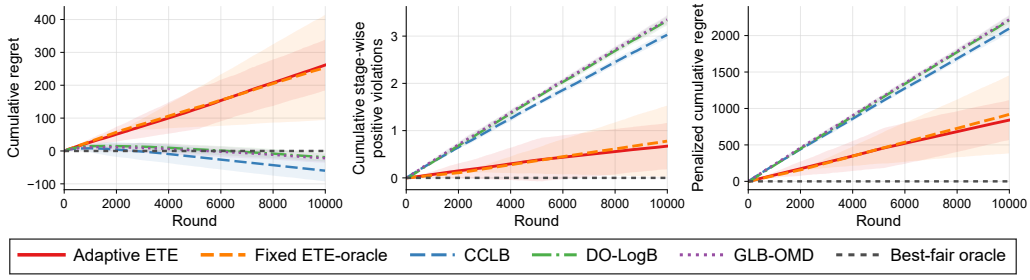
Proof Sketch. Every triggered batch targets $n_{j+1} \leq 2n_{\mathrm{crit}}$, since $n_j < n_{\mathrm{crit}}$ at a failed test. Thus at most $2n_{\mathrm{crit}}$ observations are retained, even if the last batch is truncated, and at most $J + 1$ trigger rounds are discarded. Each exploitation round contributes at most $n_j/T \leq 2n_{\mathrm{crit}}/T$ to each of regret and violation. Summing these costs proves the bound. □

6 Experiments

We compare Fixed ETE-oracle (Algorithm 1) and Adaptive ETE (Algorithm 2) with CCLB (Chen et al., 2025), DO-LogB (Chen et al., 2026), GLB-OMD (Zhang et al., 2026), and the best-fair oracle.



(a) Synthetic experiment.



(b) Structural causal model.

Figure 1: Cumulative regret, cumulative stage-wise positive violations, and penalized cumulative regret for the proposed methods and baselines.

Fixed ETE-oracle uses $V_\star$ to choose its exploration length; we estimate $V_\star$ empirically using $\theta^\star$ and the full set of offered contexts. The best-fair oracle knows $\theta^\star$ and selects the optimal fair action. We report cumulative regret, cumulative stage-wise positive violations, and penalized regret as defined by Chen et al. (2025), with zero reward assigned to unfair actions. Regret can be negative because an unfair action may yield a higher reward than the best fair action. Each experiment is repeated 10 times. Curves show means; shading spans two standard deviations on either side.

Synthetic experiment. We construct a $d = 10$ instance in which one coordinate is weakly represented in the factual features but remains visible in the counterfactual features. The coefficient of this coordinate in $\theta^\star$ is zero. Varying its factual scale therefore changes the information available for certifying counterfactual fairness without changing the mean rewards, fair action sets, or best fair actions. For the experiment in Figure 1(a), we set $K = 5$, $T = 10{,}000$, fairness threshold $\tau = 0.05$, and factual scale $\rho = 0.25$. The full experimental setup is given in Appendix A.2.

Structural causal model. We adapt the experimental setting of Chen et al. (2025) to bounded logistic features. The structural causal model contains a binary sensitive attribute A, a covariate W, a mediator M whose distribution depends on (A, W), and an action-specific variable D_a. We obtain the counterfactual feature vector by replacing A with $1 - A$ and recomputing M with the same exogenous noise. We clip and rescale features derived from (A, W, M, D_a, MD_a) and draw Bernoulli rewards with a logistic mean. For the experiment in Figure 1(b), we set $K = 20$, $T = 10{,}000$, and $\tau = 0.001$. Following Chen et al. (2025), we retain contexts for which at least 12 actions are fair. The full experimental setup is given in Appendix A.3.

Figure 1 shows that both ETE methods incur lower cumulative stage-wise positive violations and lower penalized regret than the learning baselines. Some baselines attain negative regret by choosing high-reward unfair actions, but incur higher cumulative stage-wise positive violations and higher penalized regret. Appendix A.2 presents additional experiments that vary the factual scale ρ, which controls the target-information scale $V_\star$. Appendix A.3 examines the effect of varying the fairness threshold τ.

REFERENCES

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Generalized linear bandits with safety constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3562–3566. IEEE, 2020.

Seoungbin Bae and Dabeen Lee. Neural logistic bandits. *arXiv preprint arXiv:2505.02069*, 2025.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Jiajun Chen, Jin Tian, and Christopher John Quinn. Causal logistic bandits with counterfactual fairness constraints. In *Forty-second International Conference on Machine Learning*, 2025.

Jiajun Chen, Jin Tian, and Christopher John Quinn. A causal decomposition approach for fair contextual multi-armed bandits. In *Forty-third International Conference on Machine Learning*, 2026.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 7801–7808, 2019.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

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

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

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

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

Wen Huang, Lu Zhang, and Xintao Wu. Achieving counterfactual fairness for causal bandit. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 6952–6959, 2022.

Rodolphe Jenatton, Jim Huang, and Cédric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pp. 402–411. PMLR, 2016.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pp. 5148–5157. PMLR, 2021.

Kia Khezeli and Eilyan Bitar. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10202–10209, 2020.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pp. 2328–2369. PMLR, 2020.

Yuko Kuroki, Alberto Rumi, Taira Tsuchiya, Fabio Vitale, and Nicolò Cesa-Bianchi. Best-of-both-worlds algorithms for linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 1216–1224. PMLR, 2024.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *Cambridge University*, 355, 2019.

Jacob Lear and Lu Zhang. A causal lens for learning long-term fair policies. In *International Conference on Learning Representations*, volume 2025, pp. 3344–3359, 2025.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*, 2024.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.

Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pp. 141–150. PMLR, 2020.

Blake Mason, Kwang-Sung Jun, and Lalit Jain. An experimental design approach for regret minimization in logistic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7736–7743, 2022.

Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. *Advances in neural information processing systems*, 33:11191–11201, 2020.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pp. 2827–2835. PMLR, 2021.

Drago Plecko and Elias Bareinboim. Causal fairness for outcome control. *Advances in Neural Information Processing Systems*, 36:47575–47597, 2023.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pp. 3057–3066. PMLR, 2017.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

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.

Weiqiang Wu, Jing Yang, and Cong Shen. Stochastic linear contextual bandits with diverse contexts. In *International Conference on Artificial Intelligence and Statistics*, pp. 2392–2401. PMLR, 2020.

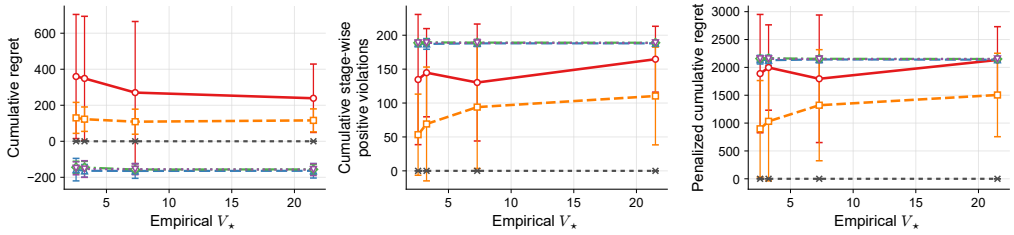
Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*, 2019.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

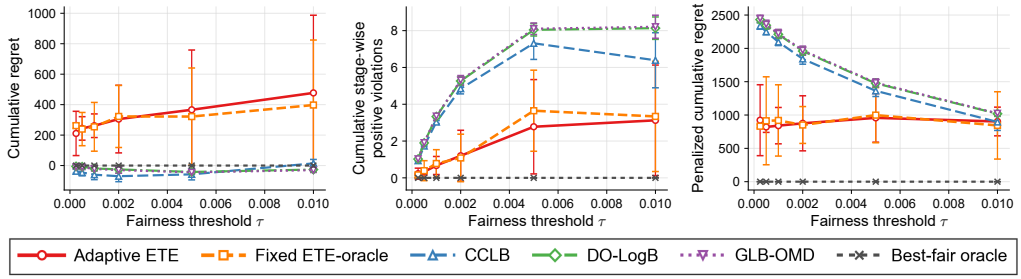
Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update. *Advances in Neural Information Processing Systems*, 38:69244–69277, 2026.

Xingyu Zhou and Bo Ji. On kernelized multi-armed bandits with constraints. *Advances in neural information processing systems*, 35:14–26, 2022.

A Experimental Details and Additional Results



(a) Synthetic experiment with varying factual scale ρ.



(b) Structural causal model with varying fairness threshold τ.

Figure 2: Comparison of cumulative regret, cumulative stage-wise positive violations, and penalized cumulative regret under varying ρ and τ.

A.1 Metrics and implementation details

We report three evaluation metrics and use the following implementation settings throughout. Curves show means over 10 independent runs; shading and error bars span two standard deviations on either side.

Let $a^\star(X_t)$ be the best fair action. In addition to cumulative regret R_t and cumulative stage-wise positive violations V_t from Section 2, we report

$$R_t^{\mathrm{pen}} = \sum_{s=1}^{t} \left[r_{\theta^\star}(X_s, a^\star(X_s)) - r_{\theta^\star}(X_s, A_s) \mathbf{1}\{A_s \in \mathcal{A}_\tau(X_s)\} \right].$$

We use the penalized cumulative regret of Chen et al. (2025). In this metric, an unfair action receives zero reward. The best-fair oracle has zero regret, violation, and penalized regret.

The logistic MLE uses ridge 10^{-6}. Both ETE algorithms begin with 100 uniform exploration rounds and use a confidence radius coefficient of 0.01. Adaptive ETE doubles the number of exploration samples used for estimation when its width test fails.

Fixed ETE-oracle sets its exploration length using an empirical version of $V_\star$ in Definition 3.3. Using $\theta^\star$ and all TK offered context–action pairs, define

$$\widehat{I} = \frac{1}{TK} \sum_{t=1}^{T} \sum_{a=1}^{K} \dot{\mu}(\phi(X_t, a)^\top \theta^\star) \phi(X_t, a) \phi(X_t, a)^\top$$

and, with $q_\Delta = q_r - q_G$,

$$\widehat{V}_\star = \max_{t \in [T], a \in [K]} \max \left\{ \|q_r(X_t, a)\|^2_{\widehat{I}^\dagger}, \|q_\Delta(X_t, a)\|^2_{\widehat{I}^\dagger} \right\}.$$

The algorithm bounds the counterfactual effect by combining separate intervals for the factual and counterfactual rewards. The corresponding empirical scale is

$$\widehat{V} = \max_{t \in [T], a \in [K]} \max \left\{ \|q_r(X_t, a)\|^2_{\widehat{I}^\dagger}, \left(\|q_r(X_t, a)\|_{\widehat{I}^\dagger} + \|q_G(X_t, a)\|_{\widehat{I}^\dagger} \right)^2 \right\},$$

where $\widehat{I}^{\dagger}$ is the Moore–Penrose inverse. The triangle inequality gives $\widehat{V}_\star \le \widehat{V} \le 9\widehat{V}_\star$. Fixed ETE-oracle explores for $m = \max\{100, \lceil(\widehat{V}T^2)^{1/3}\rceil\}$ rounds. This choice omits the theoretical factor $1024^2\ell$; the initial exploration length and radius coefficient are also empirical choices. This benchmark uses $\theta^\star$ and future contexts and is not implementable online.

For CCLB, we use the primal–dual decision rule of Chen et al. (2025) with ridge 1 and a confidence radius coefficient of 2. Its dual variable is projected onto $[0, 10]$, and the model is refitted every 10 rounds. For DO-LogB, we use the OMD update of Chen et al. (2026), replace its path-specific constraints with $|\Delta_\theta(x, a)| \le \tau$, and use 100 initial uniform exploration rounds. DO-LogB and GLB-OMD (Zhang et al., 2026) use a confidence radius coefficient of $1/4$. GLB-OMD does not use the fairness constraint. An algorithm selects an action uniformly if its optimistic fair set is empty.

