MINIMAX-OPTIMAL REGRET WITH NEAR-MINIMAL POLICY UPDATES IN LINEAR CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can linear contextual bandits attain minimax-optimal regret up to constant factors uniformly across the small- and large-action regimes? How many policy updates are necessary to attain this statistical optimum? We introduce `GATE` for d-dimensional linear contextual bandits with at most K actions per round and oblivious adversarial contexts. `GATE` achieves $\mathcal{O}(\sqrt{dT\log(1+T/d)\min\{d,\log K\}})$ regret, with $\mathcal{O}(d\log(1+T/d)\log(2+\log(2+\min\{d,\log K\}))/\log(2+\min\{d,\log K\}))$ policy updates on every sample path. To the best of our knowledge, this is the first constant-factor minimax guarantee (i.e., the sharpest regret bound) uniformly across the small- and large-action regimes, without additional logarithmic or iterated-logarithmic multipliers. At the same regret level, a feedback lower bound establishes policy-update minimality up to an iterated-logarithmic factor in $\min\{d,\log K\}$. The central technical result is a sharp Gaussian information inequality with additive, rather than multiplicative, dependence on action-set complexity and design growth. Independent Gaussian testing and exploration, direct prediction-error control, and joint threshold calibration yield the stated regret and policy-update bounds without computing Gaussian width.

1 INTRODUCTION

In linear contextual bandits (Auer, 2002; Li et al., 2010; Chu et al., 2011; Agrawal & Goyal, 2013), the pursuit of minimax-optimal regret has led to sharp upper and lower bounds (Abbasi-Yadkori et al., 2011; Foster et al., 2021; Li et al., 2021; 2024). Yet, surprisingly, a guarantee that is minimax-optimal up to constant factors uniformly across *all* action regimes has remained elusive. In the *small-action* regime ($\log K \le d$), Li et al. (2024) obtain the optimal dependence on the number of actions only up to residual iterated-logarithmic factors, so their bound is not yet minimax-optimal up to constant factors. In the *large-action* regime ($\log K > d$), the optimal regret rate is already achievable through regression reductions (Foster et al., 2021), but these reductions do not attain the optimal K-dependent rate throughout the small-action regime.

The open problem is therefore to achieve minimax optimality up to constant factors *uniformly across both the small- and large-action regimes*, including the optimal K-dependent rate in the small-action regime. This requires removing the residual iterated-logarithmic factors without worsening the dependence on d, K, or T.

Regret, however, is not the only resource at stake: attaining it also requires incorporating new feedback into the policy. Under batched feedback or constrained deployment, the policy can be updated only occasionally, so a fixed policy must serve new contexts between updates. Limited-update methods therefore treat policy changes separately from per-round action selection (Abbasi-Yadkori et al., 2011; Ruan et al., 2021; Karbasi et al., 2021; Han et al., 2026). Longer batches reduce the number of updates but delay the use of new rewards, which can inflate regret. The question is thus not merely whether few updates suffice, but how many are *necessary* to retain constant-factor minimax regret, and whether that minimum can be approached.

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.

In this work, we address the following research questions:

1. *Can linear contextual bandits attain minimax-optimal regret up to constant factors uniformly across the small- and large-action regimes?*
2. *Can this statistical optimum also be attained with nearly the minimum number of policy updates?*

Our contributions. We answer these questions with `GATE` (Gaussian-Augmented Thresholded Elimination) under Assumption 1 and the comparison ranges below.

- **First constant-factor minimax regret across the full small- and large-action regimes.** `GATE` achieves $\mathcal{O}(\sqrt{dT\log(1+T/d)\min\{d,\log K\}})$ regret. For every fixed $\delta > 0$ and $T \geq d^{2+\delta}$, a self-contained, globally normalized lower-bound construction certifies constant-factor sequential minimax optimality (Theorem 1, Proposition 1, and Section I). To our knowledge, this is the first uniform constant-factor guarantee across full small- and large-action regimes, with no additional logarithmic or iterated-logarithmic multiplier. Thus, in this range, `GATE` attains the *sharpest possible worst-case expected-regret bound up to constant factors*, even against fully sequential learners with immediate reward feedback.
- **Policy updates optimal up to an iterated-logarithmic factor.** At the same constant-factor minimax regret level, the pathwise update bound in Theorem 1 matches a feedback lower bound within an $\mathcal{O}(\log(2+\log(2+\min\{d,\log K\})))$ factor under Theorem 2's conditions. For sufficiently large d and K, the remaining factor is only $\mathcal{O}(\min\{\log\log d, \log\log\log K\})$ with no additional horizon-dependent multiplier.
- **A sharp additive information inequality and direct prediction-error control.** We prove a worst-case-sharp Gaussian information inequality relating selection under a frozen design to information gained under the accumulated design. Its *additive*, rather than multiplicative, dependence on action-set complexity and design growth permits longer batches without a multiplicative information penalty. Layer-specific updates and independent testing and exploration preserve the required conditional laws. Direct bounds on errors beyond confidence radii and joint threshold calibration then yield the stated regret and policy-update bounds without computing Gaussian width or paying an extra depth factor.
- **Horizon-free learning without extra logarithmic factors.** Our horizon-free extension takes only (d, K) and retains observations in a single run. It matches the known-horizon regret bound up to constants after any fixed number of rounds T. Its policy-update bound holds pathwise at every round, even under data-dependent termination.

Thus, `GATE` combines constant-factor minimax regret with near-minimal policy updates across both regimes. Table 1 compares limited-update methods; Sections 2 and K discuss sequential benchmarks.

2 Related Work

Sequential linear contextual bandits. Li et al. (2024) obtain $\mathcal{O}(\sqrt{dT\log T\log K})$ regret with a residual $\mathrm{poly}(\log\log(KT))$ multiplier. Li et al. (2021) give $\mathcal{O}(d\sqrt{T\log T}\log\log T)$ expected regret for infinite action sets. Regression reductions offer a complementary approach (Foster & Rakhlin, 2020). In particular, `SquareCB.Lin` (Foster et al., 2021) yields $\mathcal{O}(d\sqrt{T\log(1+T/d)})$ expected regret under our noise model. Its affine-dimension refinement also matches the minimax scale for bounded K (Section K). These results establish important special cases, but not a uniform constant-factor guarantee with the sharp $\log K$ dependence throughout the small-action regime. `GATE` closes this remaining gap and attains the resulting constant-factor optimum uniformly across the small- and large-action regimes with near-minimal policy updates.

Batched learning and limited adaptivity. Frozen estimates, elimination layers, and determinant-based updates are established tools (Ruan et al., 2021). Their multiscale policy-switch lower bound also underlies our feedback construction; Sections 5.4 and J explain the refinement needed at the exact minimax regret target. Batched Thompson sampling and prescribed-batch elimination provide alternative update mechanisms (Karbasi et al., 2021; Han et al., 2026); Table 1 states their compared

Table 1: Simultaneous regret and policy-update guarantees under oblivious adversarial contexts. Bounds use each method's native settings (Section L.5) for $T \geq d \geq 2$, subject to its source assumptions; native reward access and update semantics differ. The GATE row applies to both known-horizon and horizon-free versions, with pathwise updates.

Algorithm / Paper	Worst-case regret	Policy updates
RS-OFUL (Abbasi-Yadkori et al., 2011)	$\mathcal{O}(d\sqrt{T}\log(1+T/d))$	$\mathcal{O}(d\log(1+T/d))$
B-TS-C (Karbasi et al., 2021)	$\mathcal{O}(d\sqrt{T\min\{d,\log K\}}\log T)$	$\mathcal{O}(K\log T)$
RarelySwitch-SupLinUCB (Ruan et al., 2021)	$\mathcal{O}(\sqrt{dT}\log(dKT)\log T\log d)$	$\mathcal{O}(d\log T\log d)$
SupSBUCB† (Han et al., 2026)	$\mathcal{O}(\sqrt{dT\log(KT)}\log T)$	$\Theta(\sqrt{dT})$
GATE‡ (Ours)	$\mathcal{O}(\sqrt{dT\log(1+T/d)\min\{d,\log K\}})$	$\mathcal{O}\left(\frac{d\log(1+T/d)\Lambda_{d,K}}{\log(2+\min\{d,\log K\})}\right)$

† SupSBUCB requires $T \geq d^2$ and $K = \mathcal{O}(\mathrm{poly}(d))$; its update count uses $\lceil\sqrt{dT}\rceil$ uniform batches.
‡ $\Lambda_{d,K} := \log(2+\log(2+\min\{d,\log K\}))$.

guarantees. Fixed-action methods (Yu & Oh, 2025) and stochastic-context methods (Ruan et al., 2021; Hanna et al., 2023; Zhang et al., 2025; Yu & Oh, 2026) use additional structure to reduce adaptivity. GATE instead attains constant-factor minimax regret with a near-minimal number of policy updates under oblivious adversarial contexts, in the ranges of Section 3. Its adaptive endpoints preserve fixed within-batch policies (Yu & Oh, 2026). Interaction-dependent adaptive action sets form a stronger model (Jin, 2026), distinct from the oblivious model studied here.

Gaussian methods. Gaussian width has been used in linear-bandit experimental design (Wagenmaker et al., 2021). Our new inequality separates action-set complexity from design growth additively, and is sharp up to constants. Independent testing and exploration, together with direct prediction-error control, convert this geometric property into regret and update guarantees. We use the scalar convex-order theorem of Zhang (2026) as an external input, applied coordinatewise to independent reward noises.

3 Preliminaries and Main Results

3.1 Model and feedback protocol

At round t, the learner observes $\mathcal{A}_t = (x_{t,a})_{a\in[K_t]}$ with $x_{t,a} \in \mathbb{R}^d$ and $1 \leq K_t \leq K$, chooses $a_t \in [K_t]$, and generates the scalar reward $r_t = x_{t,a_t}^\top \theta_\star + \eta_t$. It sees no future action sets. Regret compares the selected action with the best available action in the same round:

$$R_T = \sum_{t=1}^{T}\left(\max_{a\in[K_t]} x_{t,a}^\top\theta_\star - x_{t,a_t}^\top\theta_\star\right).$$

Assumption 1. *The context sequence and $\theta_\star$ are fixed before interaction, independently of learner randomness and noise, with $\|\theta_\star\|_2, \|x_{t,a}\|_2 \leq 1$. The noises are independent, centered, and 1-sub-Gaussian: $\mathbb{E}e^{v\eta_t} \leq e^{v^2/2}$ for every $v \in \mathbb{R}$.*

Definition 1 (Fixed within-batch policy)**.** *Let S_T count policy updates over T rounds, excluding initialization and terminal feedback; the run has S_T+1 batches. For $0 = T_0 < \cdots < T_{S_T+1} = T$, a randomized kernel χ_j is fixed before the first context of batch j. Each round draws $a_t \sim \chi_j(\mathcal{A}_t)$ with fresh independent randomness. The kernel cannot read earlier contexts, actions, or rewards of that batch. Before the next context, a separate monitor may use observed history, current-batch context/action state, and consumed randomness, but no unobserved rewards. At each nonterminal batch endpoint, the learner requests and observes the previously unrevealed rewards of the actions played in that batch, then deploys the next kernel before observing the next context.*

More generally, each feedback query reveals all selected-action rewards since the previous query; Q_T counts nonterminal queries. The lower-bound class allows actions to use time and observed context/action history between queries, and is therefore larger than Definition 1. For GATE, each nonterminal feedback query deploys one updated policy, and every policy update includes such a query; hence $Q_T = S_T$. We use $\|x\|_V = (x^\top V x)^{1/2}$; logarithms are natural unless a base is displayed. Asymptotic notation hides constants, not logarithms.

3.2 Simultaneous upper bounds and their optimality

Theorem 1 (Regret and policy updates). *Under Assumption 1, for $d \geq 1$, $K \geq 2$, and every deterministic evaluation time $T \geq 1$, known-horizon and horizon-free GATE satisfy Definition 1 and*

$$\mathbb{E}R_T \leq \|R_T\|_{L_2} = \mathcal{O}\Big(\sqrt{dT\log(1+T/d)\min\{d,\log K\}}\Big), \tag{1}$$

$$S_T = \mathcal{O}\left(d\log(1+T/d)\cdot\frac{\log(2+\log(2+\min\{d,\log K\}))}{\log(2+\min\{d,\log K\})}\right) \quad \textit{pathwise}. \tag{2}$$

The horizon-free algorithm takes only (d, K) and retains its observations. Its update bound holds even when termination depends on the observed history.

Here $\|R_T\|_{L_2} = (\mathbb{E}R_T^2)^{1/2}$ averages over noise and learner randomness for a fixed instance. For the horizon-free version, T only indexes the evaluation prefix: the same algorithm satisfies (1) for every deterministic T, without knowing T in advance or restarting.

Proposition 1 (Minimax optimality of GATE). *Let $\mathfrak{R}_{\mathrm{seq}}(T, d, K)$ be the minimax expected regret with immediate reward feedback under Assumption 1. For fixed $\delta > 0$ and all integers $d, K \geq 2$, $T \geq d^{2+\delta}$,*

$$\mathfrak{R}_{\mathrm{seq}}(T,d,K) = \Theta_\delta\Big(\sqrt{dT\log(1+T/d)\min\{d,\log K\}}\Big). \tag{3}$$

Both known-horizon and horizon-free GATE attain this rate with the pathwise updates in Theorem 1; constants depend only on δ.

Combine Theorem 1 with the self-contained, globally normalized lower bound proved in Section I.

Theorem 2 (Feedback lower bound at the minimax regret target). *Fix $\delta > 0$ and a constant $C_{\mathrm{reg}} > 0$. For all integers $d \geq 4$, $K \geq 2$, and T satisfying $T \geq \max\{d^{2+\delta}, 16C_{\mathrm{reg}}^2 d^2 \min\{d, \log K\}\log(1 + T/d)\}$, every general batched-feedback learner satisfying*

$$\sup_\nu \mathbb{E}_\nu R_T \leq C_{\mathrm{reg}}\sqrt{dT\log(1+T/d)\min\{d,\log K\}}$$

requires

$$\sup_\nu \mathbb{E}_\nu Q_T = \Omega_{\delta, C_{\mathrm{reg}}}\left(\frac{d\log T}{\log(2+\min\{d,\log K\})}\right). \tag{4}$$

The implicit positive constant depends only on δ, C_{reg} and is uniform over the stated range and all admissible learners. The feedback budget is uniform over admissible instances ν.

Constant-factor regret and near-minimal policy updates. Theorem 1 holds for every fixed evaluation time; Proposition 1 identifies its regret as minimax-optimal up to constants for fixed $\delta > 0$ and $T \geq d^{2+\delta}$. The constants are independent of T, d, K. Under Theorem 2's conditions, fix the same C_{reg} large enough for the upper bound. At this regret level, the policy-update complexity is minimax-optimal up to an $\mathcal{O}(\min\{\log\log d, \log\log\log K\})$ factor for sufficiently large d and K, with no additional horizon-dependent multiplier. The comparison concerns feedback queries in the stated protocol, rather than estimator updates when rewards are already observed online.

4 Algorithm: GATE

4.1 Design principles

Three obstacles separate sharp sequential regret from a matching limited-update guarantee: multiplicative information loss from stale designs, uniform-confidence costs, and repeated cost across

Algorithm 1 GATE (Gaussian-Augmented Thresholded Elimination)

Input: horizon T, feature dimension d, upper bound K on actions per round.

1: $H \leftarrow T$; $(L, \beta_0, (\varepsilon_s, z_s)_{s<L}, (\ell_s)_{1\le s<L}) \leftarrow \mathrm{Parameters}(H, d, K)$.
2: $(V_s, b_s) = (\bar{V}_s, \bar{b}_s) = (I_d, 0)$ $(s = 0, \dots, L-1, \mathrm{all})$; $Q \leftarrow \varnothing$.
3: **for** $t = 1, \dots, T$ **do**
4: $\quad \mathcal{R}_t \leftarrow \{s < L : \log\det V_s - \log\det \bar{V}_s \ge z_s\}$.
5: $\quad$ **if** $\mathcal{R}_t \ne \varnothing$ or $\log\det V_{\mathrm{all}} - \log\det \bar{V}_{\mathrm{all}} \ge z_0$ **then**
6: $\quad\quad$ receive $(r_i)_{i\in Q}$; $b_s \leftarrow b_s + \sum_{i\in Q: J_i = s} x_{i,a_i} r_i$ $(s < L)$; $b_{\mathrm{all}} \leftarrow b_{\mathrm{all}} + \sum_{i\in Q} x_{i,a_i} r_i$.
7: $\quad\quad$ $(\bar{V}_s, \bar{b}_s) \leftarrow (V_s, b_s)$ $(s \in \mathcal{R}_t \cup \{\mathrm{all}\})$; $Q \leftarrow \varnothing$.
8: $\quad$ observe $\mathcal{A}_t$; $x_a = x_{t,a}$; $A \leftarrow [K_t]$; $J_t \leftarrow \perp$.
9: $\quad B \leftarrow \{a \in A : U_0(a) \ge \max_{b\in A} L_0(b)\}$.
10: $\quad$ **if** $\max_{a\in B} \rho_0(a) > \varepsilon_0$ **then**
11: $\quad\quad a \leftarrow \mathrm{Best}_0(B)$.
12: $\quad\quad$ **if** $\rho_0(a) \le \varepsilon_0$ and $C_0(a; B) \le 4\varepsilon_{L-1}$ **then** $a_t \leftarrow \mathrm{Best}_{\mathrm{all}}(\{b \in B : C_0(b; B) \le 4\varepsilon_{L-1}\})$; **go to 24**.
13: $\quad\quad a \leftarrow \mathrm{Best}_0(\{b \in B : \rho_0(b) > \varepsilon_0\})$.
14: $\quad\quad$ **if** $C_0^{\Delta}(a; [K_t]) > 4\rho_0(a)$ **then** $a \leftarrow \mathrm{Best}_0(\{b \in B : \rho_0(b) > \varepsilon_0,\ C_0(b; B) \le 4\rho_0(b)\})$.
15: $\quad\quad$ $(a_t, J_t) \leftarrow (a, 0)$; **go to 24**.
16: $\quad A \leftarrow B$.
17: $\quad$ **for** $s = 1, \dots, L-1$ **do**
18: $\quad\quad g_{\mathrm{test}} \sim N(0, I_d)$; compute the common radius ρ_s by (6).
19: $\quad\quad$ **if** $\rho_s > \varepsilon_s$ **then**
20: $\quad\quad\quad g_{\mathrm{sel}} \sim N(0, I_d)$ independently; $A_{\mathrm{sel}} \leftarrow \{a \in A : |g_{\mathrm{sel}}^\top \bar{V}_s^{-1/2} x_a| \ge \max_{b\in A} |g_{\mathrm{sel}}^\top \bar{V}_s^{-1/2} x_b|/\sqrt{2}\}$.
21: $\quad\quad\quad$ $(a_t, J_t) \leftarrow (\mathrm{Best}_{<s}(A_{\mathrm{sel}}), s)$; **go to 24**.
22: $\quad\quad$ read $\bar{b}_s$; $A \leftarrow \{a \in A : U_s(a) \ge \max_{b\in A} L_s(b)\}$.
23: $\quad a_t \leftarrow \mathrm{Best}_{\mathrm{all}}(A)$.
24: $\quad$ play a_t; $Q \leftarrow Q \cup \{t\}$; $V_{\mathrm{all}} \leftarrow V_{\mathrm{all}} + x_{a_t} x_{a_t}^\top$; **if** $J_t \ne \perp$ **then** $V_{J_t} \leftarrow V_{J_t} + x_{a_t} x_{a_t}^\top$.
25: receive and aggregate the remaining rewards as in Line 6; clear Q; no update.

layers. GATE's additive Gaussian information inequality permits longer batches; direct bounds on errors beyond randomized confidence radii avoid accurate width estimation; and jointly calibrated thresholds prevent an additional depth factor. Layer-specific updates and independent testing and exploration preserve the required conditional laws.

Each observation is assigned to at most one elimination layer. The first layer, indexed by 0, uses a self-normalized confidence bound each round. When it explores, it selects an uncertain action whose suboptimality is controlled by its own radius. Subsequent layers receive actions already screened by earlier layers, so their exploration regret is controlled before they choose an action. Earlier-layer reward estimates rank their exploration candidates without using the current layer's rewards. A pooled estimator uses every received selected-action observation, but ranks actions only after elimination has bounded the regret of all remaining candidates. Thus reward-based ranking coexists with the conditional independence needed for the subsequent layers' prediction bounds.

4.2 Algorithm description

Lines 1–2: parameters and estimates. The explicit map in Section G determines the depth L, confidence coefficient β_0, exploration thresholds ε_s, update thresholds z_s, and test coefficients ℓ_s from (T, d, K). There are L layers, indexed by $0, \dots, L-1$, with

$$L = \max\left\{2, \left\lceil \log_2\left(1 + \tfrac{1}{2}\log\left[d\left(1 + \log(1 + T/d)\right)\right]\right)\right\rceil\right\}.$$

Thus the depth grows doubly logarithmically in $d[1 + \log(1 + T/d)]$. Layer s accumulates $V_s = I + \sum_{i<t: J_i = s} x_{i,a_i} x_{i,a_i}^\top$ and the feature–reward sum b_s over its received observations. Bars denote statistics fixed at that layer's last update; the pool, indexed by all, omits the layer restriction. Define

$$\widehat{\theta}_j = \bar{V}_j^{-1} \bar{b}_j, \quad \widehat{\mu}_j(a) = \mathrm{clip}_{[-1,1]}(x_a^\top \widehat{\theta}_j), \quad u_j(a) = \|x_a\|_{\bar{V}_j^{-1}},$$
$$\mathrm{Best}_j(A) = \min \arg\max_{a\in A} \{\widehat{\mu}_j(a) + u_j(a)\}.$$

The buffer Q stores pending selected observations; ties use the smallest input index.

Lines 3–7: adaptive batch endpoints. Before each context, layer s requests an update after z_s log-determinant growth; the pool uses z_0. Each query reveals all pending rewards and adds each once to its assigned layer, if any, and the pool. Only requesting layers refresh their snapshots; the pool refreshes at every query. Concurrent requests count once. Accumulated designs determine endpoints; action selection uses fixed snapshots. For later layers, layer-specific refreshes prevent reward-dependent requests elsewhere from altering the snapshot training set, preserving the conditional noise law.