A.2 Synthetic experiment

This experiment varies the information available for estimating the counterfactual effect while keeping the rewards and fair action sets fixed. For every $t \in [T]$, $a \in [K]$, and $j \in [d]$, draw $S_{t,a,j}$ independently and uniformly from $\{-1, 1\}$. Set $\phi_j(X_t, a) = S_{t,a,j}/\sqrt{d}$ for $j < d$ and $\phi_d(X_t, a) = \rho S_{t,a,d}/\sqrt{d}$. The counterfactual feature of action 1 equals its factual feature, so this action is always fair. For each remaining action, the first and last coordinates of $\phi^G(X_t, a)$ are independent random permutations of $K - 1$ evenly spaced points in $[-1, 1]/\sqrt{d}$. The other coordinates equal those of $\phi(X_t, a)$. We use

$$d = 10, \quad K = 5, \quad T = 10{,}000, \quad \theta^\star = (0.75, -0.45, 0.30, -0.20, 0, \ldots, 0)^\top,$$

and set $\rho = 0.25$ and $\tau = 0.05$ for Figure 1(a). Since $\theta^\star_d = 0$, changing ρ leaves the factual and counterfactual means, fair sets, and best-fair actions unchanged. It changes only the information available in the last coordinate. The mean number of fair actions is 3.00.

Figure 2(a) varies $\rho \in \{1, 1/2, 1/4, 1/8\}$. The corresponding empirical values of $V_\star$, averaged over 10 trials, are 2.58, 3.25, 7.29, and 21.52. Fixed ETE-oracle has larger violation and penalized regret as $V_\star$ increases. Adaptive ETE also has larger values when $V_\star$ is largest.

A.3 Structural causal model

We adapt the structural causal model of Chen et al. (2025) by clipping and rescaling the features and using a logistic reward model in place of their Gaussian-noise threshold model. Draw $A \sim \mathrm{Bernoulli}(1/2)$ and $W \sim \mathcal{N}(0, 1 - A/2)$, where the second parameter of the normal distribution is the variance. Independently draw $U_M, Z_M \sim \mathcal{N}(0, 1)$ and define

$$s(a, w, u) = \begin{cases} |w|/2 + |u|/3, & a = 1, \\ |w|/3 + |u|/2, & a = 0. \end{cases}$$

Using the same U_M and Z_M in the factual and counterfactual worlds, set

$$M = s(A, W, U_M)^{1/2} Z_M, \qquad M^G = s(1 - A, W, U_M)^{1/2} Z_M.$$

For each $a \in [20]$, draw $D_a \sim \mathcal{N}(0, \max\{|W|, |M|\})$. Let $[u]^1_{-1} = \max\{-1, \min\{u, 1\}\}$. The factual feature and its counterfactual counterpart are

$$\phi(X, a) = \frac{1}{\sqrt{5}} \left(A, [W/2.5]^1_{-1}, [M/2.5]^1_{-1}, [D_a/3.5]^1_{-1}, [M/2.5]^1_{-1}[D_a/3.5]^1_{-1}\right)^\top,$$

where $\phi^G(X, a)$ is obtained by replacing (A, M) with $(1 - A, M^G)$. We set

$$\theta^\star = (0, -\sqrt{5}/2, 0, 0, 35\sqrt{5}/12)^\top$$

and draw rewards from $\mathrm{Bernoulli}(\mu(\phi(X, a)^\top \theta^\star))$. Coordinate-wise clipping and the factor $1/\sqrt{5}$ ensure that both feature vectors have norm at most one.

At each round, we use rejection sampling until at least 12 actions satisfy $|\Delta_{\theta^\star}(X_t, a)| \le \tau$, following Chen et al. (2025). Figure 1(b) uses $\tau = 0.001$. The mean number of fair actions is 18.00. Figure 2(b) uses $\tau \in \{0.00025, 0.0005, 0.001, 0.002, 0.005, 0.01\}$. Rejection sampling gives a different accepted context distribution for each value of τ. For all 6 values, one of the two ETE algorithms has the smallest mean penalized regret among the learning algorithms.

B Related Work

Causal bandits and counterfactual fairness. Early causal-bandit work uses known causal structure to share information across interventions, improving best-intervention identification (Lattimore et al., 2016; Sen et al., 2017) and cumulative regret (Lu et al., 2020). Counterfactual fairness asks whether a prediction or outcome would change under a counterfactual intervention on a protected attribute (Kusner et al., 2017); causal and path-specific criteria further distinguish the mechanisms through which disparities arise (Zhang & Bareinboim, 2018; Chiappa, 2019; Wu et al., 2019; Plecko & Bareinboim, 2023). Huang et al. (2022) bring counterfactual fairness to causal bandits using d-separation-based reward estimation and fair policy selection. Lear & Zhang (2025) use causal decomposition to study long-term fairness in Markov decision processes. For logistic rewards, Chen et al. (2025) study a counterfactual mean-reward constraint with a shared unknown parameter, while Chen et al. (2026) consider path-specific fairness. In our setting, the causal structure is supplied through known factual and counterfactual feature maps. We study the resulting constrained bandit problem: whether factual reward observations provide enough information to learn counterfactual feasibility, and the regret–violation tradeoff when they do.

Bandits with constraints. Stage-wise constrained bandits require feasibility at each round (Amani et al., 2019; Khezeli & Bitar, 2020; Moradipari et al., 2020; Amani et al., 2020; Pacchiano et al., 2021). Long-term formulations instead bound cumulative constraint values, allowing strictly feasible rounds to offset violations, often through primal–dual updates (Jenatton et al., 2016; Liu et al., 2021; Zhou & Ji, 2022). The analysis of Chen et al. (2025) uses a signed cumulative violation measure (Appendix F, Eqs. (82) and (101)). We control the stronger measure of cumulative stage-wise positive violations, $V_T = \sum_{t=1}^{T}[|\Delta_{\theta^\star}(X_t, A_t)| - \tau]_+$, which counts each positive violation separately while allowing violations during learning. The learner receives no separate constraint feedback: both reward and fairness must be estimated from the same factual observations. We establish the need for coverage and obtain matching leading upper and worst-case lower rates under a full-rank condition on the factual covariance pooled across actions.

Logistic bandits. Logistic bandits are a special case of generalized linear bandits (Filippi et al., 2010; Li et al., 2017). Subsequent work sharpens their confidence sets, curvature dependence, and computational efficiency, and extends the model to neural reward functions (Faury et al., 2020; Abeille et al., 2021; Jun et al., 2021; Faury et al., 2022; Lee et al., 2024; Zhang et al., 2026; Bae & Lee, 2025). These results estimate rewards at the available action features. In our problem, the same logistic parameter must also predict rewards at counterfactual features and the difference between the two means. The target scale $V_\star$ is the largest inverse-Fisher variance over the reward and effect directions.

Coverage assumptions. Full-rank context-covariance assumptions are used in stochastic linear contextual bandits and best-of-both-worlds analyses (Wu et al., 2020; Bastani et al., 2021; Kuroki et al., 2024). Assumption 3.1 requires the covariance generated by uniform action sampling to be nonsingular but allows the covariance of each action to be singular. Experimental-design methods choose a sampling distribution to reduce uncertainty in the directions that distinguish promising actions (Fiez et al., 2019; Wagenmaker et al., 2021; Mason et al., 2022). Our target scale $V_\star$ measures the largest uncertainty of a reward or effect direction under the Fisher information generated by uniform exploration. Hao et al. (2020) also obtain a $T^{2/3}$ rate by balancing costly exploration against estimation error. The same balance determines the exploration length in our upper bounds.

C Proof of Proposition 3.2

Take $d = 2$, $\mathcal{A} = \{1, 2\}$, and $X_t = x_0$ almost surely for every t. Define

$$\theta_0 = \begin{pmatrix} 2\log 3 \\ 0 \end{pmatrix}, \qquad \theta_1 = \begin{pmatrix} 2\log 3 \\ -2\log 21 \end{pmatrix}, \qquad \Theta = \{(1-\lambda)\theta_0 + \lambda\theta_1 : \lambda \in [0,1]\}.$$

Set $\tau = 1/8$. With e_1, e_2 denoting the standard basis of $\mathbb{R}^2$, choose

$$\phi(x_0, 1) = \frac{1}{2}e_1, \qquad \phi^G(x_0, 1) = \frac{1}{2}(e_1 + e_2), \qquad \phi(x_0, 2) = \phi^G(x_0, 2) = 0.$$

In environment $j \in \{0,1\}$, the true parameter is $\theta^\star = \theta_j$. The context distribution, feature maps, parameter set, and threshold are the same in both environments. The factual reward means satisfy

$$r_{\theta_0}(x_0,1) = r_{\theta_1}(x_0,1) = \mu(\log 3) = \frac{3}{4}, \qquad r_{\theta_j}(x_0,2) = r^G_{\theta_j}(x_0,2) = \frac{1}{2}.$$

The counterfactual reward means of action 1 are

$$r^G_{\theta_0}(x_0,1) = \mu(\log 3) = \frac{3}{4}, \qquad r^G_{\theta_1}(x_0,1) = \mu(-\log 7) = \frac{1}{8}.$$

Thus $\Delta_{\theta_0}(x_0,1) = 0$, $\Delta_{\theta_1}(x_0,1) = 5/8$, and $\Delta_{\theta_j}(x_0,2) = 0$. Both actions are fair under θ_0, so the fair oracle chooses action 1. Only action 2 is fair under θ_1. The one-round $(\text{regret}, \text{violation})$ pairs are

	action 1	action 2
θ_0	$(0,0)$	$(1/4,0)$
θ_1	$(-1/4,1/2)$	$(0,0)$.

Fix any adaptive, possibly randomized policy π, and let

$$N_1(T) = \sum_{t=1}^{T} \mathbf{1}\{A_t = 1\}.$$

Every action has the same Bernoulli reward law under θ_0 and θ_1. Since the context and observed feature maps also agree, π induces the same distribution of observed histories, and hence of $N_1(T)$, in both environments. Write $m = \mathbb{E}_{0,\pi}[N_1(T)] = \mathbb{E}_{1,\pi}[N_1(T)]$. The table gives

$$\mathbb{E}_{0,\pi}[R_T] = \frac{T-m}{4}, \qquad \mathbb{E}_{1,\pi}[V_T] = \frac{m}{2}.$$

If $m \le T/3$, then $\mathbb{E}_{0,\pi}[R_T] \ge T/6$; if $m \ge T/3$, then $\mathbb{E}_{1,\pi}[V_T] \ge T/6$. This proves the first claim.

Action 1 under θ_1 has regret $-1/4$ and violation $1/2$, so its joint loss is $1/4$. Consequently,

$$\mathbb{E}_{0,\pi}[R_T + V_T] = \frac{T-m}{4}, \qquad \mathbb{E}_{1,\pi}[R_T + V_T] = \frac{m}{4}.$$

The maximum of these two quantities is at least $T/8$, which proves the second claim.