Lines 8–10: first-layer elimination. Set $x_a = x_{t,a}$, $A = [K_t]$, and $J_t = \perp$ for an unassigned observation. The first-layer radius is $\rho_0(a) = \beta_0 u_0(a)$. Each layer uses clipped intervals

$$L_s(a) = \max\{-1, \widehat{\mu}_s(a) - \rho_s(a)\}, \qquad U_s(a) = \min\{1, \widehat{\mu}_s(a) + \rho_s(a)\}.$$

The nonempty set B retains actions whose upper endpoint reaches the largest lower endpoint. If every radius in B is at most ε_0, pass B to layer 1; otherwise first check the stopping rule below.

Lines 11–12: select a low-regret final action. Define the interval-based and action-difference suboptimality bounds

$$C_0(a; B) = \max_{b \in B} U_0(b) - L_0(a), \quad C_0^{\Delta}(a; A) = \max_{b \in A}\{(x_b - x_a)^\top \widehat{\theta}_0 + \beta_0 \|x_b - x_a\|_{\bar{V}_0^{-1}}\}. \quad (5)$$

If $a = \mathrm{Best}_0(B)$ satisfies $\rho_0(a) \le \varepsilon_0$ and $C_0(a; B) \le 4\varepsilon_{L-1}$, the pool ranks actions in B satisfying the latter bound. The set contains a, and each candidate has regret at most $4\varepsilon_{L-1}$ plus the interval-error terms in Section 5. Selection can thus stop without accurate estimates for every survivor.

Lines 13–16: first-layer exploration. If the stopping test fails, propose the highest-scoring survivor with $\rho_0(a) > \varepsilon_0$. The direct bound C_0^{Δ} uses feature differences, all incoming actions, and the unclipped ridge estimate. If it exceeds $4\rho_0(a)$, choose the highest-scoring survivor satisfying $\rho_0(a) > \varepsilon_0$ and $C_0(a; B) \le 4\rho_0(a)$. A maximum-radius survivor always satisfies the latter bound, so this fallback exists. The selected observation is assigned to layer 0; its regret is controlled by its own radius plus the errors in Lemma 5. Section B.2 gives the exact branch conventions.

Lines 17–19: test uncertainty before using rewards. For a reached layer $s \ge 1$, a fresh Gaussian gives the common radius

$$\rho_s = \max_{b \in A} \|\bar{V}_s^{-1} x_b\|_2 + \sqrt{\pi/2} \max_{b \in A} |g_{\mathrm{test}}^\top \bar{V}_s^{-1/2} x_b| + \sqrt{2\pi\ell_s} \max_{b \in A} u_s(b). \quad (6)$$

The first term controls ridge bias; the remaining terms control stochastic prediction error. The test uses features and the fixed design, before accessing $\bar{b}_s$. Rather than estimate Gaussian width accurately in each round, Section 5 directly bounds the prediction errors beyond this randomized radius.

Lines 20–21: select an exploration action with an independent draw. When $\rho_s > \varepsilon_s$, draw an independent g_{sel} and retain actions with absolute projection at least $1/\sqrt{2}$ of the maximum. The information inequality in Section 5.1 applies to any choice from this set, with only a factor-two loss. The ranking rule uses the earlier-layer estimator

$$\bar{V}_{<s} = I + \sum_{j<s} (\bar{V}_j - I), \qquad \bar{b}_{<s} = \sum_{j<s} \bar{b}_j. \quad (7)$$

The estimator uses neither layer-s rewards nor the pool; unassigned observations enter neither sum. Conditioning on the test can change the law of g_{test}; an independent selection draw retains the required Gaussian law. The selected observation is assigned to s, ending this round's cascade.

Lines 22–25: elimination and final ranking. When exploration is not requested, use the stored $\bar{b}_s$ to eliminate by interval overlap and continue. After the last layer, the pool ranks the surviving actions. Their regret bounds hold uniformly before this ranking. The played feature enters the pool and, when $J_t \neq \perp$, its assigned layer; its reward is received only at a batch boundary. Terminal feedback deploys no new policy. Incremental maintenance costs $\mathcal{O}(KLd^2 + d^2)$ amortized arithmetic per round and $\mathcal{O}(Ld^2)$ estimator storage, excluding pending observations (Section B).

5 REGRET AND POLICY UPDATE ANALYSIS

We first analyze the known-horizon algorithm with $H = T$, then explain the horizon-free extension.

5.1 ADDITIVE INFORMATION CONTROL PERMITS LONGER BATCHES

Let $\bar{V}$ be a frozen design and $V \succeq \bar{V}$ the accumulated design. The squared leverage $\|x\|^2_{V^{-1}}$ determines the log-determinant increase when $xx^\top$ is added to V. Define the analysis-only Gaussian width

$$W(A, \bar{V}) = \mathbb{E}_g \max_{a \in A} |g^\top \bar{V}^{-1/2} x_a|, \qquad g \sim N(0, I_d).$$

Lemma 1 (Gaussian-max information inequality)**.** *For $V \succeq \bar{V} \succ 0$, $|A| = k$, and $a(g) \in \arg\max_{a \in A} |g^\top \bar{V}^{-1/2} x_a|$,*

$$W(A, \bar{V})^2 \le \left[\min\{d, 2\log(2k)\} + \frac{\det V}{\det \bar{V}} - 1 \right] \mathbb{E}_g \|x_{a(g)}\|^2_{V^{-1}}. \tag{8}$$

The improvement is *addition rather than multiplication*. A direct comparison of fixed and accumulated norms multiplies the action-complexity factor by the determinant ratio. Lemma 1 instead allows that ratio to grow to order $\min\{d, \log K\}$ while keeping the information factor of the same order. Updates can therefore wait for $\log(2 + \min\{d, \log K\})$ units of log-determinant growth rather than a constant amount. This separates infrequent policy updates from the statistical cost of exploration. The dependence is sharp up to constants (Corollary 2); layer-dependent thresholds will additionally remove the cost of counting every layer equally.

For the proof, whiten the features and write $M = \bar{V}^{-1/2} V \bar{V}^{-1/2}$. The selected index and its projection sign form at most $2k$ labels. If Σ is the second-moment matrix of the Gaussian conditional mean given this label, entropy and total covariance give $\operatorname{tr}\Sigma \le \min\{d, 2\log(2k)\}$ and $\Sigma \preceq I$. Metric Cauchy–Schwarz bounds W^2 by selected information times $\operatorname{tr}(M\Sigma)$; the inequality $\operatorname{tr}(M - I) \le \det M - 1$ completes the argument. Section E proves the result, its worst-case sharpness, and its factor-two extension to the approximate selector actually used by `GATE`.

5.2 INDEPENDENT GAUSSIAN TESTS AND DIRECT PREDICTION-ERROR CONTROL

Replacing only a later layer's assigned reward noises leaves its incoming action sets, assignments, designs, and own update times unchanged. Earlier-layer scores exclude those rewards, and other requests cannot refresh its snapshot. Conditional on this restricted record, the assigned noises retain their original independent laws (Lemma 7). Prediction concentration uses the restricted pre-test history, preserving the regression law; information accounting uses the full past and current test before the independent exploration draw, preserving the Gaussian law needed for information control.

Instead of requiring simultaneous correctness of all confidence intervals, we charge only the prediction error beyond the radius. When a later layer eliminates from its incoming set A, define

$$\xi_{s,t} = \max_{a \in A} (|\widehat{\mu}_s(a) - x_{t,a}^\top \theta_\star| - \rho_s(a))_+, \qquad (v)_+ = \max\{v, 0\}.$$

Set this quantity to zero if the layer does not eliminate; layer 0 instead always uses its full incoming set. An elimination step loses at most $2\xi_{s,t}$ in the best remaining mean. With N_s the number of assignments to s, the complete policy satisfies

$$R_T \le \sum_{t: J_t = 0} 4\beta_0 u_{0,t}(a_t) + \sum_{s=1}^{L-1} 4\varepsilon_{s-1} N_s + 4T\varepsilon_{L-1} + 4\sum_{t=1}^{T} \sum_{s=0}^{L-1} \xi_{s,t} + 2TD_H. \tag{9}$$

Here D_H is the layer-0 self-normalized error beyond its confidence threshold, with $\mathbb{E}D_H^2 \le 2(eH)^{-4}$. On a round assigned to layer $s \ge 1$, all earlier layers eliminated without exploring. At layer $s - 1$, surviving radii are at most ε_{s-1}; interval overlap therefore bounds each survivor's gap from the best remaining mean by $4\varepsilon_{s-1} + 2\xi_{s-1,t}$. Adding the loss $2\xi_{j,t}$ from each preceding elimination gives per-round regret at most $4\varepsilon_{s-1} + 4\sum_{j<s} \xi_{j,t}$. First-layer exploration uses its action's own radius, with at most $2D_H$ extra for the direct action-difference bound. Final choices, including

early stopping, cost at most $4\varepsilon_{L-1}$ plus error terms, uniformly over pooled ranking. Summing these disjoint cases and bounding their counts by N_s or T gives (9) (Section C).

To balance exploration and final-choice regret, let Φ be the minimum over positive thresholds of

$$\frac{a_0}{\varepsilon_0} + \sum_{s=1}^{L-1} a_s \frac{\varepsilon_{s-1}}{\varepsilon_s^2} + 4H\varepsilon_{L-1}. \tag{10}$$

The terms correspond to layer-0 exploration, later-layer exploration, and final choices, with deterministic $a_s > 0$ from Section G. The policy caps the unconstrained final threshold $\Phi/(8H)$ at $1/2$ and sets ℓ_s via (43) using per-round error budgets

$$\tau_s = \min\left\{1, \frac{\Phi}{H}\right\} \frac{2^{s-1}}{2^{L-1}-1} \quad (1 \le s < L).$$

Lemma 2 (Single-test prediction loss)**.** *For the prescribed parameters with $\Phi \le 4H$ and $1 \le s < L$,*

$$\mathbb{E}\xi_{s,t} = \mathcal{O}(\tau_s), \qquad \mathbb{E}\xi_{s,t}^2 = \mathcal{O}(\tau_s^2), \qquad H\sum_{s=1}^{L-1} \tau_s \le \Phi. \tag{11}$$

Thus total prediction excess is $\mathcal{O}(\Phi)$ in L_2, without an extra layer factor. Conditional on the restricted pre-test record, the maximum stochastic prediction error Z and Gaussian test maximum M in (6) are independent. For $c = \sqrt{\pi/2}$ and $u = \max_{a\in A} u_s(a)$, scalar convex-order comparison (Zhang, 2026) and Gaussian concentration give $\mathbb{E}e^{v(Z-cM)} \le e^{c^2v^2u^2}$ for $v \ge 0$: the width terms $+cvW$ and $-cvW$ cancel. Tail integration controls excess beyond the radius without estimating W. Restriction to elimination rounds uses a pre-test feature-norm condition, not conditioning on elimination (Section F).

5.3 Joint threshold choice avoids additional layer factors

Let ρ_s° replace the random maximum in Equation (6) by its expectation. If $\rho_s^\circ \ge \varepsilon_s/2$, Lemma 1 and the approximate selector give

$$\mathbb{E}_{\text{sel}}\|x_{a(g)}\|^2_{V_s^{-1}} \ge \frac{8d\log(1+H/d)}{a_s}\varepsilon_s^2. \tag{12}$$

The sum of selected squared leverages is at most $2d\log(1+H/d)$, so the expected number of such assignments is $\mathcal{O}(a_s/\varepsilon_s^2)$. The complementary case requires a Gaussian fluctuation of probability at most $e^{-2\ell_s}$ and is absorbed by the error budgets. Together with the first-layer and final-choice costs, these bounds give the three terms in (10). Decreasing ε_s increases exploration at layer s but lowers the next layer's nominal regret. Weighted AM–GM with $\lambda_0 = 2^{-L}$, $\lambda_s = 2^{s-1-L}$, and terminal weight $1/2$ cancels every threshold exponent, giving

$$\Phi = 2^{3-2^{1-L}}\sqrt{H}\prod_{s<L} a_s^{\lambda_s}, \qquad \Phi^2 = \mathcal{O}(dH\log(1+H/d)\min\{d, \log K\}).$$

The first-layer confidence coefficient contributes to Φ^2 only through $(\beta_0^2)^{2^{1-L}}$. The prescribed depth makes this factor bounded by a universal constant, avoiding an additional confidence-related logarithmic multiplier. Lemmas 9 and 10 give the explicit backward recursion and parameter verification. If $\Phi > 4H$, the final threshold is $1/2$ and the pathwise bound $R_H \le 2H < \Phi/2$ suffices. For L_2 regret, a bound on accumulated squared leverage also controls the sum of its conditional expectations: $0 \le Y_t \le 1$ and $\sum_t Y_t \le B$ imply $\|\sum_t \mathbb{E}[Y_t \mid \mathcal{G}_t]\|_{L_2} \le B + \sqrt{B}$. Thus information is required in expectation, not from every realized action (Lemma 11).

Each layer update consumes z_s new units of log determinant; a pool-only update consumes z_0. Concurrent requests count once. Hence

$$\begin{aligned} S_T &\le d\log(1+T/d)\left(\sum_{s<L} z_s^{-1} + z_0^{-1}\right) \\ &= \mathcal{O}\left(\frac{d\log(1+T/d)\log(2+\log(2+\min\{d,\log K\}))}{\log(2+\min\{d,\log K\})}\right). \end{aligned} \tag{13}$$

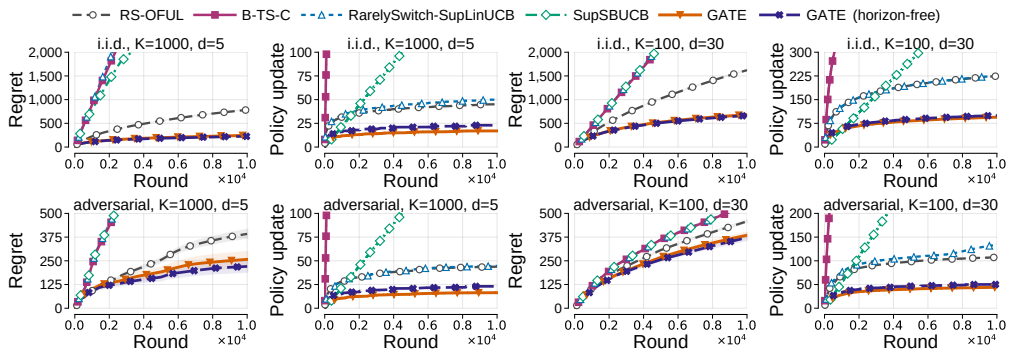

Figure 1: Synthetic results at $T = 10^4$: means and pointwise ± 1.96 SE bands. Rows: i.i.d. and oblivious rotating ("adversarial") contexts. Column pairs: large-action $(K, d) = (1000, 5)$ and small-action $(100, 30)$. Both GATE policies make no later-layer assignments. GATE denotes known horizon. Zoomed axes; full ranges and statistics: Figure 2 and Table 5.

The later-layer thresholds are $z_s = \log(2 + \min\{d, \log K\}) + 2^{(L-s-5)/2}$. Smaller exponents in the optimized product permit geometrically larger thresholds at earlier layers. Only $\mathcal{O}(\log(2 + \log(2 + \min\{d, \log K\})))$ layers have the geometric term below the common logarithmic term; the remaining reciprocals decay geometrically. This yields the iterated-logarithmic factor in $\min\{d, \log K\}$ rather than the full depth (Equation (52)).

5.4 Horizon-free learning and lower bounds

Horizon-free guarantees with retained data. Start at $H_0 = \max\{2, \lceil \log(2 + \min\{d, \log K\}) \rceil\}$. Before round $t = H$, reveal pending rewards, double H, recompute parameters, and refresh all layer and pool snapshots. Retain existing statistics and labels; initialize only new layers. Here H bounds the full history, not the epoch length. Epoch regret bounds sum geometrically in L_2. For updates, group consecutive epochs with common depth p, whose update thresholds z_s remain fixed. Charge each design's log-determinant growth across the group, not a fresh budget per epoch. For a completed group let E_p be the next group's first global round; for the group containing T, set $E_p = T$. The depth schedule ensures

$$\sum_{\text{depth groups}} \log(1 + E_p/d) \le 4 \log(1 + T/d). \tag{14}$$

Combining this with (13) bounds ordinary updates; public doubling updates fit the same order. Both guarantees follow without restarting or an extra horizon-dependent factor (Section H).

Sequential optimality and feedback necessity. We reconstruct Li et al. (2024)'s hierarchy with joint norm and scale calibration and scalar-history testing (Section I). For feedback, stage-length calibration and joint unit-norm scaling refine Ruan et al. (2021)'s multiscale construction. At expected query budget S, retaining the $\sqrt{1 + S/d}$ prefactor (Proposition 3) avoids an extra $\log \log T$ in the denominator when inverting at the exact regret target. Reward suppression handles history-dependent actions between adaptive queries; a coordinate bound completes Theorem 2's range (Section J).

6 Numerical Experiments

We compare both GATE versions with the baselines in Table 1 on cumulative pseudo-regret and nonterminal policy updates, using native parameter rules and feedback schedules. Paired runs share contexts and scalar $N(0, 1)$ noise but receive only their own selected-action rewards (Section L.1). Eight synthetic settings cover i.i.d. unit-ball and oblivious rotating contexts in both action regimes, with 50 paired replications per setting at $T = 10^4$ (Figure 1) and 20 at $T = 2 \times 10^7$ (Section L.4). Jester and KuaiRec provide two semi-synthetic controls with data-derived contexts and simulated linear rewards (50 paired replications each; Section M). Reward parameters are fixed within each replication; all replications are included. GATE has lower sample-mean terminal regret and fewer mean updates than every baseline across all ten settings; horizon-free GATE shows similar empirical regret (Sections N and N.2). Branch activity and runtime results are reported in Sections L.7 and N.3.

AI USE STATEMENT

We used AI tools solely to polish author-written text. The authors reviewed all edits and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

Section A gives a reading guide and the external mathematical inputs; Sections B, G and H specify the policies, parameters, and endpoint conventions. Complete proofs are in Sections C to J. Sections L.1, L.5 and L.6 specify the experimental protocol, implementation conventions, and statistical reporting. Section N.2 reports paired terminal comparisons. Section M describes the data sources, preprocessing, user splits, and representation learning for the semi-synthetic experiments. Section N.3 documents the software and hardware environment and the timing protocol. The supplementary material provides the source code for the experiments.

ETHICS STATEMENT

This work studies an abstract sequential decision-making model and semi-synthetic controls using public recommendation covariates. There are no interventions with participants or deployment claims. Recorded ratings and watch ratios serve representation construction; online bandit rewards are simulated. The original recommendation archives and fitted user-level arrays are not redistributed here, and reuse remains subject to the source terms. Regret guarantees under the stated assumptions do not certify fairness, privacy, safety, or effectiveness in a different deployment.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 127–135. PMLR, 2013.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214. PMLR, 2011.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 2020.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. arXiv preprint arXiv:2107.05745v1, 2021. Expanded version of the paper published in NeurIPS 2020.

Chongming Gao. KuaiRec dataset with raw features. Zenodo, 2026. Version v1.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 540–550. Association for Computing Machinery, 2022.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

Yanjun Han, Zhengqing Zhou, Zihao Hu, Jose Blanchet, Peter W. Glynn, Yinyu Ye, and Zhengyuan Zhou. Sequential batch learning in finite-action linear contextual bandits. arXiv preprint arXiv:2004.06321v2, 2026. Originally posted in 2020; to appear in Operations Research.

Osama Hanna, Lin Yang, and Christina Fragouli. Efficient batched algorithm for contextual linear bandits with large action space via soft elimination. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56772–56783, 2023.

Tianyuan Jin. Nearly minimax-optimal regret for linear contextual bandits with arbitrary adaptive action sets. arXiv preprint arXiv:2609.15170, 2026.

Amin Karbasi, Vahab Mirrokni, and Mohammad Shadravan. Parallelizing Thompson sampling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10535–10548, 2021.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670. Association for Computing Machinery, 2010.

Yingkai Li, Yining Wang, Xi Chen, and Yuan Zhou. Tight regret bounds for infinite-armed linear contextual bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 370–378. PMLR, 2021.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. *IEEE Transactions on Information Theory*, 70(1):372–388, 2024.

Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 74–87. Association for Computing Machinery, 2021. doi: 10.1145/3406325.3451004.

Andrew Wagenmaker, Julian Katz-Samuels, and Kevin Jamieson. Experimental design for regret minimization in linear bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3088–3096. PMLR, 2021.

Sanghoon Yu and Min-hwan Oh. Optimal and practical batched linear bandit algorithm. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73262–73285. PMLR, 2025.

Sanghoon Yu and Min-hwan Oh. Practical and optimal algorithm for linear contextual bandits with rare parameter updates. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. To appear.

Yihan Zhang. Sharp one-dimensional sub-Gaussian comparison in convex order. arXiv preprint arXiv:2604.26819, 2026.

Zihan Zhang, Xiangyang Ji, and Yuan Zhou. Almost optimal batch-regret tradeoff for batch linear contextual bandits. In *International Conference on Learning Representations*, 2025.

A Reading Guide, Notation, and Basic Tools

The appendices give the algorithm and proofs in the order needed to check the results. Section B explains the state and exact decisions for general positive thresholds. Section C first decomposes regret into quantities that remain to be bounded. Section D establishes the conditional noise law; Sections E and F then prove the Gaussian information and prediction-error inequalities. Section G chooses the numerical parameters and completes the fixed-horizon bounds. Section H specifies and analyzes the horizon-free algorithm. The sequential and feedback lower bounds are separate in Sections I and J. Sequential benchmarks, and all empirical results follow in Sections K to N. Forward references to the numerical parameter choice identify values to be specified later; no probability argument assumes the result it is proving.

A.1 Notation and three different notions of change

For a positive integer n, $[n] = \{1, \dots, n\}$. The notation $(v)_+ = \max\{v, 0\}$ denotes the positive part, and $\log_+ v = \max\{\log v, 0\}$ for $v > 0$. $\mathbf{1}\{E\}$ equals one on event E and zero otherwise. For a symmetric positive definite matrix V, $\|x\|_V = \sqrt{x^\top V x}$; $V \succeq W$ means $x^\top (V - W)x \geq 0$ for every vector x. An inverse square root is the symmetric one unless the Cholesky implementation is explicitly used. A feature family is indexed: two equal vectors at different indices remain two actions, and all action ties use the smallest input index.