The construction satisfies Assumptions 2.1 and 2.2. The set Θ is convex and compact, with $\|\theta\|_2 \le \|\theta_1\|_2 = 2\sqrt{(\log 3)^2 + (\log 21)^2}$ for every $\theta \in \Theta$; we take this value as the parameter bound S. All feature norms are at most one, and action 2 has zero effect for every $\theta \in \Theta$. For $\theta = (1-\lambda)\theta_0 + \lambda\theta_1$, the logits of action 1 are

$$\phi(x_0,1)^\top \theta = \log 3, \qquad \phi^G(x_0,1)^\top \theta = \log 3 - \lambda \log 21 \in [-\log 7, \log 3].$$

The logits of action 2 are zero. Therefore $\kappa = 64/7 < \infty$. Finally, the pooled factual covariance is

$$\Sigma = \frac{1}{2}\phi(x_0,1)\phi(x_0,1)^\top + \frac{1}{2}\phi(x_0,2)\phi(x_0,2)^\top = \frac{1}{8} e_1 e_1^\top.$$

Thus $\theta_1 - \theta_0 \in \ker(\Sigma)$, while $\phi^G(x_0,1)^\top(\theta_1 - \theta_0) = -\log 21 \neq 0$. The coverage assumption excludes this construction.

D Proof of Theorem 4.1

We use the following notation for both lower-bound proofs. In environment j, with true parameter θ_j, write

$$r_j = r_{\theta_j}, \qquad r^G_j = r^G_{\theta_j}, \qquad \Delta_j = \Delta_{\theta_j}, \qquad I_j = I(\theta_j).$$

Let $\mathcal{A}_{\tau,j}(x) = \{a \in \mathcal{A} : |\Delta_j(x,a)| \le \tau\}$ and $a^\star_j(x) \in \arg\max_{a \in \mathcal{A}_{\tau,j}(x)} r_j(x,a)$. Define the one-round joint cost by

$$\mathrm{cost}_j(x,a) = r_j(x, a^\star_j(x)) - r_j(x,a) + [|\Delta_j(x,a)| - \tau]_+.$$

Then $R_T + V_T = \sum_{t=1}^T \mathrm{cost}_j(X_t, A_t)$ in environment j.

D.1 Construction of the worst-case instance

Set the constants

$$v_0 = \dot{\mu}(3/4), \qquad \Delta_E = \frac{1}{2} - \mu(-3/4) - \frac{1}{128} > 0.$$

Indeed, $\mu(-3/4) < 1/3$, so $\Delta_E > 1/6 - 1/128 > 1/128$.

Fix $B \geq 1$, $\rho \in (0, 1]$, and $\alpha \in (0, 1/4]$. Define

$$v_B = \dot{\mu}(B), \qquad g_\alpha = \frac{v_0 \alpha}{8}, \qquad b_\alpha = \log \frac{1/2 + g_\alpha}{1/2 - g_\alpha}. \tag{7}$$

The context X is uniform on $\{-1, 1\}$. Define the two parameters by

$$\theta_0 = (B\sqrt{1+\rho^2}, 0)^\top, \qquad \theta_1 = (B\sqrt{1+\rho^2}, \alpha)^\top, \qquad \Theta = \mathrm{conv}\{\theta_0, \theta_1\}.$$

There are three actions: a high-reward action H, an always-fair baseline S, and an informative action E. Their features and roles are

action	$\phi(x, a)$	$\phi^G(x, a)$	role
H	0	e_2	high reward but unknown fairness
S	$-\dfrac{b_\alpha}{B\sqrt{1+\rho^2}} e_1$	$-\dfrac{b_\alpha}{B\sqrt{1+\rho^2}} e_1$	fair baseline
E	$\dfrac{-e_1 + x\rho e_2}{\sqrt{1+\rho^2}}$	$\dfrac{-e_1 + x\rho e_2}{\sqrt{1+\rho^2}}$	informative but lower reward

(8)

Finally, set the fairness threshold from Section 2 to

$$\Delta_\alpha = \mu(\alpha) - \frac{1}{2}, \qquad \tau = \tau_\alpha = \frac{\Delta_\alpha}{2}.$$

The context distribution, feature maps, parameter set, and threshold are common and known to the learner in both environments. Only the true parameter differs.

The next lemma verifies the assumptions and gives the joint cost of each action in both environments.

Lemma D.1. *For every $B \geq 1$, $\rho \in (0, 1]$, $0 < \alpha \leq 1/4$, and $x \in \{-1, 1\}$, the construction satisfies Assumptions 2.1, 2.2, and 3.1. Its one-round joint costs are*

	$\mathrm{cost}_j(x, H)$	$\mathrm{cost}_j(x, S)$	$\mathrm{cost}_j(x, E)$	$a_j^\star(x)$
$j = 0$	0	g_α	$\geq \Delta_E$	H
$j = 1$	$\geq 3g_\alpha$	0	$\geq \Delta_E$	S.

In particular, every action other than H costs at least g_α in environment 0, and the informative action costs at least Δ_E in both environments.

Proof. First note that $g_\alpha \leq 1/128$ and

$$b_\alpha = 2\,\mathrm{arctanh}(2g_\alpha) \leq \log(65/63) < 1/16.$$

Hence all factual and counterfactual feature norms are at most one, and Θ is convex and compact. Also, the definition of b_α gives

$$\mu(-b_\alpha) = \frac{1}{2} - g_\alpha.$$

Thus H and S have factual rewards $1/2$ and $1/2 - g_\alpha$ in both environments. Since $\dot{\mu}$ decreases on $[0, \infty)$ and $\alpha \leq 1/4 < 3/4$,

$$\Delta_\alpha = \int_0^\alpha \dot{\mu}(z)\, dz \geq v_0 \alpha = 8g_\alpha.$$

Therefore H is fair in environment 0. In environment 1, its effect has absolute value Δ_α, so its violation is $\Delta_\alpha - \tau_\alpha = \Delta_\alpha/2$. Its regret relative to S is $-g_\alpha$, and consequently

$$\mathrm{cost}_1(x, H) = -g_\alpha + \frac{\Delta_\alpha}{2} \geq 3g_\alpha.$$

Actions S and E have identical factual and counterfactual features, so their effects are zero and they are fair in both environments.

It remains to verify the cost of action E in both environments. Its factual logit is $-B$ in environment 0, while in environment 1 it is

$$-B + \frac{x\rho\alpha}{\sqrt{1+\rho^2}} \le -B + \alpha \le -3/4.$$

Hence

$$\mathrm{cost}_0(x, E) = \frac{1}{2} - \mu(-B) \ge \frac{1}{2} - \mu(-3/4) > \Delta_E, \tag{9}$$

$$\mathrm{cost}_1(x, E) = \frac{1}{2} - g_\alpha - \mu\left(-B + \frac{x\rho\alpha}{\sqrt{1+\rho^2}}\right) \ge \frac{1}{2} - \frac{1}{128} - \mu(-3/4) = \Delta_E. \tag{10}$$

Since $\Delta_E > 1/128 \ge g_\alpha$, these calculations prove the cost table and verify that the fair set is nonempty. Finally,

$$\Sigma = \mathbb{E}\big[\phi(X, A)\phi(X, A)^\top\big] = \frac{1}{3}\begin{pmatrix} \dfrac{1 + b_\alpha^2/B^2}{1+\rho^2} & 0 \\ 0 & \dfrac{\rho^2}{1+\rho^2} \end{pmatrix}, \quad \sigma_0^2 = \frac{\rho^2}{3(1+\rho^2)} > 0. \tag{11}$$

The first diagonal entry is no smaller than the second because $\rho \le 1$. Thus the displayed value is exactly the minimum eigenvalue, and Assumption 3.1 holds. □

We next bound $V_\star$ for this construction. Recall that

$$C = \frac{1+\rho^2}{4e^{1/4}\rho^2 v_B}.$$

Lemma D.2. *At both true parameters $j \in \{0, 1\}$,*

$$\frac{C}{3} \le V_{\star,j} \le 3C. \tag{12}$$

Moreover, for the global curvature κ of this common family,

$$\frac{e^{-1/2}}{12}\frac{\kappa}{\sigma_0^2} \le C \le \frac{e^{-1/4}}{12}\frac{\kappa}{\sigma_0^2}. \tag{13}$$

Proof. Let

$$\Sigma_E = \frac{1}{1+\rho^2}\begin{pmatrix} 1 & 0 \\ 0 & \rho^2 \end{pmatrix}$$

and let $I_E(\theta)$ be the Fisher information from action E, averaged over X. For every $\theta \in \Theta$, the logit of action E lies between $-B - 1/4$ and $-B + 1/4$. Applying (43) before taking expectation gives

$$e^{-1/4} v_B \Sigma_E \preceq I_E(\theta) \preceq e^{1/4} v_B \Sigma_E. \tag{14}$$

Because the factual feature of H is zero, the Fisher matrix under uniform action sampling is

$$I(\theta) = \frac{1}{3}\left(I_E(\theta) + \dot\mu(-b_\alpha)\phi(S)\phi(S)^\top\right).$$

We have $q_r(H; \theta) = 0$ and $q_\Delta(H; \theta) = -\dot\mu(\theta_2)e_2$ at action H. The term for action S has no e_2 component. Restricting (49) to $u = se_2$ and using the upper side of (14) gives

$$\begin{aligned} q_\Delta(H;\theta)^\top I(\theta)^{-1} q_\Delta(H;\theta) &\ge \frac{3e^{-1/4}\dot\mu(\theta_2)^2(1+\rho^2)}{v_B\rho^2} \\ &= 12\dot\mu(\theta_2)^2 C. \end{aligned}$$

At either true parameter the second component is 0 or α, so $\dot{\mu}(\theta_2) \geq v_0$. To check the constant, put $y = e^{3/4} \in (1, 3)$; then $v_0 = y/(1+y)^2 > 3/16$ because $3y^2 - 10y + 3 = 3(y - 1/3)(y - 3) < 0$. Hence $12v_0^2 > 1/3$, proving $V_{\star,j} \geq C/3$.

For $V_{\star,j} \leq 3C$, the lower side of (14) implies

$$I(\theta)^{-1} \preceq \frac{3e^{1/4}}{v_B} \Sigma_E^{-1}.$$

We now check each nonzero term in the definition of $V_\star$.

- For the effect of H, we have $\dot{\mu} \leq 1/4$, so
$$\|q_\Delta(H; \theta)\|^2_{I(\theta)^{-1}} \leq \frac{3e^{1/4}}{16 v_B} \frac{1+\rho^2}{\rho^2} < \frac{5}{4} C.$$
- The effect of S is zero. Since $\phi(S)^\top \Sigma_E^{-1} \phi(S) = b_\alpha^2 / B^2 \leq 1/16$,
$$\|q_r(S; \theta)\|^2_{I(\theta)^{-1}} \leq \frac{3e^{1/4}}{256 v_B} \leq \frac{C}{16}.$$
- The effect of E is zero. Moreover, $\phi(x, E)^\top \Sigma_E^{-1} \phi(x, E) = 2$ and $\dot{\mu}(\phi(x, E)^\top \theta) \leq e^{1/4} v_B$. Hence
$$\|q_r(x, E; \theta)\|^2_{I(\theta)^{-1}} \leq 6 e^{3/4} v_B \leq \frac{4}{3} C.$$

For the last two comparisons, use $v_B = \dot{\mu}(B) \leq \dot{\mu}(1) < 1/5$ and $\rho^2/(1+\rho^2) \leq 1/2$. Thus $3C$ bounds all terms and proves the upper bound in (12).

To bound the curvature, note that every factual and counterfactual logit has absolute value at most $B + 1/4$, and action E attains a logit of magnitude B. Since $\dot{\mu}$ is even and decreasing on $[0, \infty)$, (43) gives

$$v_B^{-1} \leq \kappa \leq e^{1/4} v_B^{-1}. \tag{15}$$

Combining (11), the definition $C = (1+\rho^2)/(4e^{1/4}\rho^2 v_B)$, and (15) gives

$$\frac{C}{\kappa/\sigma_0^2} = \frac{1}{12 e^{1/4} \kappa v_B} \in \left[\frac{e^{-1/2}}{12}, \frac{e^{-1/4}}{12} \right],$$

which is (13). □

D.2 Main proof of Theorem 4.1

We first bound the KL divergence between observations in the two environments.

Lemma D.3. *For every context $x \in \{-1, 1\}$,*

$$\mathrm{KL}(P_0(\cdot \mid x, E) \,\|\, P_1(\cdot \mid x, E)) \leq \frac{\alpha^2}{8C}.$$

The factual reward distributions of H and S are identical in the two environments.

Proof. The two logits of E, conditional on x, are

$$z_0 = -B, \qquad z_1 = -B + \frac{x\rho\alpha}{\sqrt{1+\rho^2}}.$$

Their squared difference is $\rho^2\alpha^2/(1+\rho^2)$, and the entire segment between z_0 and z_1 is within the interval $[-B - 1/4, -B + 1/4]$. Lemma G.2 and (43) therefore give

$$\mathrm{KL}(P_0 \| P_1) \leq \frac{e^{1/4} v_B \rho^2 \alpha^2}{2(1+\rho^2)} = \frac{\alpha^2}{8C}.$$

The remaining part of the claim follows from (8). □

The constants in Theorem 4.1 are

$$c_{\text{test}} = \frac{e^{-1/8} v_0}{32}, \qquad C_0 = \frac{64\Delta_E}{c_{\text{test}}}, \qquad c_1 = \Delta_E^{1/3} c_{\text{test}}^{2/3}.$$

Proof of Theorem 4.1. For a given T, choose

$$\alpha = \left(\frac{\Delta_E C}{c_{\text{test}} T} \right)^{1/3}.$$

Since $T \geq C_0 C$ and $C_0 = 64\Delta_E / c_{\text{test}}$, we have $\alpha^3 \leq 1/64$, hence $\alpha \leq 1/4$, so the construction is valid. Fix an arbitrary adaptive, possibly randomized policy π for the rest of the proof.

Let $N_E(t) = \sum_{s=1}^{t} \mathbf{1}\{A_s = E\}$. We consider two cases according to how often the policy selects the informative action.

- **Case 1: action E is used often.** If $\mathbb{E}_j N_E(T) \geq C/\alpha^2$ for one of $j \in \{0, 1\}$, then the bounds (9)–(10) in that same environment give

$$\max_{j \in \{0,1\}} \mathbb{E}_j[R_T + V_T] \geq \frac{\Delta_E C}{\alpha^2}. \tag{16}$$

- **Case 2: E is rare in both environments.** Suppose that $\mathbb{E}_j N_E(T) < C/\alpha^2$ for both j. By Lemmas G.3 and D.3, the histories immediately before round t satisfy

$$\mathrm{KL}(P_0^t \| P_1^t) \leq \frac{\alpha^2}{8C} \mathbb{E}_0 N_E(t-1) < \frac{1}{8}.$$

  By the data-processing statement in Lemma G.4, the same KL bound holds for the distributions of A_t. Applying (46) to the event $\{A_t = H\}$ yields

$$\mathbb{P}_0(A_t \neq H) + \mathbb{P}_1(A_t = H) \geq \frac{1}{2} e^{-1/8}. \tag{17}$$

  In environment 0, every action other than H costs at least g_α. In environment 1, action H costs at least $3g_\alpha$, hence at least g_α. Since the maximum is at least the average, summing (17) over t gives

$$\begin{aligned} \max_{j \in \{0,1\}} \mathbb{E}_j[R_T + V_T] &\geq \frac{1}{2} \sum_{t=1}^{T} [g_\alpha \mathbb{P}_0(A_t \neq H) + g_\alpha \mathbb{P}_1(A_t = H)] \\ &\geq \frac{e^{-1/8} g_\alpha T}{4} = c_{\text{test}} \alpha T. \end{aligned} \tag{18}$$

Finally, the choice of α makes the right-hand sides of (16) and (18) equal:

$$\frac{\Delta_E C}{\alpha^2} = c_{\text{test}} \alpha T = \Delta_E^{1/3} c_{\text{test}}^{2/3} C^{1/3} T^{2/3}.$$

This proves (1). The comparison $V_{\star,j} \asymp C$ follows from Lemma D.2. □

E Proof of Theorem 4.2

We use the environment-specific notation and one-round joint cost defined in Appendix D.

E.1 Construction of the worst-case instance

Fix $B \geq 1$, $\rho \in (0, 1]$, $d \geq 46$, and $K \geq 8$, and define

$$n = d - 1, \qquad M = \min\left\{ \left\lfloor \frac{K}{2} \right\rfloor, \left\lfloor e^{n/32} \right\rfloor \right\}, \qquad p = \frac{K - M - 1}{K}. \tag{19}$$

Here $v_B = \dot{\mu}(B)$ is the same notation as in (7). Define the scale

$$D = \frac{n(1+\rho^2)}{\rho^2 v_B}. \tag{20}$$

Since $n \geq 45$ and $K \geq 8$, we have $M \geq 4$ and $3/8 \leq p < 1$. Both terms defining M are at least 4, and $M \leq K/2$ implies $p \geq 1/2 - 1/K \geq 3/8$. Choose a hypercube packing $u_1, \ldots, u_M \in \{-1/\sqrt{n}, 1/\sqrt{n}\}^n$ satisfying $u_i^\top u_j \leq 1/2$ for $i \neq j$, whose existence is guaranteed by Lemma G.6. For $0 < \alpha \leq 1/4$, define

$$g_\alpha = \frac{v_0 \alpha}{8}, \qquad b_\alpha = \log \frac{1/2 + g_\alpha}{1/2 - g_\alpha}, \qquad c_\alpha = \left(1 + \frac{\alpha^2}{B^2(1+\rho^2)}\right)^{-1/2}. \tag{21}$$

Environment $j \in [M]$ has parameter

$$\theta_j = (B\sqrt{1+\rho^2}, -\alpha u_j)^\top, \qquad \Theta = \mathrm{conv}\{\theta_1, \ldots, \theta_M\}.$$

The context $X = (L, \xi)$ is uniform on $[n] \times \{-1, 1\}$. There are M high-reward candidate actions $H_1, \ldots, H_M$, one safe action S_0, and $K - M - 1$ informative actions (replicas) $E_1, \ldots, E_{K-M-1}$. Their factual features are

$$\begin{aligned} \phi(x, H_i) &= 0, \\ \phi(x, S_0) &= -\frac{b_\alpha}{B\sqrt{1+\rho^2}} e_1, \\ \phi((L, \xi), E_s) &= \frac{-e_1 + \xi\rho e_{L+1}}{\sqrt{1+\rho^2}}. \end{aligned}$$

The counterfactual features of S_0 and every E_s equal their factual features. For a high action,

$$\phi^G(x, H_i) = c_\alpha \left(\frac{\alpha}{B\sqrt{1+\rho^2}} e_1 + (0, u_i)^\top \right). \tag{22}$$

The choice of c_α in (21) makes $\left\|\phi^G(x, H_i)\right\|_2 = 1$. Finally choose the fairness threshold as

$$\tau = \tau_\alpha = \frac{v_0 \alpha}{8}.$$

Lemma E.1. *For every $B \geq 1$, $\rho \in (0, 1]$, $d \geq 46$, $K \geq 8$, $0 < \alpha \leq 1/4$, every environment $j \in [M]$, every context $x = (L, \xi)$, and every informative action E_s, the following hold.*

1. *The construction satisfies Assumptions 2.1, 2.2, and 3.1.*
2. *H_j is the unique fair action among high-reward actions $H_1, \ldots, H_M$. Hence H_j is the fair oracle.*
3. *Every action a that is not H_j has $\mathrm{cost}_j(x, a) \geq v_0\alpha/8$.*
4. *Every informative action satisfies $\mathrm{cost}_j(x, E_s) \geq \Delta_E$.*
5. *$\Delta_j(x, a) = 0$ and $q_\Delta(x, a; \theta) = 0$ for every safe or informative action a and every $\theta \in \Theta$.*

Proof. Item 5 is immediate because $\phi^G(x, a) = \phi(x, a)$ for $a \in \{S_0, E_1, \ldots, E_{K-M-1}\}$.

The factual reward of every high-reward action is $1/2$. At H_i in environment j, we have

$$z_{ij} = \phi^G(x, H_i)^\top \theta_j = c_\alpha \alpha (1 - u_i^\top u_j).$$

For $i = j$, $z_{jj} = 0$, so H_j has zero effect. If $i \neq j$, the construction gives $1 - u_i^\top u_j \geq 1/2$. Moreover,

$$c_\alpha^2 \geq \left(1 + \frac{1}{16}\right)^{-1} = \frac{16}{17}, \quad \text{hence } c_\alpha > \frac{3}{4}. \tag{23}$$

Since every inner product is at least -1, $0 \le z_{ij} \le 2\alpha \le 1/2$. Thus $\dot{\mu}(z) \ge v_0$ along $[0, z_{ij}]$. Also for $i \ne j$,

$$|\Delta_j(x, H_i)| = \mu(z_{ij}) - \frac{1}{2} \ge v_0 z_{ij} \ge \frac{3v_0\alpha}{8}.$$

After subtracting the threshold $v_0\alpha/8$, every wrong high-reward action has violation at least $v_0\alpha/4$. The safe action has reward $1/2 - g_\alpha$, zero effect, and hence joint cost $g_\alpha = v_0\alpha/8$. An informative action has logit

$$-B - \frac{\xi\rho\alpha u_j(L)}{\sqrt{1+\rho^2}} \le -B + \alpha \le -3/4.$$

Its effect is zero, so its joint cost relative to H_j is at least $1/2 - \mu(-3/4) > \Delta_E$. Since $\Delta_E > 1/128 \ge g_\alpha$, this proves items 2, 3, and 4.

It remains to verify the assumptions. The parameter set is convex and compact, with radius at most $\sqrt{2}B + 1/4$. The bound $b_\alpha \le \log(65/63) < 1/16$, proved in Lemma D.1, and the normalization in (22) show that every feature norm is at most one.

Finally, for coverage, average $\phi((L,\xi), E_s)\phi((L,\xi), E_s)^\top$ over (L, ξ):

$$\Sigma_E := \mathbb{E}_{(L,\xi)}\left[\phi((L,\xi), E_s)\phi((L,\xi), E_s)^\top\right] = \frac{1}{1+\rho^2}\begin{pmatrix} 1 & 0 \\ 0 & \frac{\rho^2}{n} I_n \end{pmatrix}.$$

Uniform action sampling selects an informative action with probability p. The high-reward actions have zero factual feature, so

$$\Sigma = \mathbb{E}\left[\phi(X,A)\phi(X,A)^\top\right] = p\Sigma_E + \frac{1}{K}\phi(S_0)\phi(S_0)^\top, \quad \sigma_0^2 = \frac{p\rho^2}{n(1+\rho^2)} > 0. \tag{24}$$

The S_0 term is positive-semidefinite and affects only the first coordinate. The first diagonal entry of $p\Sigma_E$ is no smaller than the other entries, so σ_0^2 is the exact minimum eigenvalue. □

Recall from (20) that

$$D = \frac{n(1+\rho^2)}{\rho^2 v_B}.$$

We identify D as the scale of $V_\star$ in this construction.