The statistical parameter $\theta_\star$ is fixed; changing feature geometry does not mean a changing reward parameter. An *action change* concerns which index is played. A *policy update* changes the kernel used to map the current action set to an action distribution. A *feedback query* reveals previously unavailable selected-action rewards. Only for `GATE`'s protocol do the last two counts coincide exactly. The lower-bound class allows more context-history-dependent actions than `GATE`, but it cannot inspect rewards before a query.

A *snapshot* is a fixed copy of sufficient statistics. A layer can have new features in its accumulated design while its snapshot remains unchanged. The monitor can use the accumulated design to stop a batch; the action rule uses only snapshots. Thus adaptive endpoints and fixed within-batch policies are compatible.

A.2 Ridge regression, bias, and prediction error

Let the rows of X be selected feature vectors, let $r = X\theta_\star + \eta$, and put $V = I + X^\top X$. Ridge regression minimizes $\|r - X\theta\|_2^2 + \|\theta\|_2^2$; setting its gradient to zero gives $V\widehat{\theta} = X^\top r$, hence $\widehat{\theta} = V^{-1}X^\top r$. Substituting the reward model gives

$$\widehat{\theta} - \theta_\star = -V^{-1}\theta_\star + V^{-1}X^\top \eta. \tag{15}$$

Indeed, $V^{-1}X^\top X - I = V^{-1}(V - I) - I = -V^{-1}$. The first term is regularization bias (fixed when the design is fixed); the second is the stochastic error. For a candidate feature x, Cauchy–Schwarz and $\|\theta_\star\|_2 \leq 1$ give

$$|x^\top V^{-1}\theta_\star| \leq \|V^{-1}x\|_2 \|\theta_\star\|_2 \leq \|V^{-1}x\|_2.$$

Since $V \succeq I$, each eigenvalue $\lambda \geq 1$ satisfies $\lambda^{-2} \leq \lambda^{-1}$; in an eigenbasis of V this gives $V^{-2} \preceq V^{-1}$ and $\|V^{-1}x\|_2^2 = x^\top V^{-2}x \leq x^\top V^{-1}x = \|x\|_{V^{-1}}^2$. Clipping a prediction to $[-1, 1]$ cannot increase its distance from the true mean $x^\top\theta_\star \in [-1, 1]$: an outside prediction moves to the nearer endpoint, and an inside prediction is unchanged. These identities hold for any realized design. A concentration bound that treats X as fixed additionally needs a valid conditional law for η; Section D supplies that law for later layers.

A.3 Determinants and accumulated information

For $V \succ 0$, the matrix determinant lemma states

$$\det(V + xx^\top) = \det(V)(1 + x^\top V^{-1}x).$$

One proof writes $V + xx^\top = V^{1/2}(I + vv^\top)V^{1/2}$ with $v = V^{-1/2}x$; $I + vv^\top$ has eigenvalues $1 + \|v\|_2^2$ and 1 in all orthogonal directions. Thus the squared leverage $x^\top V^{-1}x$ determines the

increase in log determinant. For features of norm at most one and $V \succeq I$, this leverage lies in $[0,1]$ because $V^{-1} \preceq I$. The function $2\log(1+u) - u$ is zero at $u = 0$ and has derivative $(1-u)/(1+u) \geq 0$ on $[0,1]$, proving $u \leq 2\log(1+u)$ there. For $V_t = I + \sum_{i<t} x_i x_i^\top$ and $u_t = \|x_t\|^2_{V_t^{-1}}$, the determinant identity gives $\log(1+u_t) = \log\det V_{t+1} - \log\det V_t$. Summing $u_t \leq 2\log(1+u_t)$ and using $V_1 = I$ yields

$$\sum_{t=1}^{n} \|x_t\|^2_{V_t^{-1}} \leq 2\log\det V_{n+1} \leq 2d\log(1+n/d). \tag{16}$$

The second inequality is arithmetic–geometric mean applied to the positive eigenvalues: $\det V_{n+1} \leq (\operatorname{tr} V_{n+1}/d)^d$ and $\operatorname{tr} V_{n+1} = d + \sum_{t=1}^{n} \|x_t\|_2^2 \leq d+n$. For an interval $p \leq t \leq q$, the same argument gives $\sum_{t=p}^{q} u_t \leq 2(\log\det V_{q+1} - \log\det V_p)$: the logarithmic increments telescope, even with previously accumulated data.

If $V \succeq \bar{V} \succ 0$, put $M = \bar{V}^{-1/2} V \bar{V}^{-1/2} \succeq I$. Its eigenvalues $\lambda_i \geq 1$ satisfy $\lambda_i \leq \prod_j \lambda_j = \det V / \det \bar{V}$. Thus $\det V / \det \bar{V} < e^z$ gives $M \prec e^z I$ and $M^{-1} \succ e^{-z} I$. Using $V = \bar{V}^{1/2} M \bar{V}^{1/2}$ and $V^{-1} = \bar{V}^{-1/2} M^{-1} \bar{V}^{-1/2}$ yields $V \prec e^z \bar{V}$ and $\bar{V}^{-1} \preceq e^z V^{-1}$; these congruences preserve matrix order. These are the precise matrix comparisons used for a stale snapshot.

A.4 Conditional moments and summing dependent losses

Independent centered 1-sub-Gaussian noises have second moments at most one: $\cosh y = (e^y + e^{-y})/2 \geq 1 + y^2/2$ gives, for $v \neq 0$, $\mathbb{E}\eta^2 \leq 2(\mathbb{E}\cosh(v\eta) - 1)/v^2 \leq 2(e^{v^2/2} - 1)/v^2 \to 1$ as $v \to 0$. At a round's pre-reward history, include the current action randomization; the selected feature is then determined, whereas the current noise remains independent. Writing $\mathcal{H}_t$ for this information, we have $\mathbb{E}[\eta_t \mid \mathcal{H}_t] = 0$ and $\mathbb{E}[e^{v\eta_t} \mid \mathcal{H}_t] \leq e^{v^2/2}$. This allows predictable-design self-normalization for layer 0 (Section C.2). Later-layer fixed-design arguments require a different, restricted history, specified in Section D.

For square-integrable random variables, Minkowski's inequality gives $\|\sum_i Z_i\|_{L_2} \leq \sum_i \|Z_i\|_{L_2}$, without independence: after expanding the squared sum, Cauchy–Schwarz bounds each cross term $\mathbb{E}[Z_i Z_j]$ by $\|Z_i\|_{L_2}\|Z_j\|_{L_2}$. For any nonnegative random variable Z, Tonelli's theorem gives $\mathbb{E}Z = \int_0^\infty \Pr(Z > x)\,dx$ and $\mathbb{E}Z^2 = \int_0^\infty 2x \Pr(Z > x)\,dx$. Indeed, $Z^p = \int_0^\infty p x^{p-1} \mathbf{1}\{Z > x\}\,dx$ for $p = 1, 2$; nonnegativity permits exchanging expectation and integration. The proof uses these identities for error *beyond* a radius, not for the full error on an event where the radius fails (Equations (30) and (31)).

A.5 External Gaussian comparison and concentration

Lemma 3 (Scalar comparison; Zhang (2026), Theorem 1)**.** *If X is centered and $\mathbb{E}e^{vX} \leq e^{v^2/2}$ for every real v, and $G \sim N(0,1)$, then $X \preceq_{\mathrm{cx}} cG$ with $c = \sqrt{\pi/2}$. This means that every convex function with well-defined expectations has no larger expectation under X than under cG.*

This scalar theorem is an external input. For a convex function of independent noises, fix all but one coordinate; independence preserves that coordinate's law, so scalar comparison applies. Integrating and repeating replaces the coordinates by independent cG_i (Lemma 8). This does not assert the same comparison for an arbitrary dependent sub-Gaussian vector.

The other external input is Gaussian concentration: if f is 1-Lipschitz, meaning $|f(g) - f(h)| \leq \|g - h\|_2$, and $g \sim N(0, I_d)$, then

$$\mathbb{E}\exp(v[f(g) - \mathbb{E}f(g)]) \leq e^{v^2/2} \qquad (v \in \mathbb{R}). \tag{17}$$

The Gaussian logarithmic Sobolev inequality implies this form (Boucheron et al., 2013). For Lipschitz constant $a > 0$, apply (17) to f/a at argument av to replace $v^2/2$ with $a^2v^2/2$; for $a = 0$, f is constant. In particular, $f(g) = \max_{b\in A} |g^\top V^{-1/2} x_b|$ satisfies $|f(g) - f(h)| \leq \max_{b\in A} |(g-h)^\top V^{-1/2} x_b| \leq \max_{b\in A} \|x_b\|_{V^{-1}} \|g - h\|_2$, by the triangle and Cauchy–Schwarz inequalities. Negative values of v are allowed and will be essential when the width terms of the prediction error and the independent test cancel.

A.6 Binary testing and scalar-reward relative entropy

For probability laws P, Q, total variation is $\mathrm{TV}(P,Q) = \sup_E |P(E) - Q(E)|$ and relative entropy is $\mathrm{KL}(P\|Q) = \mathbb{E}_P \log(dP/dQ)$ when $P \ll Q$. Pinsker's inequality gives $\mathrm{TV}(P,Q) \le \sqrt{\mathrm{KL}(P\|Q)/2}$ (Boucheron et al., 2013). If the first binary option is optimal under P and the second under Q, let E be the event of choosing the first. Its error sum satisfies $P(E^c) + Q(E) = 1 - [P(E) - Q(E)] \ge 1 - \mathrm{TV}(P,Q)$. Dividing by two gives the equal-weight average-error lower bound $(1 - \mathrm{TV}(P,Q))/2$. Randomized decisions are covered by including an independent random seed with the same law under both hypotheses; data processing cannot increase total variation or relative entropy.

For $Y \sim N(\mu, 1)$, the log-density ratio against $N(\mu', 1)$ is $(\mu - \mu')(Y - (\mu + \mu')/2)$; taking expectation gives $\mathrm{KL} = (\mu - \mu')^2/2$. Under independent $N(0,1)$ reward noises, write $P_\theta^{(n)}$ for the joint law of the learner's seeds, actions, and first n generated rewards, including the exposed contexts. When paired instances expose the same contexts through this prefix, the chain rule gives

$$\mathrm{KL}(P_\theta^{(n)} \| P_{\theta'}^{(n)}) = \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_\theta \left[(x_{i,a_i}^\top (\theta - \theta'))^2 \right].$$

At the same realized history, the learner's action kernels agree and contribute zero KL; their marginal action laws need not agree. With delayed feedback, this is an enlarged analysis history, not extra learner access; data processing bounds the observed-history KL by this value. A composite action still generates only *one scalar reward*; a lower bound must compute its scalar mean shift, not grant separate noisy observations for its coordinates.

B Algorithm State and Exact Decisions

B.1 What a round reads and what it changes

Start with a fixed depth $L \ge 2$, a layer-0 coefficient β_0, positive exploration thresholds $(\varepsilon_s)_{s<L}$, positive update thresholds $(z_s)_{s<L}$, and test coefficients $(\ell_s)_{1 \le s < L}$. Section G derives their deterministic numerical values from (H, d, K). For the known-horizon policy $H = T$ is fixed throughout a run. The horizon-free schedule changes H only at deterministic public times.

There are two distinct operations. Before the context arrives, the monitor checks accumulated designs, may request feedback, and may replace snapshots. After the context arrives, the action rule uses only those snapshots and fresh private randomness. It neither uses the current accumulated designs nor changes the parameters used for the current batch.

Object	Meaning and allowed changes	Use in the policy
V_s	Sum of assigned feature outer products; changes after assignments	Batch-end monitor only
b_s	Sum of received assigned feature–reward products	Copied only at authorized updates
$(\bar{V}_s, \bar{b}_s)$	Fixed snapshot of a training prefix	Prediction, tests, and action selection
$(\bar{V}_{<s}, \bar{b}_{<s})$	Union of earlier-layer snapshots	Rank exploration candidates at layer s
$(\bar{V}_{\mathrm{all}}, \bar{b}_{\mathrm{all}})$	Fixed pooled snapshot	Rank already certified terminal candidates

At layer 0, overlapping confidence intervals define survivors. If any survivor is wide, the algorithm first checks the terminal certificate, then chooses a wide proposal or a certified fallback. Otherwise the survivors pass to layer 1. At each later layer, a response-independent Gaussian test decides whether to explore. Exploration ends the round and assigns the observation to that layer. If the test permits elimination, the algorithm uses that layer's fixed reward statistics and proceeds with its survivors. The first exploration branch ends the path; an observation therefore has one layer label or the terminal label $\perp$, never multiple layer labels. Algorithm 1 is the complete known-horizon pseudocode.

B.2 State, exact branches, and randomness

At the beginning of round t,

$$V_{s,t} = I + \sum_{i<t:J_i=s} x_{i,a_i} x_{i,a_i}^\top, \qquad b_{s,t} = \sum_{\substack{i<t:J_i=s \\ r_i \text{ observed}}} x_{i,a_i} r_i.$$

The pooled sums omit the constraint $J_i = s$: only selected observations, not counterfactual rewards, enter the pool. Store the selected feature, assignment label, and round index for each pending observation; the round index identifies the reward returned at the next feedback query. Immediately before context t, form

$$\begin{aligned} \mathcal{R}_t &= \{s : \log\det V_{s,t} - \log\det \bar{V}_{s,t} \geq z_s\}, \\ P_t &= \mathbf{1}\{\log\det V_{\mathrm{all},t} - \log\det \bar{V}_{\mathrm{all},t} \geq z_0\}. \end{aligned} \tag{18}$$

If $\mathcal{R}_t \neq \varnothing$ or $P_t = 1$, receive the pending rewards and add each response exactly once to its assigned layer, if any, and to the pool. Copy (V_s, b_s) only for $s \in \mathcal{R}_t$ and always copy the pooled state. Clear the buffer and count one update. If only the pool requested, copy no layer. Concurrent requests still count once. The pool predicate is checked on all accumulated selected features, not gated by whether a terminal observation occurred since its previous copy. A terminal feedback after an external stop opens no further policy and is not counted. When $H = T$ there is no parameter recomputation within the run.

Equivalent layer-0 decision rule. Let B be the interval-overlap survivor set. The condition $\max_{a\in B} \rho_0(a) > \varepsilon_0$ is equivalent to nonemptiness of $\{a \in B : \rho_0(a) > \varepsilon_0,\ C_0(a; B) \leq 4\rho_0(a)\}$: by Lemma 4, a maximum-radius survivor belongs to that set whenever any survivor is wide. If the condition fails, all survivors pass to layer 1. Otherwise the terminal test is evaluated first. Its successful branch ranks $B_{\mathrm{cert}} = \{b \in B : C_0(b; B) \leq 4\varepsilon_{L-1}\}$ using the pool, and leaves $J_t = \perp$. If the terminal test fails, select the highest-scoring wide survivor, using the highest-scoring certified wide survivor as fallback when the action-difference certificate fails. The latter certificate uses the full incoming indices $[K_t]$ and the unclipped ridge vector.

Those candidate calculations have no side effects and consume no randomness, so postponing them until the terminal test fails leaves the selected action and assignment unchanged. Each nonterminal branch selects the same proposal or fallback under the same certificate. Induction over rounds then preserves designs, response sums, snapshots, and batch endpoints.

There is no unconditional singleton short-circuit. All-zero incoming features give zero norms and zero Gaussian projections, hence no later-layer assignment. Repeated and opposite features remain distinct indexed actions. A terminal observation writes only to the pool.

For a concrete Gaussian implementation, use the lower triangular Cholesky factor with positive diagonal, $\bar{V} = CC^\top$, and $v_a = C^{-1}x_a$. Evaluate each projection as $g^\top v_a$. This is the same joint Gaussian law as the symmetric-square-root notation, with a fully specified numerical coordinate convention. Test and selection probes are independent; every reached later-layer test receives a fresh vector, and a selection vector is used only on its assignment branch.

B.3 Computation and the meaning of exact specification

The parameter map takes $\mathcal{O}(L)$ scalar operations in log arithmetic; products are accumulated as sums and two-term exponential sums use log-sum-exp. A reached layer computes its feature leverages and bias norms through triangular solves in $\mathcal{O}(Kd^2)$. A Gaussian projection can be evaluated with one triangular solve and K dot products, or by reusing whitened features. At most one additional selection probe is needed, because an assignment terminates the cascade. The action-difference certificate is computed for one proposal against the K incoming actions, not for all K^2 pairs, and costs $\mathcal{O}(Kd^2)$.

Maintain Cholesky factors for the earlier-layer prefix sums in (7), including the union of all frozen layers for initializing future prefixes. Whenever layer j is authorized to copy, its newly incorporated observations enter each prefix $s > j$ by rank-one Cholesky updates and response-sum increments. Each observation is assigned to one layer and enters each later prefix at most once, giving $\mathcal{O}(TLd^2)$

total work and $\mathcal{O}(Ld^2)$ estimator state. A newly born empty layer uses a copy of the previously maintained full-layer prefix; terminal observations are not inserted into these prefixes. A pool-only copy changes none of them. Their sums contain each authorized assigned observation once.

Accumulated, frozen, and pooled factors also admit rank-one updates and copying; repeated matrix inversion is unnecessary. Applying a response sum through a frozen factor costs $\mathcal{O}(d^2)$. Reward aggregation costs $\mathcal{O}(Td)$, and a public full copy costs $\mathcal{O}(Ld^2)$. These costs fit $\mathcal{O}(KLd^2 + d^2)$ amortized arithmetic per round. The reward buffer is not included in estimator storage and may contain up to T feature records.

C Elimination and Regret Decomposition

We first identify the loss to be controlled, before choosing numerical parameters or invoking fixed-design concentration. Inaccurate intervals are allowed: only their error beyond the stated radius is charged.

C.1 Elimination with imperfect confidence intervals

For a nonempty indexed set A, realizable means $\mu_a \in [-1,1]$, clipped estimates $\widehat{\mu}_a \in [-1,1]$, and radii $\rho(a) \geq 0$, let $L_a = \max\{-1, \widehat{\mu}_a - \rho(a)\}$, $U_a = \min\{1, \widehat{\mu}_a + \rho(a)\}$, and $\xi(A) = \max_{a\in A}(|\widehat{\mu}_a - \mu_a| - \rho(a))_+$. By definition, $|\widehat{\mu}_a - \mu_a| \leq \rho(a) + \xi(A)$. Together with $\mu_a \in [-1,1]$, this gives $L_a - \xi(A) \leq \mu_a \leq U_a + \xi(A)$ even after clipping the endpoints.

Lemma 4 (Elimination loss)**.** *Let $B = \{a \in A : U_a \geq \max_{b\in A} L_b\}$ and $c(a;B) = \max_{b\in B} U_b - L_a$. Then B is nonempty, $0 \leq c(a;B) \leq 2$, and, for $a, b \in B$, $U_b - L_a \leq 2\rho(b) + 2\rho(a)$. Moreover*

$$\max_{a\in A} \mu_a - \max_{b\in B} \mu_b \leq 2\xi(A), \tag{19}$$

$$\max_{b\in B} \mu_b - \mu_a \leq c(a;B) + 2\xi(A) \quad (a \in B),$$

$$\max_{b\in A} \mu_b - \mu_a \leq c(a;B) + 4\xi(A) \quad (a \in B). \tag{20}$$

In particular a maximum-radius survivor has certificate at most four times its radius; if all surviving radii are at most ε, every survivor has certificate at most 4ε.

Proof. Write $\xi = \xi(A)$ and choose $j \in \arg\max_{a\in A} L_a$. Since $U_j \geq L_j = \max_{a\in A} L_a$, we have $j \in B$. Also $\max_{b\in B} U_b \geq U_j \geq L_a$ for every $a \in A$; as all endpoints lie in $[-1,1]$, this proves $0 \leq c(a;B) \leq 2$. For $a, b \in B$, $L_b \leq U_a$ and $0 \leq U_b - L_b \leq 2\rho(b)$, so $U_b - L_a = (U_b - L_b) + (L_b - U_a) + (U_a - L_a) \leq 2\rho(b) + 2\rho(a)$. Choose $a^\star \in \arg\max_{a\in A} \mu_a$. If $a^\star \in B$, the loss in (19) is zero. Otherwise $U_{a^\star} < L_j$ and $\mu_{a^\star} \leq U_{a^\star} + \xi < L_j + \xi \leq \mu_j + 2\xi$, which proves (19) since $j \in B$. For any $a \in B$ and a best index b in B, $\mu_b - \mu_a \leq (U_b + \xi) - (L_a - \xi) \leq c(a;B) + 2\xi$. Adding (19) gives (20). Finally, maximizing the pairwise endpoint bound over $b \in B$ gives $c(a;B) \leq 2\rho(a) + 2\max_{b\in B} \rho(b)$, which proves both concluding radius assertions. □

If the layer-0 wide-survivor guard is false, every survivor has radius at most ε_0. Equivalently, the certified-wide set is empty; a maximum-radius survivor witnesses the converse by Lemma 4. Thus all layer-0 survivors have radius at most ε_0. A later-layer no-assignment branch directly has common radius at most ε_s. Consequently an action assigned to layer $s \geq 1$ inherits nominal charge $4\varepsilon_{s-1}$, irrespective of which incoming action the Gaussian-max selector chooses. A final survivor has nominal charge $4\varepsilon_{L-1}$, uniformly over the pooled ranking.

C.2 Initial-layer error in second moment

For this calculation use $\beta_0 = 1 + \sqrt{d\log(1 + H/d) + 8\log(eH)}$, the deterministic layer-0 coefficient specified in Section G.2. The exploration thresholds remain arbitrary positive numbers in the regret decomposition. Here H bounds the full global history, not just the current epoch's length; the processes V_n, S_n below are not restarted at epoch boundaries.