Lemma E.2. *For every true parameter $j \in [M]$,*

$$\frac{e^{-1/4}}{20} D \le V_{\star,j} \le \frac{2D}{3}.$$

Moreover, for the global curvature κ of the common family,

$$\frac{3e^{-1/4}}{8}\frac{\kappa}{\sigma_0^2} \le D \le \frac{\kappa}{\sigma_0^2}.$$

Proof. Let $I_E(\theta)$ be the Fisher information from one E_s, averaged over (L, ξ). For every $\theta \in \Theta$, the logit of E_s lies within $1/4$ of $-B$, as in the proof of Lemma D.2. Therefore, by (43),

$$e^{-1/4} v_B \Sigma_E \preceq I_E(\theta) \preceq e^{1/4} v_B \Sigma_E. \tag{25}$$

Since every H_i has zero factual feature, we also have

$$I(\theta) = pI_E(\theta) + \frac{1}{K}\dot{\mu}(-b_\alpha)\phi(S_0)\phi(S_0)^\top.$$

Since $\phi(H_j) = 0$, we have $q_r(H_j; \theta_j) = 0$. Moreover, $\phi^G(H_j)^\top\theta_j = 0$, so $\dot{\mu}(\phi^G(H_j)^\top\theta_j) = \dot{\mu}(0) = 1/4$. Consequently,

$$q_\Delta(H_j; \theta_j) = -\frac{1}{4}\phi^G(H_j),$$

whose $(e_2, \ldots, e_{n+1})$-coordinate block is

$$q' = -\frac{c_\alpha}{4} u_j.$$

The safe term has no $(e_2, \dots, e_{n+1})$-coordinates. The upper side of (25) implies, for every $w \in \mathbb{R}^n$,

$$(0, w)^\top I(\theta_j)(0, w) \le \frac{pe^{1/4} v_B \rho^2}{n(1+\rho^2)} \|w\|_2^2 .$$

Restricting (49) to vectors $(0, w)$ yields

$$\begin{aligned} q_\Delta(H_j; \theta_j)^\top I_j^{-1} q_\Delta(H_j; \theta_j) &\ge \frac{n(1+\rho^2)}{pe^{1/4} v_B \rho^2} \|q'\|_2^2 \\ &= \frac{e^{-1/4} c_\alpha^2 D}{16p} \ge \frac{e^{-1/4} D}{20}, \end{aligned}$$

where we used (23) and $p \le 1$.

For $V_{\star,j} \le 2D/3$, the lower side of (25) implies

$$I(\theta)^{-1} \preceq \frac{e^{1/4}}{p v_B} \Sigma_E^{-1}.$$

Again we check every nonzero term in the definition of $V_\star$.

- H_i has zero q_r. Direct calculation gives

  $$\phi^G(H_i)^\top \Sigma_E^{-1} \phi^G(H_i) = c_\alpha^2 \left(\frac{\alpha^2}{B^2} + \frac{n(1+\rho^2)}{\rho^2} \right) \le 2 v_B D.$$

  The last inequality uses $\alpha^2/B^2 \le 1/16$ and $v_B D = n(1+\rho^2)/\rho^2 \ge 2n$. Since the $\dot{\mu}$ coefficient in q_Δ is at most $1/4$,

  $$\|q_\Delta(H_i; \theta)\|_{I(\theta)^{-1}}^2 \le \frac{e^{1/4} D}{8p} < \frac{2D}{3}.$$

- S_0 has zero effect and $\phi(S_0)^\top \Sigma_E^{-1} \phi(S_0) = b_\alpha^2/B^2 \le 1/16$. Thus

  $$\|q_r(S_0; \theta)\|_{I(\theta)^{-1}}^2 \le \frac{e^{1/4}}{256 p v_B} < \frac{2D}{3}.$$

- E_s has zero effect, $\phi(x, E_s)^\top \Sigma_E^{-1} \phi(x, E_s) = d$, and derivative at most $e^{1/4} v_B$. Hence

  $$\|q_r(x, E_s; \theta)\|_{I(\theta)^{-1}}^2 \le \frac{e^{3/4} v_B d}{p} < \frac{2D}{3}.$$

These comparisons follow from $p \ge 3/8$, $v_B < 1/5$, $d/n \le 46/45$, and $\rho^2/(1+\rho^2) \le 1/2$. This proves $\frac{e^{-1/4}}{20} D \le V_{\star,j} \le \frac{2D}{3}$.

To bound the curvature, use (24) to write

$$D = \frac{p}{v_B \sigma_0^2}.$$

Over Θ, every factual and counterfactual logit has absolute value at most $B + 1/4$. For an action E_s, one of the two context signs gives a logit of magnitude at least B. Hence $v_B^{-1} \le \kappa \le e^{1/4} v_B^{-1}$. Combining this with $p \in [3/8, 1]$ gives

$$\frac{D}{\kappa/\sigma_0^2} = \frac{p}{\kappa v_B} \in \left[\frac{3e^{-1/4}}{8}, 1 \right],$$

equivalently, $\frac{3e^{-1/4}}{8} \frac{\kappa}{\sigma_0^2} \le D \le \frac{\kappa}{\sigma_0^2}$. □

E.2 Main proof of Theorem 4.2

We first bound the KL divergence between observations in different environments.

Lemma E.3. *For every context $x = (L, \xi)$, every pair $j, k \in [M]$, and every informative replica E_s,*

$$\mathrm{KL}(P_j(\cdot \mid x, E_s) \,\|\, P_k(\cdot \mid x, E_s)) \le \frac{2e^{1/4}\alpha^2}{D}.$$

The factual reward distributions of all non-informative actions are identical in all environments.

Proof. Conditionally on $x = (L, \xi)$, the logits of E_s in environments j and k differ by

$$\frac{\xi\rho\alpha}{\sqrt{1+\rho^2}}\big(u_k(L) - u_j(L)\big).$$

Its squared magnitude is at most $4\rho^2\alpha^2/[n(1+\rho^2)]$. The interval between the logits remains between $-B - 1/4$ and $-B + 1/4$. Applying Lemma G.2 gives

$$\mathrm{KL}(P_j \| P_k) \le \frac{1}{2}e^{1/4}v_B\frac{4\rho^2\alpha^2}{n(1+\rho^2)} = \frac{2e^{1/4}\alpha^2}{D}.$$

The remaining part of the claim follows from Lemma E.1. □

Define the constants used in the proof by

$$a_{\mathrm{gen}} = \frac{\Delta_E D \log M}{8e^{1/4}}, \quad c_{\mathrm{gen}} = \frac{v_0}{32}, \quad \widetilde{c}_2 = \left(\frac{\Delta_E}{8e^{1/4}}\right)^{1/3}\left(\frac{v_0}{32}\right)^{2/3}, \quad C_2 = \frac{256\Delta_E}{e^{1/4}v_0}.$$

Proof of Theorem 4.2. For a given T, choose

$$\alpha = \left(\frac{a_{\mathrm{gen}}}{c_{\mathrm{gen}}T}\right)^{1/3}.$$

By (19), $\log M \le \min\{\log K, d\}$. The horizon condition in Theorem 4.2 therefore gives $T \ge C_2 D \log M$, so $\alpha^3 \le 1/64$ and $\alpha \le 1/4$. Thus the construction is valid. Fix an arbitrary adaptive, possibly randomized policy π. Let $N_E(T)$ count the selections of informative actions, and define their average over environments by

$$\overline{N}_E(T) = \frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_j N_E(T).$$

We consider two cases according to the expected number of informative actions:

- **Case 1.** If
$$\overline{N}_E(T) \ge \frac{D\log M}{8e^{1/4}\alpha^2},$$
since every informative selection costs at least Δ_E, we get
$$\max_j \mathbb{E}_j[R_T + V_T] \ge \frac{\Delta_E D\log M}{8e^{1/4}\alpha^2} = \frac{a_{\mathrm{gen}}}{\alpha^2}.$$

- **Case 2.** Now suppose
$$\overline{N}_E(T) < \frac{D\log M}{8e^{1/4}\alpha^2}.$$
Let the true environment index Z be uniform on $[M]$. Lemmas G.3 and E.3 give
$$\begin{aligned} I(Z;\mathcal{H}_t) &\le \frac{1}{M^2}\sum_{j,k=1}^{M}\mathrm{KL}(P_j^t \| P_k^t) \\ &\le \frac{2e^{1/4}\alpha^2}{D}\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_j N_E(t-1) \\ &\le \frac{2e^{1/4}\alpha^2}{D}\overline{N}_E(T) < \frac{1}{4}\log M. \end{aligned} \tag{26}$$

Infer an environment index from the action by setting

$$\widehat{Z}_t = i \quad \text{if } A_t = H_i, \quad \widehat{Z}_t = 1 \text{ if } A_t \text{ is not a high-reward action.}$$

By the data-processing statement in Lemma G.4, $I(Z;\widehat{Z}_t) \le I(Z;\mathcal{H}_t)$. Since $M \ge 4$, Fano's inequality and (26) imply

$$\mathbb{P}(\widehat{Z}_t \ne Z) \ge 1 - \frac{(\log M)/4 + \log 2}{\log M} \ge \frac{1}{4}.$$

The event $\{\widehat{Z}_t \ne Z\}$ is contained in $\{A_t \ne H_Z\}$. Hence

$$\frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_j(A_t \ne H_j) \ge \frac{1}{4}.$$

Every action in this event has cost at least $v_0\alpha/8$. Averaging over environments and summing over t gives the following bound, since the maximum is at least the average:

$$\max_j \mathbb{E}_j[R_T + V_T] \ge \frac{v_0\alpha T}{32} = c_{\mathrm{gen}}\alpha T.$$

The choice of α balances the two cases:

$$\frac{a_{\mathrm{gen}}}{\alpha^2} = c_{\mathrm{gen}}\alpha T = a_{\mathrm{gen}}^{1/3} c_{\mathrm{gen}}^{2/3} T^{2/3} = \widetilde{c}_2 (D\log M)^{1/3} T^{2/3}.$$

Thus every policy has worst-case expected joint loss at least $\widetilde{c}_2(D\log M)^{1/3}T^{2/3}$.

It remains to express the bound in terms of K and d. If $M = \lfloor K/2 \rfloor$, then $M \ge K/3$ because $K \ge 8$. Otherwise, $n \ge 45$ implies

$$M = \lfloor e^{n/32} \rfloor \ge e^{n/64}.$$

Therefore

$$\log M \ge \min\{\log(K/3), n/64\} \ge \frac{1}{128}\min\{\log K, d\},$$

where the last step uses $n = d - 1 \ge d/2$ and $\log(K/3) \ge (\log K)/3$ for $K \ge 8$. Taking $c_2 = 128^{-1/3}\widetilde{c}_2$ proves (2) for the family indexed by $\mathcal{J} = [M]$. The comparison $V_{\star,j} \asymp D$ follows from Lemma E.2. □

F Proof of Theorems 5.3 and 5.4

We first establish Lemma 5.1.

Proof of Lemma 5.1. By definition, the true effect of $a^\star(x)$ belongs to $[-\tau,\tau]$, and also belongs to its effect interval by assumed validity. Thus $a^\star(x)$ is retained. Also by validity and reward optimism,

$$r_{\theta^\star}(x,a^\star(x)) \le \overline{r}(x,a^\star(x)) \le \overline{r}(x,A), \qquad r_{\theta^\star}(x,A) \ge \underline{r}(x,A),$$

which proves the regret bound.

For the violation bound, write the selected effect interval as $[l,u]$ and set $\Delta_A = \Delta_{\theta^\star}(x,A)$. The interval contains Δ_A and intersects $[-\tau,\tau]$. If $\Delta_A > \tau$, then $u \ge \Delta_A$ and $l \le \tau$, hence $\Delta_A - \tau \le u - l$. If $\Delta_A < -\tau$, then $l \le \Delta_A$ and $u \ge -\tau$, hence $-\Delta_A - \tau \le u - l$. If $|\Delta_A| \le \tau$, the LHS is zero. □

We first prove both theorems using the valid intervals and width bound in Lemma 5.2, then prove that lemma. We work on the probability-one event that $\mathcal{A}_\tau(X_t) \ne \varnothing$ for all $t \le T$ unless otherwise stated.

Proof of Theorem 5.3 assuming Lemma 5.2. Work on the event that the confidence intervals are valid. Each of the first m rounds contributes at most one to each of regret and violation. Lemma 5.1 bounds every later contribution by the interval width. Together with Lemma 5.2, this gives

$$\max\{R_T, V_T\} \le m + 1024(T-m)\sqrt{\frac{V_\star \ell}{m}} \le m + 1024T\sqrt{\frac{V_\star \ell}{m}}$$

as desired.

For the specific choice of m, put $a = 1024^2 V_\star \ell T^2$. If that choice of m has $m < T$, then $m \le n_0 + a^{1/3} + 1$ and $m \ge a^{1/3}$. The second term in (6) equals $\sqrt{a/m} \le a^{1/3}$, so

$$\max\{R_T, V_T\} \le n_0 + 1 + 2a^{1/3}.$$

If $m = T$, then $T \le n_0 + a^{1/3} + 1$, while $R_T, V_T \le T$ implies the result. □

The following lemma bounds the sample size at which further exploration stops.

Lemma F.1. *Define* $n_{\mathrm{crit}} = \max\left\{n_0, (1024^2 V_\star \ell T^2)^{1/3}\right\}$. *On the event that Lemma 5.2 guarantees, a checkpoint with* $n_j \ge n_{\mathrm{crit}}$ *cannot trigger subsequent exploration.*

Proof. The definition of n_{crit} gives $1024\sqrt{V_\star \ell / n_j} \le n_j / T$. Combine this with (5). □

Proof of Theorem 5.4 assuming Lemma 5.2. If $n_0 \ge T$, the algorithm explores uniformly throughout, and the result follows from $R_T, V_T \le T$. Suppose $n_0 < T$ and work on the probability $1-\delta$ event of Lemma 5.2 and the probability-one event that $\mathcal{A}_\tau(X_t) \ne \varnothing$ for all $t \le T$.

If the width test never fails, set $\bar{n} = n_0$. Otherwise, let j be the active checkpoint at the last failed test and set $\bar{n} = n_{j+1}$, the target sample size of the ensuing batch, even if the horizon truncates that batch. By Lemma F.1, every failed test occurs at $n_j < n_{\mathrm{crit}}$, and $n_{j+1} \le 2n_j$. Hence

$$\bar{n} \le 2n_{\mathrm{crit}}.$$

Partition the rounds into three types.

- **Exploration rounds.** There are at most $\bar{n}$ exploration rounds. Their total contribution to each of regret and violation is at most $\bar{n}$.
- **Trigger rounds.** Each checkpoint generates at most one discarded trigger round. There are at most $J+1$ checkpoints, so their contribution is at most $J+1$.
- **Exploitation rounds.** Let $\mathcal{T}_j$ be the set of exploitation rounds on which checkpoint j is active. Lemma 5.1 and the exploitation condition $W_j(X_t) \le n_j/T$ in Algorithm 2 imply

  $$\sum_{t \in \mathcal{T}_j} \left[r_{\theta^\star}(X_t, a^\star(X_t)) - r_{\theta^\star}(X_t, A_t)\right] \le |\mathcal{T}_j| \frac{n_j}{T},$$

  and the same bound holds for the violations on $\mathcal{T}_j$. Every active checkpoint satisfies $n_j \le \bar{n}$, so

  $$\sum_j |\mathcal{T}_j| \frac{n_j}{T} \le \frac{\bar{n}}{T} \sum_j |\mathcal{T}_j| \le \bar{n}.$$

  Thus exploitation contributes at most $\bar{n}$, including when there are no exploitation rounds.

Adding the three contributions and using $\bar{n} \le 2n_{\mathrm{crit}}$ gives

$$\max\{R_T, V_T\} \le 2\bar{n} + J + 1 \le 4n_{\mathrm{crit}} + J + 1.$$

Taking the minimum with $R_T, V_T \le T$ proves the bound on $\max\{R_T, V_T\}$.

Finally, (27) implies $n_0 = \widetilde{O}(\kappa d/\sigma_0^2)$, while the definition of ℓ gives $\ell = O(\log(T/\delta) + \min\{\log K, d\})$. For more than one action, $\min\{\log K, d\}$ is bounded below by a positive universal constant, so substituting into n_{crit} and absorbing logarithmic factors gives the stated rate. With a single action, Assumption 2.2 gives $R_T = V_T = 0$, so the rate holds directly. □

F.1 Proof of Lemma 5.2

We index the adaptive algorithm's fits by j, with sample sizes n_j as in Section 5.3. For fixed-length ETE, the same argument applies to the single fit of size m, using m in place of n_j.

To prove Lemma 5.2, we specify Γ_{corr} and establish concentration and existence of the MLE at every checkpoint. Recall that we set $\delta \in (0,1)$ and defined

$$J = \lceil \log_2 T \rceil, \quad M_{\mathrm{dir}} = \min\{KT, 5^d\}, \quad \ell = \log \frac{32(J+1)(T+2)M_{\mathrm{dir}}}{\delta}.$$

Now define

$$\Gamma_{\mathrm{corr}} = \max \left\{ \begin{array}{l} 128 \left[d \log 6 + \log \dfrac{16(J+1)(T+2)M_{\mathrm{dir}}}{\delta} \right], \\ 533\ell, \\ 2 \log \dfrac{8d(J+1)}{\delta}, \\ 2d \log(4eT) + 2 \log \dfrac{16(J+1)}{\delta} \end{array} \right\} \tag{27}$$

and $n_0 = \lceil \kappa \Gamma_{\mathrm{corr}} / \sigma_0^2 \rceil$.

The following result from Jun et al. (2021) is stated with larger numerical constants.

Lemma F.2 (Jun et al. (2021))**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ satisfy $\|x_s\|_2 \le 1$. Let the responses be independent with $R_s \sim \mathrm{Bernoulli}(\mu(x_s^\top \theta^\star))$. Define*

$$H(\theta) = \sum_{s=1}^{n} \dot{\mu}(x_s^\top \theta) x_s x_s^\top$$

and assume $H(\theta^\star) \succ 0$. Let $\widehat{\theta}$ denote the logistic MLE whenever it exists as a finite vector, with $\widehat{H} = H(\widehat{\theta})$. Let n_{eff} be the number of distinct design vectors, fix $z \in \mathbb{R}^d$ with $\|z\|_2 \le 1$, and let $\varepsilon \le e^{-1}$. If

$$\max_{s \le n} \|x_s\|^2_{H(\theta^\star)^{-1}} \le \frac{1}{64\,[d \log 6 + \log((2+n_{\mathrm{eff}})/\varepsilon)]}, \tag{28}$$

then, conditional on the fixed design, there is a bad event B_ε of probability at most ε such that, on $\{\widehat{\theta}$ is finite$\} \setminus B_\varepsilon$,

$$|z^\top(\widehat{\theta} - \theta^\star)| \le 3.5 \sqrt{\log \frac{2(2+n_{\mathrm{eff}})}{\varepsilon}}\ \|z\|_{H(\theta^\star)^{-1}}, \tag{29}$$

$$\frac{1}{\sqrt{2.2}} \|u\|_{H(\theta^\star)^{-1}} \le \|u\|_{\widehat{H}^{-1}} \le \sqrt{2.2}\, \|u\|_{H(\theta^\star)^{-1}} \quad \textit{for every } u \in \mathbb{R}^d, \tag{30}$$

and, in particular,

$$|z^\top(\widehat{\theta} - \theta^\star)| \le 5.2 \sqrt{\log \frac{2(2+n_{\mathrm{eff}})}{\varepsilon}}\ \|z\|_{\widehat{H}^{-1}}. \tag{31}$$

Although the adaptive ETE requests additional exploration in response to its current confidence widths, the observations kept for fitting are not selected using their own contexts. Each batch is scheduled before those contexts are observed, and the triggering round is discarded. They can therefore be viewed as consecutive draws from a single sequence of iid uniform-explorations. The following lemma formalizes this observation.

Lemma F.3. *Let Q denote the law of one uniform factual observation $Z = (X, A, R)$, where $X \sim P_X$, $A \sim \mathrm{Unif}(\mathcal{A})$ independently of X, and*

$$R \mid X, A \sim \mathrm{Bernoulli}(r_{\theta^\star}(X, A)).$$

The adaptive ETE can be coupled with an infinite iid sequence $(Z_s^E)_{s \ge 1}$ drawn from Q such that, whenever checkpoint j is reached, the observations used for fitting are exactly $Z_1^E, \ldots, Z_{n_j}^E$.

Proof. Generate an infinite sequence of uniform exploration observations *in advance* and reveal them only when requested. The initial exploration phase reveals its first n_0 entries. If the width test fails at round t, the algorithm schedules a batch of future rounds for uniform exploration *before* observing any of their contexts. Conditional on everything observed through round t, these future contexts are fresh iid draws from P_X, their actions are independent uniform draws, and their rewards follow the Bernoulli law. The new batch can therefore be identified with the next unused entries of this sequence. □

At every checkpoint j, define the Fisher matrix at the true parameter

$$H_j^\star = \sum_{s=1}^{n_j} \dot{\mu}((\phi_s^E)^\top \theta^\star)\phi_s^E(\phi_s^E)^\top. \tag{32}$$

Here ϕ_s^E denotes the feature vector of the s-th exploration observation used to fit the MLE.

Lemma F.4. *With failure probability at most $\delta/8$, simultaneously at every checkpoint,*

$$H_j^\star \succeq \frac{n_j}{2} I_\star. \tag{33}$$

On this event,

$$(H_j^\star)^{-1} \preceq \frac{2}{n_j} I_\star^{-1} \preceq \frac{2\kappa}{n_j \sigma_0^2} I_d \tag{34}$$

and

$$\max_{s \le n_j} \left\|\phi_s^E\right\|^2_{(H_j^\star)^{-1}} \le \frac{2\kappa}{n_j\sigma_0^2} \le \frac{2}{\Gamma_{\mathrm{corr}}}. \tag{35}$$

Note that the last quantity satisfies the requirement (28) in Lemma F.2.

Proof. By Lemma F.3, the summands of (32) are iid with expectation $I_\star$. Set

$$Z_s = I_\star^{-1/2}\dot{\mu}((\phi_s^E)^\top\theta^\star)\phi_s^E(\phi_s^E)^\top I_\star^{-1/2}.$$

Then $\mathbb{E}Z_s = I_d$ (the $d \times d$ identity matrix). Since Z_s has rank at most one, the bounds $I_\star^{-1} \preceq (\kappa/\sigma_0^2)I_d$, $\|\phi\|_2 \le 1$, and $\dot{\mu} \le 1/4$ give

$$\lambda_{\max}(Z_s) = \dot{\mu}((\phi_s^E)^\top\theta^\star)\left\|\phi_s^E\right\|^2_{I_\star^{-1}} \le \frac{\kappa}{4\sigma_0^2}.$$

Matrix Chernoff with relative deviation $1/2$ now gives

$$\mathbb{P}\left(H_j^\star \not\succeq \frac{n_j}{2} I_\star\right) \le d\exp\left(-\frac{n_j\sigma_0^2}{2\kappa}\right).$$

Since $n_j \ge n_0 \ge \kappa\Gamma_{\mathrm{corr}}/\sigma_0^2$, a union bound over at most $J+1$ checkpoints is bounded by

$$d(J+1)e^{-\Gamma_{\mathrm{corr}}/2} \le \frac{\delta}{8},$$

using the third line of (27). This proves (33). Taking inverses and using $I_\star^{-1} \preceq (\kappa/\sigma_0^2)I_d$ give (34); $\|\phi\|_2 \le 1$ and $n_j \ge n_0$ give (35). □

Every later application of Lemma F.2 has $n_{\mathrm{eff}} \le T$ and

$$\varepsilon \ge \frac{\delta}{16(J+1)M_{\mathrm{dir}}}.$$

The first line of (27) therefore gives

$$\frac{2}{\Gamma_{\mathrm{corr}}} \le \frac{1}{64\left[d\log 6 + \log\frac{16(J+1)(T+2)M_{\mathrm{dir}}}{\delta}\right]} \le \frac{1}{64\left[d\log 6 + \log((2+n_{\mathrm{eff}})/\varepsilon)\right]},$$

which is (28). Moreover, $T \ge 2$ implies $J+1 \ge 2$, and every failure probability used below is at most $\delta/[16(J+1)] \le 1/32 < e^{-1}$. Thus the failure probabilities satisfy the condition in Lemma F.2.