For analysis only, place all independent learner seeds in $\mathcal{F}_0$ and set $\mathcal{F}_n = \sigma(\mathcal{F}_0, \eta_1, \ldots, \eta_n)$, including generated but undelivered noises; this does not change the policy's access. Set $x_t =$

$\mathbf{1}\{J_t = 0\}x_{t,a_t}$, $V_n = I + \sum_{t\le n} x_t x_t^\top$, and $S_n = \sum_{t\le n} x_t \eta_t$. The feature x_t is determined by $\mathcal{F}_{t-1}$, whereas η_t remains independent. Conditional sub-Gaussianity therefore gives, for fixed w,

$$\mathbb{E}\left[e^{(w^\top x_t)\eta_t - \frac{1}{2}(w^\top x_t)^2} \,\middle|\, \mathcal{F}_{t-1}\right] \le 1.$$

Thus $\exp(w^\top S_n - \frac{1}{2} w^\top (V_n - I)w)$ has conditional next-step expectation at most its current value. Averaging against $w \sim N(0, I)$ preserves this property. The Gaussian density changes the exponent to $w^\top S_n - \frac{1}{2} w^\top V_n w$; completing the square around $V_n^{-1} S_n$ gives

$$M_n = (\det V_n)^{-1/2} \exp\left(\frac{1}{2}\|S_n\|^2_{V_n^{-1}}\right), \qquad \mathbb{E}M_0 = 1, \tag{21}$$

a nonnegative supermartingale. This is the self-normalization mechanism of Abbasi-Yadkori et al. (2011). Set $B_H = \sqrt{d\log(1 + H/d) + 8\log(eH)}$ and $D_H = (\max_{n\le H} \|S_n\|_{V_n^{-1}} - B_H)_+$. For $x \ge 0$, let $\sigma_x \le H$ be the first time $\|S_n\|_{V_n^{-1}} > B_H + x$, or H if there is no crossing. On $\{D_H > x\}$, the determinant bound $\log\det V_n \le d\log(1 + H/d)$ implies $M_{\sigma_x} \ge e^{((B_H+x)^2 - d\log(1+H/d))/2}$. Bounded optional stopping and nonnegativity give

$$1 \ge \mathbb{E}M_{\sigma_x} \ge e^{((B_H+x)^2 - d\log(1+H/d))/2}\mathbb{P}(D_H > x).$$

Rearranging, and substituting the definition of B_H, yields

$$\mathbb{P}(D_H > x) \le (eH)^{-4} e^{-B_H x - x^2/2} \le (eH)^{-4} e^{-x^2/2}.$$

Using $\mathbb{E}D_H^2 = \int_0^\infty 2x\mathbb{P}(D_H > x)\,dx$ and $\int_0^\infty 2xe^{-x^2/2}\,dx = 2$ gives

$$\mathbb{E}D_H^2 \le 2(eH)^{-4}. \tag{22}$$

Let $\nu_t < t$ be the global cutoff of the frozen root snapshot, with $\nu_t = 0$ initially: it contains exactly the layer-0 assignments through round ν_t. Thus the ridge identity in Section A.2 gives

$$\bar{V}_{0,t} = V_{\nu_t}, \qquad \widehat{\theta}_{0,t} - \theta_\star = \bar{V}_{0,t}^{-1}(S_{\nu_t} - \theta_\star).$$

Since D_H controls every prefix, it also controls this random cutoff, without conditioning on ν_t. Cauchy–Schwarz, clipping, and $\|\theta_\star\|_{\bar{V}_{0,t}^{-1}} \le 1$ give, with $u = u_{0,t}(a) \le 1$ and $\beta_0 = 1 + B_H$,

$$\begin{aligned}(|\widehat{\mu}_{0,t}(a) - x_{t,a}^\top \theta_\star| - \beta_0 u)_+ &\le u(1 + \|S_{\nu_t}\|_{\bar{V}_{0,t}^{-1}} - \beta_0)_+ \\ &\le uD_H \le D_H.\end{aligned}$$

Taking the maximum over $a \in [K_t]$ gives $\xi_{0,t} \le D_H$. There are at most H rounds in the epoch, so $\sum_{t \text{ in the epoch}} \xi_{0,t} \le HD_H$ pathwise, and hence

$$\left\| \sum_{t \text{ in the epoch}} \xi_{0,t} \right\|_{L_2} \le \sqrt{2}\, H(eH)^{-2} \le \frac{\sqrt{2}}{H}. \tag{23}$$

Also $\|2HD_H\|_{L_2} \le 2\sqrt{2}/H$, by (22). Together these integrated bounds supply both initial-layer excess terms in (24).

Lemma 5 (Layer-0 certificates)**.** *Let D_H denote the root self-normalized excess defined in Section C.2, for the full causally generated prefix of global length at most H. A root action accepted by $C_0^\Delta(a; A) \le 4\rho_0(a)$ has regret at most $4\beta_0 u_0(a) + 2D_H$. A root action accepted by its soft certificate has regret at most $4\beta_0 u_0(a) + 4\xi_{0,t}$. A terminal action from B_{cert} in line 12 has regret at most $4\varepsilon_{L-1} + 4\xi_{0,t}$.*

Proof. For the frozen state at round t, write $S := S_{\nu_t} = \sum_{i\le\nu_t : J_i = 0} x_{i,a_i}\eta_i$ and $\bar{V}_0 = V_{\nu_t}$: S is the feature-weighted noise sum over the whole frozen prefix, used only in the analysis. The ridge identity gives $\theta_\star - \widehat{\theta}_0 = \bar{V}_0^{-1}(\theta_\star - S)$. Cauchy–Schwarz and $\|\theta_\star\|_{\bar{V}_0^{-1}} \le 1$ imply, for $v = x_b - x_a$,

$$v^\top(\theta_\star - \widehat{\theta}_0) \le \|v\|_{\bar{V}_0^{-1}}(1 + \|S\|_{\bar{V}_0^{-1}}) \le \beta_0\|v\|_{\bar{V}_0^{-1}} + 2D_H.$$

Indeed, $\|S\|_{\bar{V}_0^{-1}} \le B_H + D_H$ and $D_H\|v\|_{\bar{V}_0^{-1}} \le D_H\|v\|_2 \le 2D_H$ since $\bar{V}_0 \succeq I$. For a difference-certified action, adding $v^\top \widehat{\theta}_0$ and maximizing over the full incoming set $A = [K_t]$ bounds the actual round regret:

$$\max_{b \in A}(x_b - x_a)^\top \theta_\star \le C_0^\Delta(a; A) + 2D_H \le 4\beta_0 u_0(a) + 2D_H.$$

The same vector bound holds for every v, so no union bound is needed; C_0^Δ uses the unclipped ridge vector. For the soft branch, substitute $C_0(a; B) \le 4\rho_0(a)$ into (20) to obtain the second assertion. In the terminal branch, the tested action $\mathrm{Best}_0(B)$ belongs to B_{cert}, so this set is nonempty. Every action in it has soft certificate at most $4\varepsilon_{L-1}$; apply (20) again. No accuracy claim for the pooled ranking is needed. The certified-wide set remains the fallback; its maximum-radius witness proves the equivalence of the wide-survivor guard and existence of a valid exploration candidate. □

C.3 Regret decomposition

Lemma 6 (Regret decomposition for layer-dependent thresholds)**.** *For arbitrary positive exploration thresholds, the choices in Algorithm 1 obey* (24)*. (No assumption $\varepsilon_s \le \varepsilon_{s-1}$ is required)*

Define $\xi_{0,t}$ on layer 0's full incoming action set. For $s \ge 1$, define $\xi_{s,t}$ on that layer's incoming set when it uses its fixed reward statistics for elimination, and set it to zero otherwise. Let N_s count observations assigned to layer s. The asserted inequality is

$$\begin{aligned} R_T \le & \sum_{t: J_t = 0} 4\beta_0 u_{0,t}(a_t) + \sum_{s=1}^{L-1} 4\varepsilon_{s-1} N_s + 4T\varepsilon_{L-1} \\ & + 4\sum_{t=1}^{T}\sum_{s=0}^{L-1} \xi_{s,t} + 2TD_H. \end{aligned} \tag{24}$$

Proof. Consider a round assigned to layer $s \ge 1$. Let $A_{0,t} = [K_t]$ and let $A_{j+1,t}$ be the survivors after layer j eliminates. Write $m_{j,t} = \max_{a \in A_{j,t}} x_{t,a}^\top \theta_\star$. Every earlier layer $j < s$ has eliminated without exploring, so Lemma 4 gives $m_{j,t} - m_{j+1,t} \le 2\xi_{j,t}$. For $s = 1$, the root's wide-survivor guard was false; for $s > 1$, layer $s-1$ passed its no-assignment test. In either case, all surviving radii at layer $s-1$ are at most ε_{s-1}. Thus every action in $A_{s,t}$ has certificate at most $4\varepsilon_{s-1}$, and its gap from $m_{s,t}$ is at most $4\varepsilon_{s-1} + 2\xi_{s-1,t}$. Since $a_t \in A_{s,t}$, telescoping gives

$$\begin{aligned} m_{0,t} - x_{t,a_t}^\top \theta_\star &= \sum_{j<s}(m_{j,t} - m_{j+1,t}) + m_{s,t} - x_{t,a_t}^\top \theta_\star \\ &\le 4\varepsilon_{s-1} + 2\sum_{j<s}\xi_{j,t} + 2\xi_{s-1,t} \\ &\le 4\varepsilon_{s-1} + 4\sum_{j<s}\xi_{j,t}. \end{aligned}$$

This uses only the previous layer's successful elimination; it makes no comparison between ε_s and ε_{s-1}. After the final elimination the same argument gives $4\varepsilon_{L-1} + 4\sum_{j<L}\xi_{j,t}$ for every terminal survivor, regardless of how the pool ranks it. Layer-0 exploration and early terminal decisions are covered by Lemma 5. The branches are disjoint, and $N_s = \sum_{t=1}^{T} \mathbf{1}\{J_t = s\}$. All branch-specific error terms are bounded by $4\sum_{j<L}\xi_{j,t} + 2D_H\mathbf{1}\{J_t = 0\}$. Sum over rounds and bound both the terminal-round count and the layer-0 assignment count by T. This proves (24) with excess coefficient four. The proof is pathwise and applies within an epoch when its public parameters are fixed. □

The remaining proof bounds each term in this decomposition. It never needs an event on which all confidence intervals hold simultaneously.

D Conditional Independence of Layer Data

The deterministic decomposition is not yet a statistical bound. For later layers we need to fix the selected design without biasing its reward noise. This section proves that property by comparing

executions whose assigned noise values differ. It also shows why an arbitrary global history is not the right conditioning for regression.

D.1 Deferred aggregation and a frozen kernel

Before an action, the cascade reads only public parameters, current contexts, frozen layer states, and the independent random variables at reached layers. The pooled score is read only after the terminal branch and its controlled action set have been fixed, either at layer 0 or after the complete cascade. These inputs have a fixed distributional rule within the batch, so they define one kernel χ_j as required by Definition 1. An algebraic shadow process may add responses immediately after each action, provided the policy can read them only through authorized snapshots. Inducting over feedback queries shows equality of actions, assignments, accumulated designs, frozen states, and endpoints with the buffered process: at every feedback query the same previously hidden terms are added before the same snapshots are copied. This is a regrouping identity, not extra reward access for the learner.

D.2 Coordinate replacement

Fix a finite evaluation prefix, its oblivious contexts, and all private random variables. For a later layer s, record its incoming indexed sets, reached flags, assignment round/action indices, selected features when assigned, accumulated and frozen designs, snapshot training-index sets, own-refresh flags, response-read flags, public parameter versions, and layer birth time. The record excludes own-response survivor sets, later-layer response states, terminal choices, and the pooled state. With contexts and random variables fixed, this record has finite range: every design is a sum over a finite collection of indexed features and every training set is finite-valued. Let $I_s(h)$ be the assignment rounds in a candidate record h.

Lemma 7 (Replacement invariance)**.** *If two noise arrays agree outside $I_s(h)$, then one realizes record h if and only if the other does. Thus the fiber of h is measurable with respect to noise coordinates outside $I_s(h)$. Conditional on the record, the assigned noise coordinates retain their original product law.*

Proof. Proceed by induction over rounds and, within each round, layers in increasing order. The state of every earlier layer is unchanged when only noises assigned to s are replaced: earlier-layer assignments were made before any read of layer s, and no later layer can authorize an earlier-layer copy. Therefore the incoming set of s is unchanged. Its accumulated and frozen designs are functions of the same past assigned indexed features. Its test probe, common radius, and public threshold are unchanged. If it assigns, the independent selection probe and its projection scores are unchanged. The preference uses only authorized frozen states of layers $j < s$, which are unchanged by the induction. Thus its selected feature is unchanged without reading its own response or the pool.

If it instead reads, replacing its responses may alter its survivors and all subsequent layers or pooled choices. These effects are excluded from the record. They can change later-layer- or pool-triggered feedback-query times, but neither kind of request copies s or an earlier layer unless that layer has its own unchanged request. The layer-0 terminal predicate depends only on the root snapshot and public parameters, so a changed pooled ranking cannot change it. At s's next own trigger, all its previously assigned responses have been received and aggregated, so its snapshot uses exactly its recorded training set, regardless of how other queries partitioned the sum. Its response values can change, but not its own routing objects. Future contexts are fixed, so they cannot transmit a changed response back into incoming designs.

At a deterministic epoch update the time and new parameters are deterministic. Every old training set is preserved; copying all layers at this predetermined time does not introduce a noise-dependent copy decision. New bottom layers have no old assignments. The same induction therefore continues through public times and layer births. This proves both implications by symmetry.

To justify the probability statement, fix a candidate record h. Its index set $I_s(h)$ is now deterministic. Invariance means that the event of obtaining h is unchanged by replacing all coordinates in $I_s(h)$ by zero. Evaluating the measurable execution at those zero coordinates expresses that event as a measurable event involving only the complement. For bounded measurable functions f_i, independence

therefore gives

$$\mathbb{E}\left[\mathbf{1}\{\text{record} = h\} \prod_{i \in I_s(h)} f_i(\eta_i)\right] = \mathbb{P}(\text{record} = h) \prod_{i \in I_s(h)} \mathbb{E} f_i(\eta_i).$$

For a positive-probability record, division gives the conditional product law. The fixed-seed record has finite range, so this argument covers its whole conditional distribution. Integrating over the independent private seeds establishes the randomized statement. □

The retained product law is a conditional assertion on a fiber. For a fixed seed array and any positive-probability restricted layer history h, let $I_s(h)$ be its recorded assigned rounds. For bounded measurable functions $(f_i)_{i \in I_s(h)}$, Lemma 7 implies

$$\mathbb{E}\left[\prod_{i \in I_s(h)} f_i(\eta_i) \,\middle|\, h\right] = \prod_{i \in I_s(h)} \mathbb{E} f_i(\eta_i). \tag{25}$$

One does not first condition on all responses or on a successful elimination event and then treat the same noises as fresh. For a current layer test, the record is stopped before the current test; only the previously used and publicly indexed seed entries are fixed. The new test remains independent of the assigned noise under this prefix law. No assumption of independence across subsequent queries, layers, or snapshots is made.

Why an unrestricted global history is not the regression conditioning. The full history can contain past responses from layer s, so conditioning on it would not leave those responses distributed as fresh noise. The regression argument uses the restricted layer history of Lemma 7, not that history. At a current layer test, stop this restricted record immediately before the new test probe. Fix seed entries by their public indices in all earlier rounds (including unused past entries), together with the current round's earlier-layer random-variable entries. Do not condition on a later layer's realized seed-consumption pattern. Conditional on these entries and the restricted record, snapshot training indices and the incoming set are fixed, assigned noises have product law, and the current test vector is independent. The same fiber induction applies to this prefix. Averaging the resulting fixed-design estimate proves a bound for the adaptively reached layer. One must not first fix the current probe and then invoke its Gaussian concentration as though it remained random.

For expected leverage accounting, by contrast, it is legitimate to condition on all past reward noises and random variables, the current contexts, and the current test probe, *before* the current independent selection probe. Regression concentration is not applied under this conditioning; only the Gaussian-max information inequality is. Distinguishing these two conditionings prevents a selection-bias error.

D.3 Action-dependent noise and the auxiliary pool

For the extension, generate independent vectors $(\eta_{t,a})_{a \in [K_t]}$ over rounds; every coordinate is centered 1-sub-Gaussian. The selected coordinate is observed. Conditional dependence among coordinates of one vector is allowed. In Lemma 7, replace the *entire* vectors for rounds in $I_s(h)$. The action at an assigned round remains fixed before any coordinate of that round is read. Thus the selected coordinates have their corresponding fixed marginal laws and remain independent across assigned rounds. All fixed-design calculations remain valid, even when the marginal law depends on action. The initial-layer martingale argument also holds because the current vector is independent of the pre-action history. No analogous product-law claim is made for arbitrary cross-round martingale noise.

The pooled estimator is never a fixed-design confidence object. It may contain fully adaptively selected observations, but its score chooses only inside a set controlled by (20), including the certified terminal set in Lemma 5. Changing a terminal choice can change pooled data and future pool-triggered endpoints, but cannot change a layer's assignments or authorize a nonrequesting layer's snapshot. The pool is copied at every feedback query and may open an additional batch through (18); such a pool-only query copies no upstream layer. The root terminal predicate is fixed before the pool

is read. Noninitial reward preference uses only authorized frozen states of earlier layers, never this pool. These restrictions preserve the replacement proof while permitting data reuse; the additional endpoints are paid explicitly in Equations (13) and (52).

E GAUSSIAN-MAX INFORMATION INEQUALITIES

We compare the exploration direction chosen using a fixed design $\bar{V}$ with the information it supplies under the accumulated design V. The goal is additive dependence on action-set complexity and design growth.

E.1 FULL PROOF OF THE GAUSSIAN-MAX INFORMATION INEQUALITY

For Lemma 1, set $v_a = \bar{V}^{-1/2}x_a$ and $M = \bar{V}^{-1/2}V\bar{V}^{-1/2} \succeq I$. Set $a(g) = \min \arg\max_{a\in A} |g^\top v_a|$ and $J = (a(g), \operatorname{sign}(g^\top v_{a(g)}))$, with sign $+1$ at zero. For $j = (a, \sigma)$, write $v_j = \sigma v_a$; then $g^\top v_J = \max_{a\in A} |g^\top v_a|$. Any fixed measurable maximizer tie rule also satisfies the unprojected inequality. The event $J = j$ is measurable. Discard zero-probability labels and put $p_j = \mathbb{P}(J = j)$, $m_j = \mathbb{E}[g \mid J = j]$. The conditional law has density $\mathbf{1}\{J = j\}/p_j$ relative to standard Gaussian measure γ_d, so

$$\mathrm{KL}(\mathcal{L}(g \mid J = j)\|\gamma_d) = \log(1/p_j).$$

For every $w \in \mathbb{R}^d$, conditional Jensen and $\mathbf{1}\{J = j\} \le 1$ give

$$e^{w^\top m_j} \le \mathbb{E}[e^{w^\top g} \mid J = j] \le p_j^{-1}\mathbb{E}e^{w^\top g} = p_j^{-1}e^{\|w\|_2^2/2}.$$

Taking logarithms and $w = m_j$ gives $\|m_j\|_2^2 \le 2\log(1/p_j)$. For $\Sigma = \sum_j p_j m_j m_j^\top$, total covariance gives

$$I = \mathrm{Cov}(g) = \mathbb{E}[\mathrm{Cov}(g \mid J)] + \mathrm{Cov}(\mathbb{E}[g \mid J]) = \mathbb{E}[\mathrm{Cov}(g \mid J)] + \Sigma.$$

Here $\mathbb{E}g = 0$, so the covariance of the conditional mean is exactly Σ; the other covariance term is positive semidefinite. Thus $\Sigma \preceq I$. Moreover, $\operatorname{tr}\Sigma = \sum_j p_j\|m_j\|_2^2 \le 2\,\mathrm{H}(J)$, where $\mathrm{H}(J) = \sum_j p_j \log(1/p_j) \le \log(2|A|)$ by concavity of log. Combining this with $\operatorname{tr}\Sigma \le d$ gives

$$\operatorname{tr}\Sigma \le \min\{d, 2\,\mathrm{H}(J)\} \le \min\{d, 2\log(2|A|)\}.$$

Since v_J is determined by J, conditioning gives $W = \mathbb{E}[g^\top v_J] = \mathbb{E}[m_J^\top v_J]$. Cauchy–Schwarz for the random vectors $M^{-1/2}v_J$ and $M^{1/2}m_J$ yields

$$W^2 \le \mathbb{E}[v_J^\top M^{-1} v_J]\,\mathbb{E}[m_J^\top M m_J] = \mathbb{E}\|x_{a(g)}\|_{V^{-1}}^2 \operatorname{tr}(M\Sigma).$$

Here $M^{-1} = \bar{V}^{1/2}V^{-1}\bar{V}^{1/2}$, the sign squares to one, and $\mathbb{E}[m_J m_J^\top] = \Sigma$. For $B, C \succeq 0$, $\operatorname{tr}(BC) = \operatorname{tr}(B^{1/2}CB^{1/2}) \ge 0$. Apply this to $B = M - I$ and $C = I - \Sigma$ to obtain

$$\begin{aligned}\operatorname{tr}(M\Sigma) &= \operatorname{tr}\Sigma + \operatorname{tr}((M - I)\Sigma) \\ &\le \min\{d, 2\log(2|A|)\} + \operatorname{tr}(M - I).\end{aligned}$$

If $\mu_i \ge 1$ are the eigenvalues of M, expanding $\prod_i(1 + (\mu_i - 1))$ and retaining only its constant and linear terms gives $\operatorname{tr}(M - I) = \sum_i(\mu_i - 1) \le \prod_i \mu_i - 1 = \det M - 1$. Finally $\det M = \det V/\det\bar{V}$. This proves (8). A nonempty zero feature set has width and selected leverage both zero. Repeated features, opposite features, and positive-probability ties do not affect the argument: only the measurable finite signed label is used. Orthogonal changes from the symmetric root to the Cholesky whitening convention preserve the calculation.

E.2 PROOF OF THE PROJECTED SELECTION-COVARIANCE BOUND

Proposition 2 (Selection covariance and active directions)**.** *With M, J, Σ from the preceding proof, let P be the orthogonal projector onto $\operatorname{span}\{\bar{V}^{-1/2}x_a : a \in A\}$ and $r = \operatorname{rank} P$. Under the fixed-index tie rule, writing $\mathcal{L} = \mathbb{E}_g\|x_{a(g)}\|_{V^{-1}}^2$, one has*

$$W^2 \le \mathcal{L}\operatorname{tr}(M\Sigma) \le \mathcal{L}\left[\min\{r, 2\,\mathrm{H}(J)\} + \operatorname{tr}\{P(M - I)P\}\right]. \tag{26}$$

Since $v_a = Pv_a$, every score satisfies $g^\top v_a = (Pg)^\top v_a$, so the fixed-index rule makes J a function of Pg. The jointly Gaussian vectors Pg and $(I-P)g$ are independent: their cross-covariance is $P(I-P) = 0$. Thus $\mathbb{E}[(I-P)g \mid J] = 0$ and $m_J = \mathbb{E}[Pg \mid J]$. Since $\mathrm{Cov}(Pg) = P$, the same total-covariance identity gives

$$0 \preceq \Sigma = \mathbb{E}[m_J m_J^\top] \preceq P.$$

The entropy argument above gives $\mathrm{tr}\,\Sigma \le 2\,\mathrm{H}(J)$; the covariance bound gives $\mathrm{tr}\,\Sigma \le r$. The same metric Cauchy–Schwarz calculation yields $W^2 \le \mathcal{L}\,\mathrm{tr}(M\Sigma)$. Because $M - I \succeq 0$ and $P - \Sigma \succeq 0$,

$$\mathrm{tr}(M\Sigma) \le \min\{r, 2\,\mathrm{H}(J)\} + \mathrm{tr}((M-I)P) = \min\{r, 2\,\mathrm{H}(J)\} + \mathrm{tr}\{P(M-I)P\}.$$

Trace cyclicity and $P^2 = P$, not matrix commutativity, prove the last identity. This proves Proposition 2. If every feature is zero, both width and leverage vanish and the inequality is immediate.