The next lemma shows that the MLE and estimated Fisher inverse are well defined at every checkpoint with high probability.

Lemma F.5. *Apart from an event of probability at most $\delta/8$, every checkpoint satisfying* (33) *has a unique finite MLE, and its estimated Fisher matrix is positive definite.*

Proof. Fix a checkpoint and write its features as $x_s = \phi_s^E$. Write $p_s = \mu(x_s^\top \theta^\star)$. The negative log-likelihood is

$$L(\theta) = \sum_s \left[\log(1 + e^{x_s^\top \theta}) - R_s x_s^\top \theta\right].$$

Define its recession function in direction v by

$$L_\infty(v) := \lim_{c\to\infty} \frac{L(\theta + cv) - L(\theta)}{c} = \sum_{s:R_s=0} (x_s^\top v)_+ + \sum_{s:R_s=1} (-x_s^\top v)_+. \tag{36}$$

where $(a)_+ := \max(a, 0)$. The limit in (36) does not depend on θ.

The lower bound in (33) implies that $\{x_1, x_2, \ldots, x_{n_j}\}$ spans $\mathbb{R}^d$. If some nonzero v has $L_\infty(v) = 0$, then

- $x_s^\top v \le 0$ for $R_s = 0$;
- $x_s^\top v \ge 0$ for $R_s = 1$.

Equivalently,

- $x_s^\top v < 0$ implies $R_s = 0$;
- $x_s^\top v > 0$ implies $R_s = 1$.

Hence the hyperplane $x^\top v = 0$ separates the observations with labels $R_s = 0$ and $R_s = 1$. At least one projection is nonzero because $\{x_1, x_2, \ldots, x_{n_j}\}$ spans $\mathbb{R}^d$. For each nonzero projection, the corresponding term in the loss strictly decreases as c increases. Thus moving in direction v decreases the loss, so no finite θ attains its minimum.

Conversely, if $L_\infty(v) > 0$ for every unit vector v, continuity of L_∞ and compactness of the unit sphere give a positive lower bound for L_∞. Then $L \to \infty$ as $\|\theta\|_2 \to \infty$, thus L attains a finite minimum inside some sufficiently large compact ball.

We have established that the following are equivalent:

- the MLE is not finite;
- there exists a direction with $L_\infty(v) = 0$;
- there exists a hyperplane that separates the two labels.

Fix a unit direction v and let $S(v) = \{s : x_s^\top v \neq 0\}$. For $s \in S(v)$, let $r_s = p_s$ when separation requires label one (i.e., $x_s^\top v > 0$) and $r_s = 1 - p_s$ when it requires label zero (i.e., $x_s^\top v < 0$). Then

$$\mathbb{P}(\text{all required labels match} \mid x_{1:n_j}) = \prod_{s\in S(v)} r_s.$$

Using $-\log r_s \ge 1 - r_s \ge p_s(1 - p_s)$ gives

$$\mathbb{P}(\text{all required labels match} \mid x_{1:n_j}) \le \exp\left(-\sum_{s\in S(v)} p_s(1 - p_s)\right). \tag{37}$$

Since $|x_s^\top v| \le 1$, on (33),

$$\sum_{s\in S(v)} p_s(1 - p_s) \ge \sum_s p_s(1 - p_s)(x_s^\top v)^2$$
$$= v^\top H_j^\star v \ge \frac{n_j}{2} v^\top I_\star v \ge \frac{n_j \sigma_0^2}{2\kappa}.$$

Hence

$$\mathbb{P}(\text{one particular separating pattern}) \le \exp\left(-\frac{n_j\sigma_0^2}{2\kappa}\right).$$

To take a union bound over separating directions, we count the possible sign patterns. The required labels depend only on

$$\operatorname{sign}(x_1^\top v), \dots, \operatorname{sign}(x_{n_j}^\top v)$$

and these signs are constant on each face of the arrangement of hyperplanes $x_s^\top v = 0$. The zero-dimensional face contributes one. For $m \ge 1$, a face whose span has dimension m is contained in the intersection of $d-m$ linearly independent homogeneous hyperplanes; there are at most n_j^{d-m} choices for those hyperplanes. Inside an m-dimensional intersection, we have at most $2\sum_{k=0}^{m-1}\binom{n_j-1}{k} \le 2(d+1)n_j^m$ chambers. Adding the zero-dimensional face and summing over $m = 1, \dots, d$ yields the bound

$$N_{\text{face}} \le 2(d+1)^2 n_j^d \le (4en_j)^d$$

which includes lower-dimensional faces. Combining the two orientations, (37), and $n_j \le T$ gives

$$\mathbb{P}(\text{no finite MLE and (33)} \mid x_{1:n_j}) \le 2(4eT)^d \exp\left(-\frac{n_j\sigma_0^2}{2\kappa}\right).$$

The fourth line of (27) gives

$$2(J+1)(4eT)^d e^{-\Gamma_{\text{corr}}/2} \le \frac{\delta}{8}.$$

Now a union bound over checkpoints proves the finite-MLE part of the claim.

Finally, the Hessian at every finite parameter is

$$\sum_s \dot\mu(x_s^\top\theta) x_s x_s^\top$$

which is positive definite because $\dot\mu(x_s^\top\theta) > 0$ and the design vectors span $\mathbb{R}^d$. Hence L is strictly convex, the finite minimizer is unique, and $\widehat{H}_j$ is invertible. □

The next lemma combines these events with logit confidence bounds that hold simultaneously at the contexts where each checkpoint is used. For adaptive ETE, let $\mathcal{C}_{j,t}$ be the event that, before observing X_t, checkpoint j has been fitted and the algorithm is about to compute $W_j(X_t)$ in the while loop of Algorithm 2. This event is determined by the history before X_t. It includes both exploitation and trigger rounds, and excludes scheduled exploration rounds. For fixed-length ETE, use its sole checkpoint of size m and set $\mathcal{C}_{0,t} = \{t > m\}$.

Lemma F.6. *With probability at least $1-\delta$, every checkpoint MLE that the algorithm calculates is finite and unique, its estimated Fisher matrix is invertible, and simultaneously for every such checkpoint j,*

$$\frac{1}{\sqrt{2.2}}\|u\|_{(H_j^\star)^{-1}} \le \|u\|_{\widehat{H}_j^{-1}} \le \sqrt{2.2}\,\|u\|_{(H_j^\star)^{-1}} \quad \textit{for all } u \in \mathbb{R}^d, \tag{38}$$

and, at every round with $\mathcal{C}_{j,t}$, for all $a \in \mathcal{A}$ and $z \in \{\phi(X_t,a), \phi^G(X_t,a)\}$,

$$|z^\top(\widehat{\theta}_j - \theta^\star)| \le 11\sqrt{\ell}\,\|z\|_{\widehat{H}_j^{-1}} = b_j(z). \tag{39}$$

Proof. Work on the events of Lemmas F.4 and F.5. Their failure probabilities are at most $\delta/8$ each, and (35) ensures Lemma F.2 is applicable. Let $\mathcal{E}_{j,\text{base}}$ denote the event that checkpoint j satisfies the Fisher lower bound (33) and has a finite MLE.

At each checkpoint, apply Lemma F.2 with any fixed unit direction and failure probability $\delta/[16(J+1)]$. The global comparison (30) belongs to the same event. A union bound over checkpoints costs at most $\delta/16$ and proves (38).

We bound the logit errors using one of two arguments, according to which gives the smaller logarithmic factor.

- **Action-wise route:** when $KT \le 5^d$. Fix (j,t,a) and one of the two deterministic maps $z(x) = \phi(x,a)$ or $z(x) = \phi^G(x,a)$. For every deterministic x, apply (31) with

$$\varepsilon = \frac{\delta}{16(J+1)KT}.$$

Because $n_{\text{eff}} \le T$,

$$\log \frac{2(2+n_{\text{eff}})}{\varepsilon} \le \log \frac{32(J+1)KT(T+2)}{\delta} = \ell.$$

Let $F(x)$ be the event of the failure of the resulting bound. The bad-event formulation of Lemma F.2, first conditional on the checkpoint design and then averaged over it, gives $\mathbb{P}(\mathcal{E}_{j,\text{base}} \cap F(x)) \le \varepsilon$ for every fixed x. Since, on $\mathcal{C}_{j,t}$, the base event and estimator are determined before X_t arrives, the tower property gives

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_{j,\text{base}} \cap \mathcal{C}_{j,t} \cap F(X_t)) \\
&= \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_{j,\text{base}} \cap \mathcal{C}_{j,t}} \int \mathbf{1}_{F(x)}\, P_X(dx)\right] \\
&= \int \mathbb{P}(\mathcal{E}_{j,\text{base}} \cap \mathcal{C}_{j,t} \cap F(x))\, P_X(dx) \\
&\le \int \mathbb{P}(\mathcal{E}_{j,\text{base}} \cap F(x))\, P_X(dx) \le \varepsilon.
\end{aligned}$$

There are at most $2(J+1)KT$ choices of direction, so the total failure probability is at most $\delta/8$, with coefficient $5.2 < 11$.

- **Dimension-wise route:** when $5^d < KT$. For each checkpoint choose a $1/2$-net $\mathcal{N}$ of the Euclidean unit sphere with $|\mathcal{N}| \le 5^d$. Conditional on the design, for $h \in \mathcal{N}$ define

$$z_{j,h} = \frac{(H_j^\star)^{1/2} h}{\left\|(H_j^\star)^{1/2} h\right\|_2}.$$

This direction is allowed to depend on the design because Lemma F.2 is applied after conditioning on that design. Apply (29) with $\varepsilon = \delta/[16(J+1)5^d]$. The same logarithmic calculation gives

$$|h^\top (H_j^\star)^{1/2} (\widehat{\theta}_j - \theta^\star)| \le 3.5\sqrt{\ell} \quad (h \in \mathcal{N}).$$

The $1/2$-net inequality $\|v\|_2 \le 2\max_{h\in\mathcal{N}} |h^\top v|$ therefore yields

$$\left\|\widehat{\theta}_j - \theta^\star\right\|_{H_j^\star} \le 7\sqrt{\ell}.$$

For every $z \in \mathbb{R}^d$, Cauchy–Schwarz and (38) give

$$\begin{aligned}
|z^\top (\widehat{\theta}_j - \theta^\star)| &\le \|z\|_{(H_j^\star)^{-1}} \left\|\widehat{\theta}_j - \theta^\star\right\|_{H_j^\star} \\
&\le 7\sqrt{2.2\ell}\, \|z\|_{\widehat{H}_j^{-1}} < 11\sqrt{\ell}\, \|z\|_{\widehat{H}_j^{-1}}.
\end{aligned}$$

This argument has total failure probability at most $\delta/16$.

The failure probabilities are bounded as follows:

event	failure probability
Fisher concentration	$\delta/8$
finite MLEs	$\delta/8$
empirical/true Fisher comparison	$\delta/16$
selected directional route	$\le \delta/8$

whose sum is $7\delta/16 < \delta$. □

We can now prove Lemma 5.2.