Why the tie rule is part of the statement. For duplicate actions $x_1 = x_2 = e_1$ in dimension two and $\bar{V} = I$, a tie rule that chooses an index from the sign of g_2 would make the signed label encode both signs of g. The four equally likely conditional means are $\sqrt{2/\pi}(\pm 1, \pm 1)$, so $\Sigma = (2/\pi)I_2$ is not bounded by $P = \mathrm{diag}(1, 0)$. Our fixed-index rule excludes this irrelevant information. The original $\Sigma \preceq I$ bound remains true under that alternative rule; it is the refined $\Sigma \preceq P$ statement that needs score-measurable tie breaking. Orthogonal changes between symmetric and Cholesky whitening preserve the argument.

E.3 The information inequality for the selection rule

Corollary 1. *Let $0 < \alpha \le 1$ and let $a(g)$ be any measurable choice satisfying*

$$|g^\top \bar{V}^{-1/2} x_{a(g)}| \ge \alpha \max_{b \in A} |g^\top \bar{V}^{-1/2} x_b|.$$

Under the matrix assumptions of Lemma 1,

$$W(A, \bar{V})^2 \le \alpha^{-2} \left[\min\{d, 2\log(2|A|)\} + \frac{\det V}{\det \bar{V}} - 1 \right] \mathbb{E}_g \|x_{a(g)}\|^2_{V^{-1}}. \tag{27}$$

If the selection, including its tie rule, depends on g only through the projections on the whitened feature span, Proposition 2 also holds with its left-hand side replaced by $\alpha^2 W^2$.

Proof. Let $J = (a(g), \mathrm{sign}(g^\top v_{a(g)}))$ with sign $+1$ at zero. Define p_j, m_j, Σ anew from this chosen label. The defining inequality gives $\alpha W \le \mathbb{E}[g^\top v_J] = \mathbb{E}[m_J^\top v_J]$. The entropy and total-covariance bounds above use only the finite label, not exact maximization. The same Cauchy–Schwarz calculation therefore gives

$$\alpha^2 W^2 \le \mathbb{E}\|x_{a(g)}\|^2_{V^{-1}}\ \mathrm{tr}(M\Sigma).$$

The same trace–determinant calculation proves (27). Projection-measurable selection gives $\Sigma \preceq P$ by the independence of the orthogonal Gaussian component, so the preceding projected proof also applies. Zero width and repeated features are handled by the same conventions. □

In Algorithm 1, $\alpha = 1/\sqrt{2}$ and the earlier-layer scores are fixed under the pre-selector conditioning. They depend on neither unused components of g_{sel} nor the layer's own responses. Thus the selection rule satisfies both assertions of Corollary 1; it pays factor two rather than assuming exact maximization. The uniform sharpness result below concerns the original exact-max inequality itself, not a claim of optimality of this factor for the reward-preferring policy.

E.4 Uniform sharpness

Corollary 2 (Uniform sharpness)**.** *For $d \ge 1$, $k \ge 2$, and $D \ge 1$, the right-hand factor in Lemma 1 is at most $2[\min\{d, \log(2k)\} + D - 1]$ whenever $|A| \le k$ and $\det V / \det \bar{V} \le D$. Conversely, there are at most k unit-norm features, at least one nonzero, and $V \succeq \bar{V} \succ 0$ with this determinant constraint such that*

$$\frac{W(A, \bar{V})^2}{\mathbb{E}_g \|x_{a(g)}\|^2_{V^{-1}}} \ge \frac{\min\{d, \log(2k)\} + D - 1}{4\pi}. \tag{28}$$

Proof. The upper bound follows from Lemma 1 and $\min\{d, 2\log(2k)\} \le 2\min\{d, \log(2k)\}$. For the complexity lower bound put $r = \min\{d, \lfloor \log_2 k \rfloor\} \ge 1$ and use the 2^r features

$$x_\epsilon = r^{-1/2}(\epsilon_1, \dots, \epsilon_r, 0, \dots, 0), \quad \epsilon \in \{-1, 1\}^r, \qquad V = \bar{V} = I.$$

Each feature has norm one. The triangle inequality gives $\max_\epsilon |g^\top x_\epsilon| \le r^{-1/2} \sum_{i=1}^r |g_i|$, with equality for $\epsilon_i = \operatorname{sign}(g_i)$. Since $\mathbb{E}|g_i| = \sqrt{2/\pi}$, $W = \sqrt{2r/\pi}$; thus width squared is $2r/\pi$ and selected squared leverage is one. For the drift lower bound use $A = \{e_1, -e_1\}$, $\bar{V} = I$, and $V = \operatorname{diag}(D, 1, \dots, 1)$. Width squared is $2/\pi$ and selected squared leverage is $1/D$. Both examples satisfy the same constraints; choose whichever has the larger ratio. That ratio is at least $\max\{2r/\pi, 2D/\pi\} \ge (r + D)/\pi$. If $r = d$, then $\min\{d, \log(2k)\} \le r$. Otherwise $k < 2^{r+1}$ and $\min\{d, \log(2k)\} \le \log(2k) < (r + 2)\log 2 \le 4r$. Thus $\min\{d, \log(2k)\} \le 4r$ in either case, and $(r + D)/\pi \ge [\min\{d, \log(2k)\} + D - 1]/(4\pi)$ since $D \ge 1$. This proves (28). □

These examples establish worst-case sharpness of the matrix inequality, not necessity of its factor for every policy's update count. The projected refinement likewise does not establish a local-rank version of Theorem 1.

F Single-Test Calibration and Prediction Excess

By Lemma 7, the restricted pre-test record fixes a later layer's snapshot training indices, design, and incoming features, but not its response sum. Conditional on this record, assigned noises retain their product law and the fresh Gaussian test remains independent. We first bound stochastic prediction error for such a deterministic design, then combine that bound with a fresh Gaussian test. The scalar comparison and Gaussian concentration used here are stated in Section A.5.

F.1 Fixed-design prediction errors

Lemma 8 (Numerical fixed-design bound)**.** *Let $V = I + X^\top X$, with X deterministic and independent centered 1-sub-Gaussian noise coordinates η_i. For a fixed incoming feature set, put*

$$Z = \max_{a \in A} |x_a^\top V^{-1} X^\top \eta|, \qquad W = \mathbb{E}_g \max_{a \in A} |g^\top V^{-1/2} x_a|.$$

For $v, w \ge 0$ and $c = \sqrt{\pi/2}$,

$$\begin{aligned} \mathbb{E}e^{vZ} &\le \exp\left(cvW + \tfrac{1}{2} c^2 v^2 \max_{a \in A} \|x_a\|_{V^{-1}}^2 \right), \\ \mathbb{P}\left\{ Z > cW + c\sqrt{2w} \max_{a \in A} \|x_a\|_{V^{-1}} \right\} &\le e^{-w}. \end{aligned} \tag{29}$$

Proof. For $v \ge 0$, the function $\eta \mapsto \exp(v \max_a |x_a^\top V^{-1} X^\top \eta|)$ is convex: the maximum absolute linear form is convex and the exponential is convex and nondecreasing. Hold all noise coordinates but the first fixed and use Lemma 3; independence permits integration over the others. Repeat for each coordinate with fresh independent Gaussian replacements. This bounds the original expectation by its value at c times an independent standard Gaussian noise vector. All exponential moments are finite by the finite action set and the sub-Gaussian hypotheses. Let $G \sim N(0, I_n)$, where n is the number of rows of X, and let $G' \sim N(0, I_d)$ be independent. Set $Y = V^{-1} X^\top G$ and $Q = V^{-1} G'$. Their covariances are $V^{-1} - V^{-2}$ and V^{-2}, respectively, so $Y + Q \sim N(0, V^{-1})$. For the convex function $F(h) = \exp(cv \max_{a \in A} |x_a^\top h|)$, conditional Jensen gives

$$F(Y) = F(\mathbb{E}[Y + Q \mid Y]) \le \mathbb{E}[F(Y + Q) \mid Y].$$

Taking expectations after the coordinate replacements therefore gives

$$\mathbb{E}e^{vZ} \le \mathbb{E} \exp\left(cv \max_{a \in A} |g^\top V^{-1/2} x_a| \right).$$

Write $u = \max_{a\in A} \|x_a\|_{V^{-1}}$ and $f(g) = \max_{a\in A} |g^\top V^{-1/2} x_a|$. Since $|f(g)-f(h)| \le \max_a |(g-h)^\top V^{-1/2} x_a| \le u\|g-h\|_2$, scaled (17) at argument cv proves the moment bound. For $u > 0$, exponential Markov gives, for $y \ge 0$,

$$\mathbb{P}\{Z > cW + y\} \le \inf_{v\ge 0} e^{-vy + c^2v^2u^2/2} = e^{-y^2/(2c^2u^2)}.$$

The minimum is attained at $v = y/(c^2u^2)$; take $y = cu\sqrt{2w}$. When $u = 0$, $Z = W = 0$ and the assertions hold directly. □

F.2 Prediction excess with one test sample

Integrate excess, not truncated error. For arbitrary nonnegative random variables Z, B, Tonelli's theorem gives

$$\begin{aligned} \mathbb{E}(Z-B)_+ &= \int_0^\infty \mathbb{P}(Z > B + x)\, dx, \\ \mathbb{E}[(Z-B)_+^2] &= \int_0^\infty 2x\, \mathbb{P}(Z > B + x)\, dx. \end{aligned} \tag{30}$$

These identities need no independence between the error and the radius. In contrast, for deterministic $a \ge 0$,

$$\mathbb{E}[Z\mathbf{1}\{Z > a\}] = a\mathbb{P}(Z > a) + \int_a^\infty \mathbb{P}(Z > x)\, dx. \tag{31}$$

The boundary term cannot be dropped. The deterministic elimination lemma charges (error − radius)$_+$, so (30) is the relevant identity, even for a randomized radius.

Fix a restricted pre-test record as above; suppress this conditioning in expectations and probabilities below. For its fixed design and incoming set, the ridge identity is

$$x_a^\top(\widehat{\theta} - \theta_\star) = -x_a^\top V^{-1}\theta_\star + x_a^\top V^{-1} X^\top \eta.$$

The bias is bounded in absolute value by $b = \max_{a\in A} \|V^{-1}x_a\|_2$ since $\|\theta_\star\|_2 \le 1$. Write $u = \max_{a\in A} \|x_a\|_{V^{-1}}$, $M = \max_{a\in A} |g_{\text{test}}^\top V^{-1/2} x_a|$, and $E_{\max} = \max_{a\in A} |\widehat{\mu}(a) - x_a^\top \theta_\star|$. Clipping gives $E_{\max} \le b + Z$. Under this conditioning, Z and M are independent and $\mathbb{E}M = W$. For $v \ge 0$, scaled Gaussian concentration at argument $-vc$ gives

$$\mathbb{E}\exp\left(-vc \max_{a\in A} |g_{\text{test}}^\top V^{-1/2} x_a|\right) \le \exp\left(-vcW + \tfrac{1}{2}c^2v^2 \max_{a\in A} \|x_a\|_{V^{-1}}^2\right).$$

Independence gives $\mathbb{E}e^{v(Z-cM)} = (\mathbb{E}e^{vZ})(\mathbb{E}e^{-vcM})$. Multiplication with (29) cancels $cvW - vcW$. For $u > 0$, exponential Markov bounds the tail by $e^{-vx + c^2v^2u^2}$, minimized at $v = x/(2c^2u^2)$. Thus

$$\begin{aligned} \mathbb{E}\exp\left(v\left[Z - c\max_{a\in A} |g_{\text{test}}^\top V^{-1/2} x_a|\right]\right) &\le \exp\left(c^2v^2 \max_{a\in A} \|x_a\|_{V^{-1}}^2\right), \\ \mathbb{P}\left\{Z - c\max_{a\in A} |g_{\text{test}}^\top V^{-1/2} x_a| > x\right\} &\le \exp\left(-\frac{x^2}{4c^2 \max_{a\in A} \|x_a\|_{V^{-1}}^2}\right) \quad (x \ge 0). \end{aligned} \tag{32}$$

A zero maximum norm has zero excess and is treated directly. Otherwise, for $\ell \ge 0$, define the auxiliary excess, whether or not a read occurs,

$$\xi = (E_{\max} - b - cM - 2c\sqrt{\ell}\, u)_+ \le (Z - cM - 2c\sqrt{\ell}\, u)_+.$$

For $x \ge 0$, $\{\xi > x\}$ therefore requires $Z - cM > 2c\sqrt{\ell}\, u + x$. Apply (32) and use $(2c\sqrt{\ell}\, u + x)^2 \ge 4c^2\ell u^2 + x^2$ to obtain

$$\begin{aligned} \mathbb{P}\{\xi > x\} &\le \exp\left[-\frac{(2c\sqrt{\ell} \max_{a\in A} \|x_a\|_{V^{-1}} + x)^2}{4c^2 \max_{a\in A} \|x_a\|_{V^{-1}}^2}\right] \\ &\le e^{-\ell} \exp\left(-\frac{x^2}{4c^2 \max_{a\in A} \|x_a\|_{V^{-1}}^2}\right). \end{aligned} \tag{33}$$

In (30), use $\int_0^\infty e^{-x^2/(4c^2u^2)}\,dx = \sqrt{\pi}\,cu$ and $\int_0^\infty 2xe^{-x^2/(4c^2u^2)}\,dx = 4c^2u^2$ to obtain

$$\mathbb{E}\xi \le \sqrt{\pi}\,ce^{-\ell}\max_{a\in A}\|x_a\|_{V^{-1}}, \qquad \mathbb{E}\xi^2 \le 4c^2e^{-\ell}\max_{a\in A}\|x_a\|^2_{V^{-1}}. \tag{34}$$

No independence among prediction errors at different actions is used.

For a reached layer $s \ge 1$, let $\mathcal{H}_{s,t}$ denote the restricted pre-test information of Section D, and substitute $V = \bar{V}_s$, $\ell = \ell_s$. Then u is fixed given $\mathcal{H}_{s,t}$. The layer reads exactly when $\rho_s \le \varepsilon_s$, and $\rho_s = b + cM + 2c\sqrt{\ell_s}u \ge 2c\sqrt{\ell_s}u$. Hence

$$\xi_{s,t} = \xi\mathbf{1}\{\rho_s \le \varepsilon_s\} \le \xi\mathbf{1}\{u \le \varepsilon_s/(2c\sqrt{\ell_s})\}.$$

The last indicator, unlike the read outcome, is fixed before the test. Thus (34) gives

$$\begin{aligned}\mathbb{E}[\xi^2_{s,t} \mid \mathcal{H}_{s,t}] &\le \mathbf{1}\{u \le \varepsilon_s/(2c\sqrt{\ell_s})\}4c^2e^{-\ell_s}u^2 \\ &\le \frac{\varepsilon_s^2 e^{-\ell_s}}{\ell_s} \le \tau_s^2.\end{aligned}$$

Indeed, (43) implies $\varepsilon_s e^{-\ell_s/2} \le \max\{\varepsilon_s, \varepsilon_{s-1}\}e^{-\ell_s/2} \le \tau_s$, and $\ell_s \ge 1$. Averaging over the restricted record gives $\mathbb{E}\xi^2_{s,t} \le \tau_s^2$; Jensen then gives $\mathbb{E}\xi_{s,t} \le (\mathbb{E}\xi^2_{s,t})^{1/2} \le \tau_s$. An unreached layer contributes zero excess. This proves the moment claims of Lemma 2 without conditioning on the read outcome. Section G and Lemma 10 later verify that inequality and the sum of budgets for the numerical parameter choice. The calibration identity holds on both cap branches, although the regret proof only needs it on the uncapped branch. It does not require simultaneous relative accuracy of Gaussian tests over all queries.

G Parameter Choice and Fixed-Horizon Guarantees

The previous sections provide the two ingredients for selecting parameters: an action explored at layer $s \ge 1$ costs $4\varepsilon_{s-1}$ plus prediction excess, and its expected information depends on ε_s^2. Layer 0 instead contributes a term proportional to $1/\varepsilon_0$. These different powers explain why the exploration thresholds are optimized jointly rather than chosen as a decreasing geometric sequence. We now specify the exact numerical policy, verify its inequalities, and combine the bounds. None of the following values is selected from experimental outcomes.

G.1 Determinant staleness and potential

Before action selection, the pre-context trigger guarantees

$$\bar{V}_s \preceq V_s \prec e^{z_s}\bar{V}_s, \qquad x^\top \bar{V}_s^{-1}x \le e^{z_s}x^\top V_s^{-1}x. \tag{35}$$

Indeed all eigenvalues of $\bar{V}_s^{-1/2}V_s\bar{V}_s^{-1/2}$ are at least one and their product is below e^{z_s} unless the snapshot has just been copied, in which case the claim is immediate. The pool satisfies the same inequality with s replaced by all and z_s by z_0, because its own trigger is checked before every context. For any layer's assigned sequence, accumulated-design leverages are at most one and the matrix determinant lemma gives

$$\sum_{t:J_t=s}\|x_{t,a_t}\|^2_{V_{s,t}^{-1}} \le 2\sum_{t:J_t=s}\log(1+\|x_{t,a_t}\|^2_{V_{s,t}^{-1}}) \le 2d\log(1+H/d). \tag{36}$$

The final determinant is at most $(1+H/d)^d$ by trace–determinant comparison. For a subinterval the middle sum telescopes between its starting and ending designs, so (36) remains true with causally retained data, provided the full global prefix through its end has at most H rounds.

G.2 Deterministic numerical parameters

For public bound $H \ge 1$, dimension $d \ge 1$, and action-count bound $K \ge 2$, put $D = d\log(1+H/d)$, $G = d + D$, and $c = \sqrt{\pi/2}$. Set

$$\begin{aligned}L &= \max\left\{2, \left\lceil \log_2\left(1 + \tfrac{1}{2}\log G\right)\right\rceil\right\}, \\ \lambda_0 &= 2^{-L}, \qquad \lambda_s = 2^{s-1-L} \quad (1 \le s < L).\end{aligned} \tag{37}$$

The formula gives $L \geq 2$, but does not fix $L = 2$ for all public inputs. The weights satisfy $\sum_{s<L} \lambda_s = 1/2$, $1 \leq \sum_{s<L} \sqrt{\lambda_s} < \sqrt{3}$, and $\sum_{s=1}^{L-1}(L-s)\lambda_s < 1$. Define

$$\begin{aligned}
\beta_0 &= 1 + \sqrt{d\log(1+H/d) + 8\log(eH)},\\
z_0 &= \max\left\{\frac{\log(1+(1+\log(e+1))^{-1/2})}{\sqrt{\lambda_0}\sum_{s<L}\sqrt{\lambda_s}},\ \log(1+\min\{d,\log K\})\right\}, \qquad (38)\\
a_0 &= 8d\log(1+H/d)\beta_0^2 e^{z_0},\\
z_s &= \log(2+\min\{d,\log K\}) + 2^{(L-s-5)/2},\\
a_s &= 2^{14}c^2 d\log(1+H/d)[2(2+\min\{d,\log K\}) + e^{z_s}]2^{L-s} \quad (1 \leq s < L). \qquad (39)
\end{aligned}$$

The pool uses threshold z_0. The same root formulas are evaluated with the depth and weights in (37); they are not separately recalibrated from a failure probability or an observed performance criterion.

The minimum of the uncapped accounting objective and the implemented thresholds are

$$\Phi = 2^{3-2^{1-L}}\sqrt{H}\prod_{s=0}^{L-1} a_s^{\lambda_s}, \qquad (40)$$

$$\begin{aligned}
\varepsilon_{L-1} &= \min\left\{\frac{1}{2}, \frac{\Phi}{8H}\right\},\\
\varepsilon_{s-1} &= \frac{\lambda_s\Phi}{a_s}\max\left\{1, \frac{\Phi}{4H}\right\}\varepsilon_s^2 \quad (s = L-1,\dots,1). \qquad (41)
\end{aligned}$$

Every backward step uses the cap multiplier. Compute the budgets

$$\tau_s = \min\left\{1, \frac{\Phi}{H}\right\}\frac{\lambda_s}{1/2 - 2^{-L}} \quad (1 \leq s < L), \qquad (42)$$

then set

$$\ell_s = \max\left\{1,\ 2\log\frac{\max\{\varepsilon_s, \varepsilon_{s-1}\}}{\tau_s}\right\} \quad (1 \leq s < L). \qquad (43)$$

Thus the order is $(L, \lambda_s, \beta_0, z_s, a_s)$, then Φ, then the backward thresholds, then τ_s, and finally ℓ_s. The budgets enter the parameter specification through (43); they are not an extra live input to the deployed kernel. The coefficient 2^{14} pays for the selection rule in Algorithm 1, including its factor-two information cost. Lemma 10 proves $1 \leq \ell_s \leq 20\,2^{L-s}$ without defining a_s circularly from ℓ_s. All quantities depend only on public inputs. Exponentials and products can be evaluated through their logarithms.

G.3 Initial-layer exploration cost

A layer-0 assignment has $\beta_0\|x_t\|_{\bar{V}_0^{-1}} > \varepsilon_0$ and nominal regret certificate at most $4\beta_0\|x_t\|_{\bar{V}_0^{-1}}$. Writing $u = \|x_t\|_{\bar{V}_0^{-1}}$, the guard gives $4\beta_0 u \leq (4\beta_0^2/\varepsilon_0)u^2$. Before any action, determinant monitoring gives $\|x_t\|^2_{\bar{V}_0^{-1}} \leq e^{z_0}\|x_t\|^2_{V_0^{-1}}$. The accumulated squared leverages sum to at most $2d\log(1+H/d)$. Consequently the total nominal cost of these assignments is at most

$$\frac{4\beta_0^2 e^{z_0}}{\varepsilon_0}\sum_{t:J_t=0}\|x_t\|^2_{V_{0,t}^{-1}} \leq \frac{8d\log(1+H/d)\beta_0^2 e^{z_0}}{\varepsilon_0} = \frac{a_0}{\varepsilon_0}.$$

Thus the coefficient in (10) is exactly the one in (38); no additional initial-layer cost is omitted.

G.4 Exact optimizer, including the cap

Lemma 9 (Dyadic optimizer)**.** *For arbitrary positive coefficients $a_0,\dots,a_{L-1}$, the minimum of* (10) *over positive thresholds is* (40)*. Its unique minimizer is $\varepsilon_{L-1} = \Phi/(8H)$ and $\varepsilon_{s-1} = (\lambda_s\Phi/a_s)\varepsilon_s^2$. With the constraint $\varepsilon_{L-1} \leq 1/2$, the unique minimizer is* (41)*.*

Proof. Apply weighted AM–GM to the $L+1$ summands of (10), divided by their respective weights $\lambda_0, \dots, \lambda_{L-1}, 1/2$. The weights sum to one. In their weighted product, the exponent of ε_0 is $-\lambda_0 + \lambda_1 = 0$; the exponents of the intermediate thresholds are $-2\lambda_j + \lambda_{j+1} = 0$ for $1 \le j \le L-2$; and the last is $-2\lambda_{L-1} + 1/2 = 0$. Consequently all threshold variables cancel. Using $-\sum_{s<L} \lambda_s \log_2 \lambda_s = 3/2 - 2^{1-L}$, the remaining product is

$$\frac{(4H)^{1/2} \prod_{s<L} a_s^{\lambda_s}}{(1/2)^{1/2} \prod_{s<L} \lambda_s^{\lambda_s}} = 2^{3-2^{1-L}} \sqrt{H} \prod_{s<L} a_s^{\lambda_s} = \Phi.$$

Equality holds when every objective summand is its weight times Φ. The terminal summand gives $\varepsilon_{L-1} = \Phi/(8H)$, and the layer-$s$ summand gives the stated backward recurrence. The recurrence determines every threshold uniquely, and the product identity then forces the root summand to equal $\lambda_0 \Phi$ as well. Thus the lower bound is attained at exactly one point.

Fix $\varepsilon_{L-1} > 0$. Applying AM–GM to the first L summands with weights $2\lambda_s$ cancels every threshold except ε_{L-1}. Their weighted product is $P(\varepsilon_{L-1}) = \frac{1}{2}[\Phi/\sqrt{8H\varepsilon_{L-1}}]^2 = \Phi^2/(16H\varepsilon_{L-1})$, attained by the same backward equality argument. The remaining one-variable objective is therefore

$$f(\varepsilon_{L-1}) = \frac{\Phi^2}{16H\varepsilon_{L-1}} + 4H\varepsilon_{L-1}, \qquad f'(\varepsilon_{L-1}) = 4H - \frac{\Phi^2}{16H\varepsilon_{L-1}^2}.$$

It decreases up to $\varepsilon_{L-1} = \Phi/(8H)$ and increases afterwards. Hence the constrained optimum is $\varepsilon_{L-1} = \min\{1/2, \Phi/(8H)\}$. If $\Phi > 4H$, then $\varepsilon_{L-1} = 1/2$ and each prefix summand equals $2\lambda_s P(1/2) = \lambda_s \Phi^2/(4H)$. Thus every backward coefficient is $\lambda_s \Phi^2/(4Ha_s)$, exactly as in (41). The full capped minimum is $\Phi^2/(8H) + 2H$. We do not use this larger value for regret: the pathwise ceiling $2H < \Phi/2$ suffices on this branch. □

Lemma 10 (Verification of the numerical parameters)**.** *For all public inputs,*

$$\begin{aligned} \Phi^2 &\le 2^{34} c^2 dH \log(1 + H/d)(2 + \min\{d, \log K\}), \\ \frac{\Phi^2}{Ha_s} &\le \frac{2^{20}}{3}\, 2^{-(L-s)} \quad (1 \le s < L), \qquad c = \sqrt{\pi/2}. \end{aligned} \tag{44}$$

On both cap branches,

$$\begin{aligned} 1 \le \ell_s \le 20\, 2^{L-s}, \qquad \max\{\varepsilon_s, \varepsilon_{s-1}\} &\le \tau_s e^{\ell_s/2} \quad (1 \le s < L), \\ H \sum_{s=1}^{L-1} \tau_s &\le \Phi. \end{aligned} \tag{45}$$

Proof. Write $D = d \log(1 + H/d)$, and $G = d + D$. As $(1 + H/d)^d \ge 1 + H$ for integer d, $\log(eH) \le 1 + D$. Using $(1 + \sqrt{u})^2 \le 2 + 2u$ and $d \ge 1$ gives $\beta_0^2 \le 2 + 2[D + 8(1 + D)] \le 20G$. The depth gives $2^L \ge 1 + \frac{1}{2} \log G$ and $L \ge 2$. Consequently

$$\log((\beta_0^2)^{2\lambda_0}) \le 2^{1-L} \log(20G) \le 4 + \tfrac{1}{2} \log 20.$$

The last inequality uses $2^{1-L} \le 1/2$ for the $\log 20$ term and $2^{1-L} \log G \le 4$ for the other term. Since $\sum_s \sqrt{\lambda_s} \ge 1$ and $\log(1 + (1 + \log(e+1))^{-1/2}) < 1$, $z_0 \le \log(2 + \min\{d, \log K\}) + \lambda_0^{-1/2}$. For $j = L - s \ge 1$, put $B_s = 2(2 + \min\{d, \log K\}) + e^{z_s}$. Then

$$\begin{aligned} a_0 &\le 8D(2 + \min\{d, \log K\}) \beta_0^2 e^{\lambda_0^{-1/2}}, \\ 3(2 + \min\{d, \log K\}) \le B_s &\le 3(2 + \min\{d, \log K\}) e^{1/(8\sqrt{\lambda_s})}. \end{aligned}$$

Put $C_* = 2^{14} c^2 D(2 + \min\{d, \log K\})$. Since $8 \le 2^{14} c^2$, we may enlarge the root coefficient to $2^{14} c^2$ for the product estimate below. Expanding (40) squared, the common coefficient contributes $C_*^{2\sum_s \lambda_s} = C_*$, while the remaining factors give

$$\begin{aligned} \frac{\Phi^2}{HC_*} &\le 2^6 (\beta_0^2)^{2\lambda_0} 3^{2\sum_{s\ge1} \lambda_s} 2^{2\sum_{s\ge1}(L-s)\lambda_s} \\ &\quad \times e^{2\sqrt{\lambda_0} + \frac{1}{4}\sum_{s\ge1}\sqrt{\lambda_s}} \\ &\le 2^6 \cdot 3 \cdot 4\, e^{4 + (\log 20)/2 + 1 + \sqrt{3}/4} < 2^{20}. \end{aligned}$$

Here $\sum_{s\ge1}\lambda_s \le 1/2$, $\sum_{s\ge1}(L-s)\lambda_s < 1$, and $\sum_s \sqrt{\lambda_s} < \sqrt{3}$; multiplying by HC_* proves the first bound. The root threshold exponent contributes at most $2\sqrt{\lambda_0} \le 1$, and the higher threshold exponents at most $\frac{1}{4}\sum_{s\ge1}\sqrt{\lambda_s} < \sqrt{3}/4$. Dividing by Ha_s and using $B_s \ge 3(2+\min\{d,\log K\})$ proves the second bound.

Let $\delta_\lambda = 1/2 - 2^{-L} \ge 1/4$ and $m = \max\{1, \Phi/(4H)\}$. Directly from the three ranges $\Phi/H \le 1$, $1 < \Phi/H \le 4$, and $\Phi/H > 4$,

$$\frac{\varepsilon_{L-1}}{\tau_{L-1}} \le 1, \qquad \Phi m \min\{1, \Phi/H\} \le \Phi^2/H.$$

Thus the multiplier in a backward budget ratio obeys

$$M_s := \frac{2\lambda_s \Phi m \tau_s}{a_s} \le \frac{2\lambda_s^2}{\delta_\lambda}\frac{\Phi^2}{Ha_s} < 2^{20-3j}, \qquad j = L-s.$$

For $s \ge 2$, $\tau_{s-1} = \tau_s/2$ and $\varepsilon_{s-1}/\tau_{s-1} = M_s(\varepsilon_s/\tau_s)^2$. For this paragraph only introduce $\widetilde{\tau}_0 = \tau_1/2$; then the same recurrence holds for the last step to layer 0. Let $v_j = \log_+(\varepsilon_{L-j}/\tau_{L-j})$ for $1 \le j < L$ and $v_L = \log_+(\varepsilon_0/\widetilde{\tau}_0)$. Starting at $v_1 = 0$, the recurrence gives

$$\begin{aligned} v_{j+1} &\le 2v_j + (20-3j)_+ \log 2, \\ \frac{v_j}{2^j} &\le (\log 2)\sum_{k=1}^{j-1}\frac{(20-3k)_+}{2^{k+1}} \le (\log 2)\sum_{k=1}^{6}\frac{20-3k}{2^{k+1}}. \end{aligned} \tag{46}$$

To bound ℓ_s, we need the positive logarithms of the budget ratios. For $j = L-s$, the parent budget is $\tau_s/2$ (including $\widetilde{\tau}_0$), so, writing $C_{\mathrm{rat}} = (225/32)\log 2$ for the last sum above,

$$\log_+ \frac{\max\{\varepsilon_s, \varepsilon_{s-1}\}}{\tau_s} \le \max\{v_j, v_{j+1}\} \le C_{\mathrm{rat}} 2^{j+1}.$$

Thus (43) gives $1 \le \ell_s \le \max\{1, 4C_{\mathrm{rat}}2^j\} \le 20\,2^j$, since $4C_{\mathrm{rat}} < 20$. The parameter inequality in (45) then follows directly from the definition of ℓ_s. Finally, $\sum_{s\ge1}\lambda_s = \delta_\lambda$ gives $H\sum_s \tau_s = H\min\{1, \Phi/H\} \le \Phi$. The coefficient proof used no realized value of ℓ_s to define a_s; the public computation is noncircular. □

G.5 Informative assignments

For each reached later layer, let $W = W(A, \bar{V}_s)$, $b = \max_{a\in A}\|\bar{V}_s^{-1}x_a\|_2$, and $\rho_s^\circ = b + cW + 2c\sqrt{\ell_s}\max_{a\in A}\|x_a\|_{\bar{V}_s^{-1}}$ with $c = \sqrt{\pi/2}$. Call an assignment *informative* when $\rho_s^\circ \ge \varepsilon_s/2$. The remaining assignments have $\rho_s^\circ < \varepsilon_s/2$ and are controlled by a Gaussian tail. This classification is for analysis only; the policy does not evaluate ρ_s°.

Write $u = \max_{a\in A}\|x_a\|_{\bar{V}_s^{-1}}$. An action attaining u has Gaussian projection $N(0,u^2)$, so $W \ge \mathbb{E}|N(0,u^2)| = u/c$; also $b \le u$ by Section A.2. Hence

$$\rho_s^\circ \le cW(2+2c\sqrt{\ell_s}), \qquad W^2 \ge \frac{\varepsilon_s^2}{4c^2(2+2c\sqrt{\ell_s})^2} \quad \text{on an informative assignment.}$$

At the information-accounting conditioning of Section D, the indicator, incoming set, designs, and earlier-layer preference are fixed; only the independent selector is unconditioned. With $\alpha = 1/\sqrt{2}$, Corollary 1 gives

$$W^2 \le \alpha^{-2}[\min\{d, 2\log(2|A|)\} + e^{z_s} - 1]\mathbb{E}_{\mathrm{sel}}\|x_{a(g)}\|^2_{V_s^{-1}}.$$

Put $D = d\log(1+H/d)$, $j = L-s$, and $B_s = 2(2+\min\{d,\log K\}) + e^{z_s}$. The bracket is at most B_s. To obtain (12), it suffices that

$$a_s \ge 32Dc^2(2+2c\sqrt{\ell_s})^2 B_s/\alpha^2.$$

By Lemma 10, $\ell_s \le 20\,2^j$ and $j \ge 1$, so

$$\frac{32(2+2c\sqrt{\ell_s})^2}{\alpha^2 2^j} \le \frac{128}{\alpha^2}(1/\sqrt{2} + c\sqrt{20})^2 < 10{,}200 < 2^{14}.$$

Thus the implemented a_s in (39) pays for the factor-two selection loss and proves (12). For fixed s, let I_t indicate an informative assignment and put $Y_t = I_t \|x_{t,a_t}\|^2_{V_{s,t}^{-1}}$, with $I_t = Y_t = 0$ if none occurs. Let $\mathcal{G}_t$ contain the full past and the current cascade through layer s's test, but not its independent selector; if the round stops earlier, use its end-of-round history. These histories are nested and contain all earlier Y_i. Thus I_t is $\mathcal{G}_t$-measurable and

$$p_t := \mathbb{E}[Y_t \mid \mathcal{G}_t] \geq \kappa_s I_t, \qquad \kappa_s = 8D\varepsilon_s^2/a_s.$$

Since $N_s^{\inf} = \sum_t I_t$ and $\sum_t Y_t \leq 2D$, the tower property gives $\kappa_s \mathbb{E}N_s^{\inf} \leq \sum_t \mathbb{E}p_t = \mathbb{E}\sum_t Y_t \leq 2D$, hence

$$\mathbb{E}N_s^{\inf} \leq \frac{a_s}{4\varepsilon_s^2}. \tag{47}$$

Lemma 11 (Bounded potential controls the compensator in L_2)**.** *Let $0 \leq Y_t \leq 1$ satisfy $\sum_t Y_t \leq B$ pathwise, and let $A = \sum_t \mathbb{E}(Y_t \mid \mathcal{G}_t)$. The pre-observation sigma-fields are nested, with Y_t measurable with respect to $\mathcal{G}_{t+1}$. Then $\|A\|_{L_2} \leq B + \sqrt{B}$.*

Proof. Put $p_t = \mathbb{E}[Y_t \mid \mathcal{G}_t]$ and $\Delta_t = Y_t - p_t$. Then $\mathbb{E}[\Delta_t \mid \mathcal{G}_t] = 0$, and for $i < t$, Δ_i is $\mathcal{G}_t$-measurable, so $\mathbb{E}[\Delta_i \Delta_t] = 0$. Thus, for $M = \sum_t \Delta_t$,

$$\mathbb{E}M^2 = \sum_t \mathbb{E}\Delta_t^2 = \sum_t (\mathbb{E}Y_t^2 - \mathbb{E}p_t^2) \leq \mathbb{E}\sum_t Y_t \leq B,$$

using $Y_t^2 \leq Y_t$. Since $A = \sum_t Y_t - M$, Minkowski gives $\|A\|_{L_2} \leq B + \sqrt{B}$. □

For the Y_t, p_t, κ_s above, $0 \leq Y_t \leq 1$, $B = 2D$, and $A = \sum_t p_t \geq \kappa_s N_s^{\inf}$ on every path. Therefore

$$\|N_s^{\inf}\|_{L_2} \leq \frac{2D + \sqrt{2D}}{\kappa_s} = \frac{a_s}{4\varepsilon_s^2}\left(1 + \frac{1}{\sqrt{2D}}\right) \leq \frac{a_s}{2\varepsilon_s^2},$$

since $D \geq \log 2 > 1/2$. In particular,

$$\|N_s^{\inf}\|_{L_2} = \mathcal{O}\left(\frac{a_s}{\varepsilon_s^2}\right). \tag{48}$$

This is a compensator argument, not a deterministic lower bound on the information in every realized exploratory action.

G.6 Gaussian-triggered assignments and prediction excess

Condition immediately before a test probe. Write $u = \max_{a \in A} \|x_a\|_{\bar{V}_s^{-1}}$ and let M be the Gaussian maximum in (6). Suppose $\rho_s^\circ < \varepsilon_s/2$. Since $\rho_s^\circ \geq 2c\sqrt{\ell_s}u$ and $\rho_s = \rho_s^\circ + c(M - W)$,

$$u < \frac{\varepsilon_s}{4c\sqrt{\ell_s}}, \qquad \{\rho_s > \varepsilon_s\} \subseteq \{M - W > \varepsilon_s/(2c)\}.$$

For $u > 0$, (17) gives, under this pre-test conditioning,

$$\mathbb{P}\{M - W > \varepsilon_s/(2c)\} \leq e^{-\varepsilon_s^2/(8c^2u^2)} \leq e^{-2\ell_s}.$$

If $u = 0$, no assignment is possible. Write $I_{s,t}^{\mathrm{err}}$ for the noninformative-assignment indicator, zero if not reached. Averaging the conditional bound gives $\|I_{s,t}^{\mathrm{err}}\|_{L_2} = \sqrt{\mathbb{P}(I_{s,t}^{\mathrm{err}} = 1)} \leq e^{-\ell_s}$. The verified parameter inequalities then give

$$\|4\varepsilon_{s-1} I_{s,t}^{\mathrm{err}}\|_{L_2} \leq 4\varepsilon_{s-1} e^{-\ell_s} \leq 4\tau_s e^{-\ell_s/2} \leq 4\tau_s. \tag{49}$$

Thus the sum over at most H rounds is bounded in L_2 by $4H\sum_s \tau_s \leq 4\Phi$. For read excess, the proof of Lemma 2 in Section F gives $\|\xi_{s,t}\|_{L_2} \leq \tau_s$. Minkowski therefore yields

$$\left\| \sum_{t,s \geq 1} \xi_{s,t} \right\|_{L_2} \leq \sum_{t,s \geq 1} \|\xi_{s,t}\|_{L_2} \leq H \sum_{s \geq 1} \tau_s \leq \Phi. \tag{50}$$

Neither argument assumes independence across queries, snapshots, rounds, or layers. Equation (23) and $\|2HD_H\|_{L_2} \leq 2\sqrt{2}/H$ control the separate initial-layer terms.

G.7 Completion of the fixed-bound regret proof

The initial-layer calculation above bounds its total nominal cost by a_0/ε_0 pathwise. By (47), informative assignments have expected nominal cost at most $\sum_{s\ge 1} a_s\varepsilon_{s-1}/\varepsilon_s^2$. By (48) the L_2 cost is at most a universal constant times the same expression. Terminal cost is at most $4H\varepsilon_{L-1}$. On the uncapped branch, the accounting objective, not the random costs, equals Φ. The explicit bound preceding (48) therefore makes the combined nominal L_2 cost at most 2Φ. The Gaussian-triggered costs add at most 4Φ; the later-layer excess adds at most 4Φ, including its coefficient four in (24). The two root excess terms add at most $6\sqrt{2}/H$ by (23) and $\|2HD_H\|_{L_2} \le 2\sqrt{2}/H$. Minkowski thus gives $\|R\|_{L_2} \le 10\Phi + 6\sqrt{2}/H$, proving

$$\|R\|_{L_2} = \mathcal{O}(\Phi + H^{-1}) = \mathcal{O}\Big(\sqrt{Hd\log(1+H/d)\min\{d,\log K\}}\Big). \tag{51}$$

The last inequality uses $\min\{d,\log K\} \ge \log 2$, (44), and $Hd\log(1+H/d)\min\{d,\log K\} \ge (\log 2)^2$ to absorb the constant. On the capped branch, $R \le 2H < \Phi/2$ directly proves the same statistical bound. Taking $H = T$ proves the moment statement of Theorem 1. The proof of (51) also applies to a subinterval with at most H rounds and causally generated retained data of total global length at most H: all potentials are incremental, the regression law is the full-prefix law, and every queried snapshot has at most H observations. It is not an arbitrary-initial-state regret assertion.

G.8 Pathwise counting

Every ordinary refresh of a fixed layer consumes at least z_s new units of its log determinant. A pool-only endpoint consumes at least z_0 new units of the pooled log determinant. Copies at a different trigger can discard unspent increments but never reuse spent increments. Therefore

$$S_T \le \sum_{s<L} \frac{\log\det V_{s,T+1}}{z_s} + \frac{\log\det V_{\mathrm{all},T+1}}{z_0} \le d\log(1+T/d)\left(\sum_{s<L} z_s^{-1} + z_0^{-1}\right).$$

This upper bound may count concurrent requests more than once, whereas the controller counts their common endpoint only once.

Put $x = \log(2+\min\{d,\log K\})$. Since $\min\{d,\log K\} \ge \log 2$ implies $(1+\min\{d,\log K\})^2 \ge 2+\min\{d,\log K\}$, $z_0 \ge \log(1+\min\{d,\log K\}) \ge x/2$. For $j = L-s \ge 1$, the higher-layer threshold is $x + 2^{(j-5)/2}$. The summand is decreasing, so comparison with an integral bounds all possible higher layers at once:

$$\begin{aligned}\sum_{j=1}^{\infty} \frac{1}{x+2^{(j-5)/2}} &\le \frac{1}{x+1/4} + \int_1^{\infty} \frac{dt}{x+2^{(t-5)/2}} \\ &\le \frac{1}{x} + \frac{2}{x\log 2}\log(1+4x) \\ &\le \frac{1}{x} + \frac{4}{x\log 2}\log(2+x).\end{aligned}$$

The integral follows by substituting $y = 2^{(t-5)/2}$; the last step uses $1+4x \le (2+x)^2$. Including the root adds at most $2/x$; including the pool adds another $2/x$. Since $\log(2+x) \ge \log 2$,

$$\sum_{s<L} z_s^{-1} \le \frac{3}{x} + \frac{4\log(2+x)}{x\log 2} \le \frac{7}{\log 2}\frac{\log(2+x)}{x} < 14\frac{\log(2+x)}{x},$$

$$\sum_{s<L} z_s^{-1} + z_0^{-1} \le \frac{5}{x} + \frac{4\log(2+x)}{x\log 2} \le \frac{9}{\log 2}\frac{\log(2+x)}{x} < 16\frac{\log(2+x)}{x}.$$

Consequently, the thresholds satisfy the following bounds with logarithmic regularizer 2:

$$\begin{aligned}\sum_{s<L} z_s^{-1} &\le 14\frac{\log(2+\log(2+\min\{d,\log K\}))}{\log(2+\min\{d,\log K\})}, \\ \sum_{s<L} z_s^{-1} + z_0^{-1} &\le 16\frac{\log(2+\log(2+\min\{d,\log K\}))}{\log(2+\min\{d,\log K\})}.\end{aligned} \tag{52}$$

This proves (13), including every pool-only endpoint. For $d \geq 3$, $K \geq 16$, one has $\min\{d, \log K\} \geq \log 16 > e$; continuity on bounded intervals and the ratio limit at infinity show

$$\log(2 + \log(2 + \min\{d, \log K\})) = \Theta\big(\log\log(\min\{d, \log K\})\big).$$

This last equivalence is used only inside asymptotic bounds; the displayed numerical inequalities retain their regularized logarithms.