Proof of Lemma 5.2. Work on the event of Lemma F.6, where each MLE is finite and each $\widehat{H}_j$ is invertible. By monotonicity of μ, the bound (39) makes the factual and counterfactual reward intervals valid. Subtracting their endpoints gives a valid effect interval.

To prove the width bound (5), combine (38) and (34). For every queried direction z, this gives

$$b_j(z) \le 11\sqrt{2.2\ell}\,\|z\|_{(H_j^\star)^{-1}} \le 11\sqrt{\frac{4.4\ell}{n_j}}\,\|z\|_{I_\star^{-1}}.$$

For $\|z\|_2 \le 1$, using $I_\star^{-1} \preceq (\kappa/\sigma_0^2)I_d$ and $n_j \ge n_0$,

$$b_j(z)^2 \le \frac{121 \cdot 4.4\,\ell\kappa}{n_j\sigma_0^2} \le \frac{121 \cdot 4.4\,\ell}{\Gamma_{\mathrm{corr}}} \le \frac{532.4}{533} < 1.$$

Fix z and write $\eta^\star = z^\top\theta^\star$, $\widehat{\eta} = z^\top\widehat{\theta}_j$, and $b = b_j(z)$. On the confidence event, $|\widehat{\eta} - \eta^\star| \le b \le 1$. Every point u in $[\widehat{\eta} - b, \widehat{\eta} + b]$ therefore satisfies $|u - \eta^\star| \le 2$. Lemma G.1 gives $\dot{\mu}(u) \le e^2\dot{\mu}(\eta^\star)$ throughout that interval. Hence

$$\begin{aligned}\mu(\widehat{\eta} + b) - \mu(\widehat{\eta} - b) &= \int_{\widehat{\eta}-b}^{\widehat{\eta}+b} \dot{\mu}(u)\,du \\ &\le 2e^2\dot{\mu}(\eta^\star)b \\ &\le 341\sqrt{\frac{\ell}{n_j}}\,\|\dot{\mu}(\eta^\star)z\|_{I_\star^{-1}}, \end{aligned} \tag{40}$$

because

$$2e^2 \cdot 11\sqrt{4.4} = 340.987\ldots < 341.$$

For $z = \phi(x,a)$, the last vector in (40) is $q_r(x,a)$. Thus

$$\overline{r}_j(x,a) - \underline{r}_j(x,a) \le 341\sqrt{\frac{\ell}{n_j}}\,\|q_r(x,a)\|_{I_\star^{-1}}. \tag{41}$$

For $z = \phi^G(x,a)$, the last vector is $q_G(x,a)$. The length of the effect interval is the sum of the factual and counterfactual interval lengths, so

$$\overline{\Delta}_j(x,a) - \underline{\Delta}_j(x,a) \le 341\sqrt{\frac{\ell}{n_j}}\left(\|q_r(x,a)\|_{I_\star^{-1}} + \|q_G(x,a)\|_{I_\star^{-1}}\right). \tag{42}$$

Using $\|q_r(x,a)\|_{I_\star^{-1}} \le \sqrt{V_\star}$ and $\|q_r(x,a)\|_{I_\star^{-1}} + \|q_G(x,a)\|_{I_\star^{-1}} \le 3\sqrt{V_\star}$, the right-hand side of (41) is at most $341\sqrt{V_\star\ell/n_j}$, and that of (42) is at most

$$3 \cdot 341\sqrt{\frac{V_\star\ell}{n_j}} = 1023\sqrt{\frac{V_\star\ell}{n_j}} < 1024\sqrt{\frac{V_\star\ell}{n_j}}.$$

Taking the maximum over actions proves (5). □

G TECHNICAL LEMMAS

Lemma G.1. *For all $u, v \in \mathbb{R}$,*

$$e^{-|u-v|}\dot{\mu}(v) \le \dot{\mu}(u) \le e^{|u-v|}\dot{\mu}(v). \tag{43}$$

Proof. Differentiation gives

$$\frac{d}{du}\log\dot{\mu}(u) = 1 - 2\mu(u),$$

whose absolute value is at most one. Thus $\log\dot{\mu}$ is 1-Lipschitz. Integrating between u and v and exponentiating proves (43). □

Lemma G.2. *For all $z_0, z_1 \in \mathbb{R}$,*

$$\mathrm{KL}(\mathrm{Bernoulli}(\mu(z_0)) \,\|\, \mathrm{Bernoulli}(\mu(z_1))) \le \frac{1}{2} \sup_{z \in [z_0, z_1]} \dot{\mu}(z)(z_1 - z_0)^2,$$

where $[z_0, z_1]$ denotes the segment between z_0 and z_1. Note that $z_1 < z_0$ is also possible, but we abuse the interval notation here.

Proof. Write $P_z = \mathrm{Bernoulli}(\mu(z))$ and set $h = z_1 - z_0$. For $A(z) = \log(1 + e^z)$, we have $A'(z) = \mu(z)$. Thus

$$\begin{aligned}
\mathrm{KL}(P_{z_0} \| P_{z_1}) &= A(z_1) - A(z_0) - A'(z_0)h \\
&= h^2 \int_0^1 (1 - s) A''(z_0 + sh)\, ds \\
&\le \frac{1}{2} h^2 \sup_{z \in [z_0, z_1]} \dot{\mu}(z).
\end{aligned}$$

□

For a randomized policy, let $\mathcal{H}_t$ denote the history immediately before the round-t action is randomized, including the current context X_t. Conditional on $\mathcal{H}_t$ and A_t, R_t is a new Bernoulli draw.

Lemma G.3. *Consider environments j, k with the same context law and the same policy. Let P_j^t, P_k^t be the respective distributions of $\mathcal{H}_t$. Then*

$$\mathrm{KL}(P_j^t \| P_k^t) = \mathbb{E}_j \left[\sum_{s=1}^{t-1} \mathrm{KL}(P_j(R_s \mid \mathcal{H}_s, A_s) \,\|\, P_k(R_s \mid \mathcal{H}_s, A_s)) \right]. \tag{44}$$

If only actions in a set $\mathcal{E}$ have different reward distributions and every such conditional observation has KL at most γ, then

$$\mathrm{KL}(P_j^t \| P_k^t) \le \gamma\, \mathbb{E}_j N_{\mathcal{E}}(t-1), \quad \text{where } N_{\mathcal{E}}(s) = \sum_{u=1}^{s} \mathbf{1}\{A_u \in \mathcal{E}\}. \tag{45}$$

Proof. Apply the chain rule for KL divergence to the variables X_s, A_s, R_s. The context terms vanish because the contexts are iid with the same distribution. Conditional on the realized history, the policy uses the same action distribution in both environments, so the action terms also vanish. Only the conditional reward terms remain, giving (44). Bounding each nonzero term by γ gives (45). □

We use the following information-theoretic inequalities.

Lemma G.4. *The following statements hold.*

1. Bretagnolle & Huber (1979); see also Lattimore & Szepesvári (2019, Theorem 14.2). *For probability measures P, Q and every event A,*

$$P(A^c) + Q(A) \ge \frac{1}{2} e^{-\mathrm{KL}(P \| Q)}. \tag{46}$$

2. Data processing (Lattimore & Szepesvári, 2019, Exercise 14.10). *If Y is a measurable function of H, then $\mathrm{KL}(P_Y \| Q_Y) \le \mathrm{KL}(P_H \| Q_H)$. Adjoining an independent random seed with the same law under P and Q does not change KL.*

3. Pairwise mutual-information bound. *If $m \ge 2$, Z is uniform on $[m]$, and $H \mid \{Z = j\} \sim P_j$, then*

$$I(Z; H) \le \frac{1}{m^2} \sum_{j,k=1}^{m} \mathrm{KL}(P_j \| P_k). \tag{47}$$

4. Fano (Cover et al., 1991, Theorem 2.10.1). *Under the setup of item 3, for every estimator $\widehat{Z}(H)$ taking values in $[m]$,*

$$\mathbb{P}(\widehat{Z} \neq Z) \geq 1 - \frac{I(Z;H) + \log 2}{\log m}. \tag{48}$$

For item 3, with

$$\overline{P} = \frac{1}{m}\sum_{k=1}^{m} P_k,$$

we have

$$\begin{aligned} I(Z;H) &= \frac{1}{m}\sum_{j=1}^{m} \mathrm{KL}(P_j \| \overline{P}) \\ &\leq \frac{1}{m}\sum_{j=1}^{m}\frac{1}{m}\sum_{k=1}^{m} \mathrm{KL}(P_j \| P_k), \end{aligned}$$

where the inequality follows from convexity of $Q \mapsto \mathrm{KL}(P_j\|Q)$. This proves (47). For the second claim in item 2, if $U \sim \rho$ is independent of H under both P and Q, then

$$\mathrm{KL}(P_H \otimes \rho \| Q_H \otimes \rho) = \mathrm{KL}(P_H\|Q_H) + \mathrm{KL}(\rho\|\rho) = \mathrm{KL}(P_H\|Q_H).$$

Consequently, if a randomized policy is represented as $A_t = f_t(\mathcal{H}_t, U_t)$ using an environment-independent seed U_t, data processing gives

$$\mathrm{KL}(P_j(A_t \in \cdot) \,\|\, P_k(A_t \in \cdot)) \leq \mathrm{KL}(P_j^t \| P_k^t).$$

Finally, the form (48) follows from Fano's inequality. Writing $p_e = \mathbb{P}(\widehat{Z} \neq Z)$ and using $\mathsf{H}(Z) = \log m$,

$$\log m - I(Z;H) = \mathsf{H}(Z \mid H) \leq h_2(p_e) + p_e \log(m-1) \leq \log 2 + p_e \log m.$$

Rearranging gives (48).

Lemma G.5. *For every positive-definite matrix M and vector q,*

$$q^\top M^{-1} q = \sup_{u \in \mathbb{R}^d} \left(2q^\top u - u^\top M u\right). \tag{49}$$

Proof. Complete the square:

$$2q^\top u - u^\top M u = q^\top M^{-1} q - \left\|u - M^{-1}q\right\|_M^2.$$

The supremum is attained at $u = M^{-1}q$. □

Lemma G.6. *For every integer $n \geq 45$ and every $m \leq \lfloor e^{n/32} \rfloor$, there exist vectors*

$$u_1, \ldots, u_m \in \left\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right\}^n$$

such that $u_i^\top u_j \leq 1/2$ whenever $i \neq j$.

Proof. Draw the vectors independently and uniformly. For a fixed pair $i \neq j$,

$$u_i^\top u_j = \frac{1}{n}\sum_{k=1}^{n} \xi_k,$$

where the ξ_k are independent Rademacher variables, taking values $+1$ and -1 with equal probability. Hoeffding's inequality gives

$$\mathbb{P}(u_i^\top u_j > 1/2) = \mathbb{P}\left(\sum_{k=1}^{n} \xi_k > \frac{n}{2}\right) \leq e^{-n/8}.$$

A union bound over fewer than $m^2/2$ unordered pairs gives a failure probability of at most

$$\frac{m^2}{2} e^{-n/8} \leq \frac{1}{2} e^{-n/16} < 1.$$

Hence some choice of vectors satisfies all the required pairwise inequalities. □

USE OF GENERATIVE-AI TOOLS

The authors defined the parameter $V_\star$ that quantifies instance difficulty (Definition 3.3), developed Algorithm 1, and established its matching upper bound. GPT-5.6 Sol assisted in developing the hard instances, the lower-bound proofs, Algorithm 2, and the proof of its upper bound. It also assisted with manuscript polishing, reference exploration, and simulation code development. The authors take responsibility for the final arguments and presentation.