H Horizon-Free Analysis with Retained Statistics

We first give the controller and endpoint conventions, then prove both guarantees for that single process. The estimator regularization remains one, and observations are retained rather than replaced by independent epoch restarts.

H.1 Horizon-free wrapper and endpoint conventions

Fix d, K and set

$$H_0 = \max\{2, \lceil \log(2 + \min\{d, \log K\}) \rceil\}. \tag{53}$$

The only additional scheduler state is the current integer global bound H. The parameter formula at that bound is exactly (37)–(43); it is not a separately tuned family.

Algorithm 2 GATE (horizon-free)

Input: feature dimension d, upper bound K on actions per round.

1: $H \leftarrow \max\{2, \lceil \log(2 + \min\{d, \log K\}) \rceil\}$; $(L, \beta_0, (\varepsilon_s, z_s)_{s<L}, (\ell_s)_{1\leq s<L}) \leftarrow \mathrm{Parameters}(H, d, K)$.
2: $(V_s, b_s) = (\bar{V}_s, \bar{b}_s) = (I_d, 0)$ $(s = 0, \ldots, L-1, \mathrm{all})$; $Q \leftarrow \varnothing$.
3: **for** $t = 1, 2, \ldots$ **do**
4: $\quad \mathcal{R}_t \leftarrow \{s < L : \log\det V_s - \log\det \bar{V}_s \geq z_s\}$; $P_t \leftarrow \mathbf{1}\{\log\det V_{\mathrm{all}} - \log\det \bar{V}_{\mathrm{all}} \geq z_0\}$.
5: $\quad$ **if** $t = H$ or $\mathcal{R}_t \neq \varnothing$ or $P_t = 1$ **then**
6: $\quad\quad$ receive $(r_i)_{i\in Q}$; $b_s \leftarrow b_s + \sum_{i\in Q: J_i = s} x_{i,a_i} r_i$ $(s < L)$; $b_{\mathrm{all}} \leftarrow b_{\mathrm{all}} + \sum_{i\in Q} x_{i,a_i} r_i$.
7: $\quad\quad$ **if** $t = H$ **then** $L_{\mathrm{old}} \leftarrow L$; $H \leftarrow 2H$; $(L, \beta_0, (\varepsilon_s, z_s)_{s<L}, (\ell_s)_{1\leq s<L}) \leftarrow \mathrm{Parameters}(H, d, K)$.
8: $\quad\quad$ $(V_s, b_s) \leftarrow (I_d, 0)$ $(L_{\mathrm{old}} \leq s < L)$; $\mathcal{R}_t \leftarrow \{0, \ldots, L-1\}$.
9: $\quad\quad$ $(\bar{V}_s, \bar{b}_s) \leftarrow (V_s, b_s)$ $(s \in \mathcal{R}_t)$; $(\bar{V}_{\mathrm{all}}, \bar{b}_{\mathrm{all}}) \leftarrow (V_{\mathrm{all}}, b_{\mathrm{all}})$; $Q \leftarrow \varnothing$.
10: $\quad$ execute Algorithm 1, Lines 8–24, with the current parameters and state.
11: on external stop: repeat Line 6; $Q \leftarrow \varnothing$; no policy update.

Reading Algorithm 2 line by line. **Line 1** sets the initial public bound and then computes its parameters without a final horizon. **Line 2** initializes the statistics and buffer. **Lines 3–6** combine public, layer, and pool requests into one pre-context feedback query and aggregate every buffered response once.

Line 7 tests $t = H$ before changing the bound; only then does it save the old depth, double the global bound, and recompute parameters at the new bound. **Line 8** initializes only new bottom layers and requests all current layers on that same public-update branch. No old statistic or assignment label is reset, discarded, or relabeled; new layers receive no old observations, including old terminal data. **Line 9** copies the requested layers and pool once. Thus a public copy absorbs concurrent ordinary requests, while an ordinary pool-only request copies no layer.

Line 10 executes only the indicated observe–select–play–write block of Algorithm 1, at the current global round with the current parameters; it does not execute its initialization, finite-horizon loop, or terminal feedback. **Line 11** is the termination handler after an externally stopped action and its design writes; it deploys no new policy. A pre-context update is counted only if that round occurs. The first public interval is $1, \ldots, H_0 - 1$; later intervals at bound H are $H/2, \ldots, H - 1$, or a prefix. Hence H bounds the global sample count, not just the current interval's length.

H.2 Moment bound with retained data

Lemma 7 applies to the entire prefix generated by Algorithm 2. Public times and parameter changes do not depend on replaced noises, and existing layer labels are preserved. For a later-layer snapshot,

let X stack its recorded training features, including earlier epochs, and write $V = \bar{V}_s = I + X^\top X$. Its stored response sum is $X^\top(X\theta_\star + \eta)$, so

$$\widehat{\theta} - \theta_\star = -V^{-1}\theta_\star + V^{-1}X^\top\eta.$$

The vector η retains its product law conditional on the restricted pre-test record of Section D, not on the full epoch-start history. For layer 0, Section C.2 instead controls all global prefixes through H via D_H. Thus retention introduces no extra warm-start term. The accumulated potential spent inside an epoch is

$$\sum_{t \text{ in epoch: } J_t = s} \|x_t\|^2_{V_{s,t}^{-1}} \le 2(\log\det V_{s,\mathrm{end}} - \log\det V_{s,\mathrm{start}}) \le 2d\log(1 + H/d).$$

Here $x_t = x_{t,a_t}$; the increments telescope and $V_{s,\mathrm{start}} \succeq I$, while at most H observations enter $V_{s,\mathrm{end}}$. After initialization or a deterministic epoch copy $V_s = \bar{V}_s$, and subsequently (35) holds with the current epoch's fixed threshold. Consequently (51) bounds the L_2 regret of that epoch, or any deterministic prefix of it, by $C\sqrt{Hd\log(1 + H/d)\min\{d, \log K\}}$ without restarting the process or conditioning on its initial state.

For $T \ge H_0$, let H_{last} be the last active parameter bound. Its epoch starts at $H_{\mathrm{last}}/2$, so $H_{\mathrm{last}} \le 2T$. Also $d\log(1+H_{\mathrm{last}}/d) \le 2d\log(1+T/d)$ since $1+2u \le (1+u)^2$. Each bound with $H = 2^{-j}H_{\mathrm{last}}$ is at most $2^{-j/2}$ times the last bound, by monotonicity of the logarithm. Minkowski, which requires no independence between epochs, therefore gives

$$\begin{aligned}
\|R_T\|_{L_2} &\le C \sum_{\substack{H \le H_{\mathrm{last}} \\ H/H_0 \text{ a power of } 2}} \sqrt{dH\log(1 + H/d)\min\{d, \log K\}} \\
&\le \frac{C\sqrt{dH_{\mathrm{last}}\log(1 + H_{\mathrm{last}}/d)\min\{d, \log K\}}}{1 - 2^{-1/2}} \\
&\le C'\sqrt{dT\log(1 + T/d)\min\{d, \log K\}}.
\end{aligned}$$

For $1 \le T < H_0$, note that $T \le \max\{1, \log(2 + \min\{d, \log K\})\} \le 2\min\{d, \log K\}$ (using $\min\{d, \log K\} \ge \log 2$). As $d \ge \min\{d, \log K\}$, $T/d \le 2$. For $u \ge 0$, $\log(1 + u) = \int_0^u (1 + v)^{-1}\,dv \ge u/(1 + u)$, so $d\log(1 + T/d) \ge T/(1 + T/d) \ge T/3$. Therefore $\sqrt{Td\log(1 + T/d)\min\{d, \log K\}} \ge T\sqrt{(\log 2)/3}$ and the pathwise ceiling $R_T \le 2T$ proves the same order. This early-prefix argument does not assert that ordinary determinant triggers are absent. Combining the epoch bound with this short-prefix argument proves the moment part of Theorem 1 at every fixed T. It does not assert a moment inequality at arbitrary data-dependent stopping times with a random horizon substituted into the right-hand side.

H.3 Summing depth blocks instead of restarting budgets

The depth is nondecreasing in H by (37). Group consecutive epochs of the stopped run with equal depth p into a constant-depth block. For a completed block, let E_p be its last global parameter bound H; the next block starts immediately before round E_p, so $E_p \le T$. For the current block set $E_p = T$, which is not necessarily the current parameter bound. Skipped depths have no block.

Lemma 12 (Depth-block summability)**.** *For the schedule in Algorithm 2,*

$$\sum_{\textit{constant-depth blocks}} \log(1 + E_p/d) \le 4\log(1 + T/d). \tag{54}$$

Proof. For $w = \log(1 + H/d)$ the depth is

$$L(H) = \max\left\{2, \left\lceil \log_2\{1 + \tfrac{1}{2}\log d + \tfrac{1}{2}\log(1 + w)\}\right\rceil\right\}.$$

Set

$$b_p = \exp(2^{p+1} - 2)/d - 1 \qquad (p \ge 1).$$

For $p \ge 3$, $L(H) = p$ means $2^{p-1} < 1 + \frac{1}{2}\log[d(1 + w)] \le 2^p$; rearranging and exponentiating gives $b_{p-1} < w \le b_p$. For the clamped depth $p = 2$, only $w \le b_2$ is needed. Whenever $b_p > 0$ and $p \ge 2$,

$$\frac{b_p}{b_{p+1}} = \frac{\exp(2^{p+1} - 2)/d - 1}{\exp(2^{p+1})\exp(2^{p+1} - 2)/d - 1} \le e^{-2^{p+1}} \le e^{-8}.$$

Thus $(b_p)_+ \le e^{-8}(b_{p+1})_+$ also when $b_p \le 0$; iteration gives the geometric-series bound below. If there is no completed block, the conclusion is immediate. If the last completed depth is $p_\star = 2$, there are no earlier completed blocks, so the last completed and current contributions suffice. Otherwise $p_\star \ge 3$, and

$$b_{p_\star - 1} < \log(1 + E_{p_\star}/d) \le \log(1 + T/d).$$

Earlier completed blocks have depths at most $p_\star - 1$; their sum is at most

$$\sum_{2 \le p < p_\star} (b_p)_+ \le \frac{(b_{p_\star - 1})_+}{1 - e^{-8}} \le 2 \log(1 + T/d).$$

If that threshold is nonpositive, there are no earlier completed blocks. The last completed and current blocks each contribute at most $\log(1 + T/d)$. Skipping a depth only removes summands. This proves (54). □

Within a constant-depth block, every z_s and the pool threshold z_0 are identical across epochs: they depend on d, K and depth, not on β_0, the resolutions, or other changing coefficients. Each block begins with initialization or a public full copy. An ordinary refresh of layer s consumes at least $z_s(p)$ new units of its log determinant since its last copy. These charged segments are disjoint and lie within the block. Other copies may discard unspent growth but cannot recreate spent growth. For pool-only endpoints, use the segments since the last pool copy, including copies at layer requests. Every ordinary endpoint either has a layer request or is pool-only; counting concurrent requests separately only enlarges the bound. Thus its count in this block is at most

$$\sum_{s<p} \frac{\log\det V_{s,\mathrm{end}} - \log\det V_{s,\mathrm{start}}}{z_s(p)} + \frac{\log\det V_{\mathrm{all,end}} - \log\det V_{\mathrm{all,start}}}{z_0(p)}$$
$$\le 16 d \log(1 + E_p/d) \cdot \frac{\log(2 + \log(2 + \min\{d, \log K\}))}{\log(2 + \min\{d, \log K\})}.$$

The last inequality uses (52) and the trace–determinant bound: starting log determinants are nonnegative, and ending ones are at most $d \log(1 + E_p/d)$. There are fewer than E_p observations at a completed block endpoint, and at most T at the current one. Summing and applying Lemma 12 gives at most $64 d \log(1 + T/d) \log(2 + \log(2 + \min\{d, \log K\}))/\log(2 + \min\{d, \log K\})$ ordinary updates.

H.4 Deterministic epoch updates, including short horizons

There are no deterministic epoch updates for $T < H_0$. Otherwise their number is $N_{\mathrm{pub}} = 1 + \lfloor \log_2(T/H_0) \rfloor$; the update before round H_0 is counted. For $x \ge 1$, $1 + \log_2 x \le (2/\log 2)\log(1+x)$ because $2x \le (1+x)^2$. Since $H_0 \ge \log(2 + \min\{d, \log K\})$,

$$N_{\mathrm{pub}} \le \frac{2}{\log 2} \log\left(1 + \frac{T}{\log(2 + \min\{d, \log K\})}\right).$$

We have $\log(2 + \min\{d, \log K\}) \le 2\min\{d, \log K\} \le 2d$. For fixed T and $x > 0$, $f(x) = x \log(1 + T/x)$ satisfies $f'(x) = \log(1 + T/x) - T/(x + T) \ge 0$, by $\log(1 + u) \ge u/(1 + u)$. Applying this monotonicity at $x = \log(2 + \min\{d, \log K\})/2 \le d$ gives

$$d \log(1 + T/d) \ge \frac{\log(2 + \min\{d, \log K\})}{2} \log\left(1 + \frac{2T}{\log(2 + \min\{d, \log K\})}\right)$$
$$\ge \frac{\log(2 + \min\{d, \log K\})}{2} \log\left(1 + \frac{T}{\log(2 + \min\{d, \log K\})}\right).$$

Combining the inequalities yields

$$N_{\mathrm{pub}} \le \frac{4}{\log 2} \frac{d \log(1 + T/d)}{\log(2 + \min\{d, \log K\})}$$
$$< 6 d \log(1 + T/d) \cdot \frac{\log(2 + \log(2 + \min\{d, \log K\}))}{\log(2 + \min\{d, \log K\})}.$$

Here $\log(2+\log(2+\min\{d,\log K\})) > 1$ because $\min\{d,\log K\} \geq \log 2$, so the numerical coefficient 6 is valid. Adding this to the ordinary-update count proves the pathwise bound

$$S_T \leq 70d\log(1+T/d)\cdot\frac{\log(2+\log(2+\min\{d,\log K\}))}{\log(2+\min\{d,\log K\})},$$

which establishes (2). These counting inequalities hold for every realized finite prefix; a data-dependent stopping rule merely selects one such prefix. The starting offset ensures the same counting order even at short horizons.

H.5 Horizon-blindness and computational cost

For fixed (d,K) and one infinite context/noise/randomness sequence, Algorithm 2 defines one process. Executing it to $T_2 > T_1$ and restricting to the first T_1 actions is identical to executing it with an external stop after T_1; only the uncounted terminal feedback differs. The kernel uses no time argument; time belongs solely to the monitor that authorizes deterministic epoch copies. Monotonicity of $L(H)$ and $H \leq 2T$ after the initial interval gives the same layer-storage order as the known-horizon configuration. Initially $H_0 = \mathcal{O}(1+\log(2+d))$, so its depth has the same order even for a short prefix. Parameter recomputation and copying cost $\mathcal{O}(Ld^2)$ at a public time and do not change the stated amortized arithmetic order.

I Sequential Statistical Lower Bound

For the sequential comparison, fix $\delta > 0$ and integers $d \geq 2$, $K \geq 2$, and $T \geq d^{2+\delta}$ throughout this proof.

I.1 Minimax classes and the use of a prior

Let $\mathcal{I}(T,d,K)$ be the deterministic instances satisfying Assumption 1. A fully sequential learner knows T, d, K, sees only the current and past contexts, and receives each selected reward immediately. For the more restricted feedback class, let

$$\mathfrak{R}_{\mathrm{fb}}(T,d,K,S) = \inf_{\pi\in\Pi_{\mathrm{fb}}:\,\sup_{\nu\in\mathcal{I}}\mathbb{E}_\nu Q_T^\pi \leq S}\ \sup_{\nu\in\mathcal{I}}\mathbb{E}_\nu R_T^\pi,$$

where Π_{fb} permits time- and context-history-dependent actions between queries. Fully sequential learning has no larger minimax regret. Restricting Π_{fb} to fixed within-batch policies can only increase its minimax regret.

Our lower bounds choose a finite random parameter and a complete context sequence *before* interaction, independently of learner randomness and noise. The parameter and its precommitted contexts may be correlated; each realization is an admissible fixed instance. For any learner π and finite prior $\mathcal{P}$ on $\mathcal{I}$,

$$\sup_{\nu\in\mathcal{I}}\mathbb{E}_\nu R_T^\pi \geq \mathbb{E}_{\nu\sim\mathcal{P}}\mathbb{E}_\nu R_T^\pi.$$

A lower bound on the right that holds for every π therefore survives the infimum over sequential learners. Similarly, a uniform worst-case expected-query budget implies the same budget averaged over the prior. Thus no minimax interchange is needed. The constructions use independent $N(0,1)$ scalar noise, a subfamily of Assumption 1.

Relation to earlier lower bounds. Our construction builds on the hierarchical-testing idea of Li et al. (2024), but proves the lower half of Proposition 1 without invoking their lower-bound theorem. Its Section I-A imposes the same global unit-norm constraints on the parameter and features. Theorem 2 states an $\Omega_\epsilon(\sqrt{dT\log K\log(T/d)})$ lower bound for $K \leq 2^{d/2}$ and $T \geq d(\log_2 K)^{1+\epsilon}$, with fixed small $\epsilon > 0$; their action-count notation is n, here written as K. The proof of Theorem 2 in Section III-D applies their Theorem 21 on disjoint coordinate blocks in successive phases. Such a transfer must preserve one fixed, globally unit-norm parameter, not merely unit norms within individual phases.

For an even local dimension h, the parameter formula in their Section III-C1 gives $\vartheta_{2s}^{(b)} = 1/(2\sqrt{h})$ for $s \in [h/2]$. Unrescaled copies on B disjoint coordinate blocks therefore satisfy

$$\left\| \bigoplus_{b=1}^{B} \vartheta^{(b)} \right\|_2^2 = \sum_{b=1}^{B} \|\vartheta^{(b)}\|_2^2 \geq B\,\frac{h}{2}\left(\frac{1}{2\sqrt{h}}\right)^2 = \frac{B}{8}.$$

Thus this concatenation is not globally unit-norm for $B > 8$; zero-filling inactive context coordinates does not change its parameter norm. Dividing the parameter by $\sqrt{B}$ reduces every local mean gap by $1/\sqrt{B}$. Compensating with a $\sqrt{B}$ increase in the active features restores those means but requires renewed verification of the full feature norm and the admissible feature scales. Consequently, the phasewise horizon condition alone does not verify the global instance or preserve its gap calculation with dimension-independent constants. This observation concerns the displayed transfer, not the possibility of a different construction establishing a wider horizon range.

What the reconstruction establishes. We jointly budget the parameter norm across all m groups and the action norm across the r simultaneously active groups before the testing argument. The reciprocal scalings in (57) and (58) preserve local mean differences. The stage count and length in (55) then make the full-action norm bound (59) and the stage-independent gap (61) hold simultaneously. The resulting construction uses one fixed parameter and one scalar reward per round. Its history KL bound, (62), includes immediate feedback within the current stage and compares action kernels at the same realized history; it does not assume that their marginal action laws agree. The single choice of r covers both action regimes, including bounded K, and the integer choices and zero padding cover every stated d, K, T.

Role in the minimax characterization. Throughout this section, the standing assumptions remain fixed $\delta > 0$, integers $d, K \geq 2$, and $T \geq d^{2+\delta}$. Within this same range, (60) and (64) yield the lower bound (65), whose constant depends only on δ, uniformly in d, K, T. Their Theorem 21 already addresses the exponential-action endpoint under polynomial horizon separation; we do not claim that endpoint lower-bound order as new. Rather, Proposition 1 combines the new upper bound with a matching lower bound established here by a fully specified, globally normalized construction, without relying on the phasewise transfer. This sufficient comparison condition is not asserted to be necessary and does not restrict Theorem 1.

I.2 Construction for every $T \geq d^{2+\delta}$

Building on the hierarchical-testing idea of Li et al. (2024), we prove the lower bound needed for Proposition 1 by placing independent binary decisions at several feature scales. At each stage the current contexts reveal earlier signs, but not the sign currently being tested. Small scalar-reward KL keeps that sign hard to identify even with immediate feedback. All norms below are for the full parameter and full composite action.

For the fixed δ, d, K, T above, choose

$$r = \min\{\lfloor \log_2 K \rfloor, \lfloor d/2 \rfloor\}, \qquad m = r\left\lfloor \frac{d}{2r} \right\rfloor,$$
$$J = \max\{j \in \mathbb{N}_{\geq 1} : 2m^2 j 16^{j-1} \leq T\}, \qquad N = \left\lfloor \frac{Tr}{mJ} \right\rfloor. \tag{55}$$

Here r is the number of two-dimensional groups used simultaneously, m is their total number, J is the number of stages per block, and N is the number of rounds per stage. Since $r \geq 1$ and $\lfloor x \rfloor \geq x/2$ for $x \geq 1$,

$$\frac{d}{4} \leq m \leq \frac{d}{2}, \qquad r \mid m, \qquad 2^r \leq K, \qquad r \geq \frac{1}{3}\min\{d, \log K\}.$$

The last inequality follows from $\lfloor d/2 \rfloor \geq d/3$ and $\lfloor \log_2 K \rfloor \geq \log K/(2\log 2)$. The set defining J is nonempty because $2m^2 \leq d^2/2 \leq T$, and finite because $j16^{j-1}$ grows without bound. Its defining inequality also gives $Tr/(mJ) \geq 2mr16^{J-1} \geq 2$, hence

$$N \geq \frac{Tr}{2mJ}, \qquad \frac{m}{r}JN \leq T. \tag{56}$$

Split $2m$ coordinates into m two-dimensional groups and use zeroes in unused coordinates. Draw independent uniform signs $u_{g,j} \in \{-1, 1\}$ for $g \in [m]$, $j \in [J]$. Set

$$\psi_{g,j} = \frac{1}{4} + \frac{1}{32} \sum_{i=1}^{j} u_{g,i} 4^{-i}, \qquad z_j = \frac{4^{j-1}}{\sqrt{N}}, \qquad \theta_\star = \frac{1}{\sqrt{m}} \bigoplus_{g=1}^{m} (\psi_{g,J}, 1/2). \tag{57}$$

Here $\bigoplus$ concatenates the two-coordinate groups. Define $\psi_{g,0} = 1/4$. In this lower-bound construction, z_j is a local feature scale, not an algorithmic log-determinant threshold. The cancellation $z_j 4^{-j} = 1/(4\sqrt{N})$ will give stage-independent gaps, while J limits the largest feature norm. Partition the groups, independently of the signs, into m/r ordered blocks of r groups. For a block G, at stage j offer all 2^r actions indexed by $b \in \{1, 2\}^G$, with b_g selecting the local component:

$$x_{j,b}^{(g)} = \begin{cases} \sqrt{m}\,(z_j, 0), & g \in G,\ b_g = 1, \\ \sqrt{m}\,(0, 2 z_j \psi_{g,j-1}), & g \in G,\ b_g = 2, \\ (0, 0), & g \notin G. \end{cases} \tag{58}$$

Repeat that indexed set for N rounds. Fill the remaining $T - mJN/r$ rounds with two identical zero features. The full context sequence is a deterministic function of the signs sampled before interaction, not of learner actions or rewards.

I.3 Global norms and the number of stages

The geometric sum gives $|\psi_{g,j} - 1/4| \le (1/32) \sum_{i \ge 1} 4^{-i} = 1/96$. Consequently

$$\|\theta_\star\|_2^2 = \frac{1}{m} \sum_{g=1}^{m} (\psi_{g,J}^2 + 1/4) \le (25/96)^2 + 1/4 < 1, \qquad 2|\psi_{g,j}| < 1.$$

There are r active groups on disjoint coordinates; each contributes at most $m z_j^2$ to the squared action norm because $2|\psi_{g,j-1}| < 1$. Using (56) and the definition of J,

$$\|x_{j,b}\|_2^2 \le m r z_J^2 = \frac{m r 16^{J-1}}{N} \le \frac{2m^2 J 16^{J-1}}{T} \le 1. \tag{59}$$

This verifies the norm of the composite action, rather than merely checking each group separately.

Maximality of J implies

$$T < 2m^2 (J+1) 16^J \le \frac{d^2}{2} (J+1) 16^J.$$

For every integer $J \ge 1$, $J + 1 \le 2^J$, and hence $1 + (J+1)16^J/2 \le 1 + 32^J/2 \le 32^J$. It follows that

$$\log(1 + T/d^2) \le J \log 32. \tag{60}$$

Equation (60) gives a logarithmic lower bound on the number of testing stages throughout the assumed range $T \ge d^{2+\delta}$.

I.4 Mean gaps and immediate-feedback KL

Within an active group g, the action's $\sqrt{m}$ and the parameter's $1/\sqrt{m}$ cancel, giving mean contributions $z_j \psi_{g,J}$ and $z_j \psi_{g,j-1}$. Their difference is

$$\frac{z_j}{32} \left(u_{g,j} 4^{-j} + \sum_{i=j+1}^{J} u_{g,i} 4^{-i} \right).$$

Since $|\sum_{i=j+1}^{J} u_{g,i} 4^{-i}| \le 4^{-j}/3$, this difference has sign $u_{g,j}$: component 1 is optimal for $u_{g,j} = 1$, and component 2 for $u_{g,j} = -1$. The incorrect-component gap is at least

$$\frac{z_j 4^{-j}}{32} \left(1 - \sum_{i \ge 1} 4^{-i} \right) = \frac{1}{192\sqrt{N}}. \tag{61}$$

Write b_t for the selected component vector and let $e_{t,g}$ indicate an incorrect component. Since every combination is offered, the best action optimizes each group separately, so round regret is at least $(192\sqrt{N})^{-1}\sum_{g\in G} e_{t,g}$.

Fix (g,j) and all other signs, not a realized history. Write u for the array with $u_{g,j}=1$ and $u^{\oplus(g,j)}$ for its flip. All contexts through the end of stage j are identical: the stage-i contexts with $i\le j$ depend only on $\psi_{g,i-1}$. Previous blocks do not activate group g. Flipping $u_{g,j}$ changes a stage-i scalar reward mean by $2(1/32)4^{-j}z_i$ when component 1 of group g is chosen, and by zero otherwise. This is the full scalar reward mean shift, conditional on the chosen composite action and all other signs.

Let P_u^{t-1} and $P_{u^{\oplus(g,j)}}^{t-1}$ denote the laws of the complete immediate-feedback histories before action t of stage j, including current contexts and private randomization through that action. By Section A.6, each scalar observation contributes the expected squared conditional mean difference divided by two to history KL. Dropping the component-1 indicator and using at most N observations from each stage $i\le j$, including the elapsed part of stage j, give

$$\begin{aligned}\mathrm{KL}(P_u^{t-1}\|P_{u^{\oplus(g,j)}}^{t-1}) &\le 2(1/32)^2 N 4^{-2j}\sum_{i=1}^{j} z_i^2\\ &= 2(1/32)^2\sum_{i=1}^{j}16^{i-1-j} \le \frac{2}{32^2}\frac{1/16}{1-1/16}=\frac{1}{7680}.\end{aligned} \tag{62}$$

Action kernels agree at the same realized history and hence add no KL; their marginal action laws need not agree. Future contexts, which may reveal $u_{g,j}$, are unavailable here. For $E=\{b_{t,g}=1\}$, Section A.6 and (62) give

$$\frac{P_u^{t-1}(E^c)+P_{u^{\oplus(g,j)}}^{t-1}(E)}{2}\ge\frac{1-\sqrt{1/15360}}{2}\ge\frac{1}{4}.$$

Averaging over the other signs gives $\mathbb{E}e_{t,g}\ge 1/4$ under the original prior. Independence between component errors is not needed.

Summing (61) times the error-probability lower bound $1/4$ over $(m/r)JN$ rounds and r components per round gives the first inequality below. The next two use $N\ge Tr/(2mJ)$, then $mr\ge d\min\{d,\log K\}/12$ and $J\ge\log(1+T/d^2)/\log 32$:

$$\begin{aligned}\mathbb{E}R_T &\ge \frac{mJ\sqrt{N}}{768}\ge\frac{\sqrt{mrTJ}}{768\sqrt{2}}\\ &\ge\frac{1}{768\sqrt{24\log 32}}\sqrt{dT\min\{d,\log K\}\log(1+T/d^2)}.\end{aligned} \tag{63}$$

This is a lower bound for every randomized immediate-feedback learner averaged over a finite prior. Since every realization is admissible, (63) also lower-bounds $\mathfrak{R}_{\mathrm{seq}}(T,d,K)$ throughout the assumed range $T\ge d^{2+\delta}$.

I.5 Minimax optimality under the same horizon assumption

The standing assumption $T\ge d^{2+\delta}$ gives $\log d\le\delta^{-1}\log(T/d^2)$, and $1+T/d\le d(1+T/d^2)$. Hence

$$\begin{aligned}\log(1+T/d)&\le\log d+\log(1+T/d^2)\\ &\le(1+\delta^{-1})\log(1+T/d^2),\end{aligned} \tag{64}$$

whereas $\log(1+T/d^2)\le\log(1+T/d)$. Combining (64) with (63) and the finite-prior argument gives

$$\mathfrak{R}_{\mathrm{seq}}(T,d,K)\ge c_\delta\sqrt{dT\log(1+T/d)\min\{d,\log K\}}, \tag{65}$$

where $c_\delta=\sqrt{\delta/(1+\delta)}/(768\sqrt{24\log 32})>0$ depends only on δ. This proves the lower half of (3). For the upper half, a sequential learner can run `GATE` while ignoring rewards until its authorized queries. Theorem 1 applies to that learner. This proves matching for every $T\ge d^{2+\delta}$. The feedback comparison retains its own stronger conditions; neither (64) nor this statistical result removes them.

J FEEDBACK LOWER BOUND

The sequential lower bound limits statistical performance even with every reward available. Here the additional resource is feedback queries, in the larger class Π_{fb} defined in Section I. For fixed $\delta > 0$, the comparison holds for every integer horizon satisfying the combined condition in Theorem 2. The hierarchy below gives the dependence on the query budget. An independent-coordinate calculation supplies the linear-in-dimension bound that yields the numerical constant 16 in the horizon condition of Theorem 2 and covers the horizons below the hierarchy's applicability threshold. Stages and blocks below are deterministic construction intervals, not learner-selected batches. We use the finite-prior reduction and binary-testing inequalities in Sections A.6 and I.

Relation to Ruan et al. (2021). Our hidden-sign hierarchy builds on their multiscale policy-switch construction. The refinement is its calibration and analysis at a constant-factor minimax regret target. Theorem 8 of their full arXiv version gives the expected-regret lower bound

$$B_{\mathrm{R}}(M) = \frac{\sqrt{dT}}{32d}\left(\frac{2T}{d}\right)^{1/(16M/d+2)}$$

for $K = 2$, even $d \geq 2$, T larger than a sufficiently large constant times d, and expected policy-switch budget $20d \leq M \leq d\log T/48$. Let $q = \min\{d, \log K\}$. Even at the level of this displayed expression, imposing $B_{\mathrm{R}}(M) \leq C_{\mathrm{reg}}\sqrt{dTq\log(1+T/d)}$ gives

$$\frac{\log(2T/d)}{16M/d+2} \leq \log\Big(32C_{\mathrm{reg}}d\sqrt{q\log(1+T/d)}\Big).$$

The right side retains $\log d$ and $\frac{1}{2}\log\log(1+T/d)$; direct inversion of the stated bound therefore does not give (4) uniformly. This compares the regret–budget expressions, not the two resource models.

Our feature scale $z_j = v^{-(j+1)}/\sqrt{N}$ is calibrated to the stage length N, not the full local horizon NJ. This gives $\Delta \gtrsim (v\sqrt{N})^{-1}$ while keeping the pre-stage history KL bounded by a constant. Joint scaling in Lemma 14 preserves the full parameter and context unit-norm constraints with one scalar reward per round. For $r = 1$, $m = \lfloor d/2 \rfloor$, and $N = \lfloor T/(mJ) \rfloor$, aggregation retains $mJ\sqrt{N} \asymp \sqrt{dTJ}$. Choosing $J \asymp 1 + S/d$ yields the prefactor $\sqrt{dT(1+S/d)}$ in Proposition 3. In that proposition's range, $T \geq d^{2+\delta}$ gives $\log(1+T/d) = \mathcal{O}_\delta(n_\star)$, where $n_\star = \log(T/d^2)$. Writing $x = n_\star/(1+S/d)$, comparison at the same regret target gives

$$\frac{e^{c_2 x}}{\sqrt{x}} \lesssim_{\delta, C_{\mathrm{reg}}} \sqrt{q} \quad \Longrightarrow \quad x \lesssim_{\delta, C_{\mathrm{reg}}} \log(2+q).$$

Discarding $\sqrt{1+S/d}$ would instead leave $e^{c_2 x} \lesssim_{\delta, C_{\mathrm{reg}}} \sqrt{n_\star q}$, reintroducing $\log n_\star$ into the denominator after inversion. The binary construction suffices for every $K \geq 2$; q enters through the regret target, not through a need for more actions in the hard instance.

The feedback protocol also requires a separate testing argument. Their rare-switch model permits immediate rewards but fixes the current-context policy between switches; our class Π_{fb} allows time- and context-history-dependent actions without new reward access. Reward suppression in Lemma 13 gives a roundwise comparison by suppressing delivery at the first internal query, without fixing a within-batch kernel or conditioning on a future no-query event. Lemma 15 supplies $S \geq d/2$ under Theorem 2's conditions, absorbs the additive $-d$ in (73), and covers $n_\star < 8$. Together with the upper bound, this yields $\Theta(d\log T)$ feedback complexity for bounded K at the same regret level in the stated range.

J.1 REWARD-SUPPRESSED STAGE TESTING

For KL only, enlarge the pre-stage history to include all past contexts, actions, consumed private randomness, and every generated reward, including unobserved rewards. This enlargement does not give the learner additional information. Its actual state and any feedback query before the stage's first action are measurable from the enlarged history.

Lemma 13 (Stage test for general batched feedback)**.** *Consider J stages of N rounds, each repeating an indexed set encoding r binary choices. Conditional on all other hidden signs and any side information independent of the sign array, each current bit is uniform; flipping it leaves the stage set unchanged, reverses its optimal choice, and gives enlarged pre-stage history KL at most $1/4$. Suppose per-round regret is at least Δ times the number of wrong coordinates. Then every general batched-feedback learner satisfies*

$$\mathbb{E}R_{NJ} \geq rN\Delta\,[J/4 - \bar{s}]_+, \tag{66}$$

where $\bar{s} = \mathbb{E}Q_{\mathrm{block}}$ and Q_{block} counts feedback boundaries internal to this block; expectation also averages over the prior. If the enlarged pre-stage history KL is at most $1/8$, the same conclusion holds with $J/4$ replaced by $3J/8$.

Proof. Fix a stage j, a coordinate, the other signs, and any independent side information, but not the realized pre-stage history. Let H_j be the enlarged pre-stage history, U_j independent stage seeds, and P_+, P_- their joint laws under the two signs; appending the same independent seed law does not change KL. The current stage set is identical under both signs. Start from the learner's actual state determined by H_j and use U_j. Simulate its reward-free transitions until its first request strictly between that stage's first and last actions. At the request, suppress reward delivery and choose a fixed indexed action for the remaining stage rounds. Thus each comparison bit is a common function of (H_j, U_j) under both signs, although the laws of H_j may differ. No stage reward is used. For E the event that this function chooses the $+$-optimal bit, data processing and Section A.6 give

$$\frac{P_+(E^c) + P_-(E)}{2} \geq \frac{1 - \sqrt{\mathrm{KL}(P_+\|P_-)/2}}{2} \geq \frac{1}{4}.$$

Averaging over the fixed signs and side information gives comparison error at least $1/4$ at every coordinate and round.

Let F_j denote no internal real feedback request in stage j. Couple the real and comparison processes under each instance. On F_j they coincide. For any coordinate and stage round,

$$\mathbf{1}\{\text{real wrong}\} \geq \mathbf{1}\{\text{comparison wrong}\} - \mathbf{1}\{F_j^c\}.$$

Writing R_j for regret in stage j, averaging gives

$$\mathbb{E}R_j \geq rN\Delta\big(1/4 - \mathbb{P}(F_j^c)\big), \qquad \sum_j \mathbf{1}\{F_j^c\} \leq Q_{\mathrm{block}}. \tag{67}$$

The indicator inequality is pathwise: on F_j the actions agree, and on F_j^c its right side is nonpositive. Thus no independence between F_j and the current signs is required. The first request is determined before its requested rewards become available, so its simulation uses no hidden stage reward. Previously buffered rewards are already part of the enlarged pre-stage history in the KL comparison. Assign each queried stage one of its internal queries. These queries are distinct and all counted by Q_{block}, even if a query also delivers older rewards. Sum the stage inequalities, use $\sum_j \mathbb{P}(F_j^c) \leq \mathbb{E}Q_{\mathrm{block}} = \bar{s}$, and use nonnegative regret when the resulting right side is negative. For the stronger assertion, the KL bound $1/8$ improves the comparison-error lower bound at every coordinate and round to

$$\frac{1 - \sqrt{(1/8)/2}}{2} = \frac{3}{8}.$$

Replace $1/4$ by $3/8$ in (67) and repeat the same summation. This changes neither the feedback protocol nor the coupling argument. □

J.2 An amplified hierarchy across coordinate groups

Let $N, J \geq 1$ be integers, $0 < \rho_x \leq 1$ the local context radius, and $\zeta = 1/8$. Put $v = (N\rho_x^2)^{-1/(2(J+1))}$ and assume $v \leq 1/4$. Draw all independent uniform signs $u_i \in \{-1, 1\}$ before interaction; they determine the fixed parameter and complete context sequence. Put $\psi_0 = 1/4$ and

$$\psi_j = \tfrac{1}{4} + \zeta \sum_{i=1}^{j} u_i v^i, \quad z_j = v^{-(j+1)}/\sqrt{N}, \quad \theta = (\psi_J, 1/2).$$

At stage j the actions are $(z_j, 0)$ and $(0, 2z_j\psi_{j-1})$. Since $|\psi_j - 1/4| \le \zeta v/(1-v) \le 1/24$, we have $\|\theta\|_2^2 \le (7/24)^2 + 1/4 < 1$ and $2|\psi_{j-1}| < 1$. Also $v^{-(J+1)} = \sqrt{N}\rho_x$, so both context norms are at most $z_J = \rho_x$. The difference of their means is $z_j(\psi_J - \psi_{j-1}) = \zeta z_j v^j(u_j + \sum_{k=1}^{J-j} u_{j+k}v^k)$. The tail sum has absolute value at most $v/(1-v) \le 1/3$, so action 1 is optimal for $u_j = 1$ and action 2 for $u_j = -1$. Since $z_j v^j = 1/(v\sqrt{N})$,

$$\Delta \ge \zeta z_j v^j \left(1 - \frac{v}{1-v}\right) \ge \frac{2\zeta}{3v\sqrt{N}} = \frac{1}{12v\sqrt{N}}. \tag{68}$$

Fix all signs except u_j. Contexts through stage j depend only on earlier signs, so flipping u_j leaves them unchanged and changes ψ_J by $2\zeta v^j$. A stage-$i < j$ scalar mean therefore shifts by $2\zeta z_i v^j$ for action 1 and zero for action 2. At the same realized history, the learner's action and query rules agree; only the reward law contributes KL. By Section A.6, each observation contributes at most $(2\zeta z_i v^j)^2/2$. With N observations per earlier stage, the enlarged pre-stage KL is at most

$$2\zeta^2 N \sum_{i<j} z_i^2 v^{2j} = 2\zeta^2 \sum_{i<j} v^{2(j-i-1)} \le \frac{2\zeta^2}{1-v^2} \le \frac{1}{30}. \tag{69}$$

Here $\sum_{i<j} v^{2(j-i-1)} = \sum_{h=0}^{j-2}(v^2)^h$, with an empty sum for $j = 1$. Appending the same seed law, independent of the complete history under both signs, leaves KL unchanged. Including all N stage-j rewards would add $2\zeta^2/v^2$ to this upper bound; suppressing those rewards is therefore essential to the stage test.

Lemma 14 (Unit-norm block lower bound)**.** *Suppose integers $m, r, J \ge 1$ and a budget $S \ge 0$ satisfy $r \mid m$, $2m \le d$, $2^r \le K$, $J \ge \max\{1, \lceil 4Sr/m \rceil\}$, and $3T/(4m^2J) \ge 16^{J+1}$. Then*

$$\mathfrak{R}_{\mathrm{fb}}(T, d, K, S) \ge \frac{\sqrt{3}}{192}\sqrt{mrTJ}\left(\frac{3T}{4m^2J}\right)^{1/(2(J+1))}. \tag{70}$$

Proof. Fix any learner with $\sup_\nu \mathbb{E}_\nu Q_T \le S$. Take $N = \lfloor Tr/(mJ) \rfloor$, $v = (N/(mr))^{-1/(2(J+1))}$, and m/r consecutive blocks of NJ rounds, using $(m/r)NJ \le T$ rounds. Unused rounds have identical zero contexts. The horizon assumption gives $Tr/(mJ) \ge (4mr/3)16^{J+1} \ge 4$. For $x \ge 4$, $\lfloor x \rfloor \ge x - 1 \ge 3x/4$, so

$$N \ge \frac{3Tr}{4mJ}, \qquad \frac{N}{mr} \ge \frac{3T}{4m^2J} \ge 16^{J+1}.$$

In particular, $v \le 1/4$. Use m independent local hierarchies with radius $(mr)^{-1/2}$ and amplitude $\zeta = 1/8$. Partition their groups into m/r disjoint sets of size r, independently of the signs, and activate each set in exactly one block. Concatenate their parameters and divide by $\sqrt{m}$. Its squared norm is the average squared local norm, at most one. At a block activating r groups, form every combination of local actions multiplied by $\sqrt{m}$ on those groups, zero elsewhere. Its squared norm is at most $rm(mr)^{-1} = 1$.

The feature factor $\sqrt{m}$ cancels the parameter factor $1/\sqrt{m}$ in each active group's inner product. Thus the mean is the sum of the r local contributions, but the reward contains one scalar unit Gaussian noise. Conditional on all other signs, a bit flip changes only its own contribution; earlier blocks do not activate this group and contribute zero KL. Thus (69) still controls the full pre-stage history uniformly over actions. Every action combination is offered, so the best action optimizes each group separately and regret adds over wrong bits, with gap at least (68). Let s_ℓ be the prior-averaged expected internal feedback in block ℓ. Since the KL bound in (69) is $1/30 < 1/8$, apply the $3/8$ form of Lemma 13. Together with (68), this gives

$$\begin{aligned}
\mathbb{E}R_T &\ge \frac{r\sqrt{N}}{12v}\sum_{\ell=1}^{m/r}[3J/8 - s_\ell]_+ \\
&\ge \frac{r\sqrt{N}}{12v}\left(\frac{3mJ}{8r} - \sum_{\ell=1}^{m/r} s_\ell\right) \\
&\ge \frac{mJ\sqrt{N}}{96v} \ge \frac{\sqrt{3}}{192v}\sqrt{mrTJ}.
\end{aligned}$$

Here $\sum_\ell s_\ell \le \mathbb{E}Q_T \le S \le mJ/(4r)$: the uniform budget also holds after averaging over the joint finite prior. Hence $3mJ/(8r) - \sum_\ell s_\ell \ge (3/8 - 1/4)mJ/r = mJ/(8r)$, giving the factor $1/96$. The bound $N \ge 3Tr/(4mJ)$ gives the last inequality. Finally $v^{-1} = (N/(mr))^{1/(2(J+1))}$ is at least $(3T/(4m^2J))^{1/(2(J+1))}$. Every prior realization is admissible, so worst-case regret dominates this prior average; taking the infimum over budget-feasible learners proves (70). □

Proposition 3 (Binary feedback lower bound)**.** *For $d \ge 4$, $K \ge 2$, $n_\star = \log(T/d^2) \ge 8$, and $0 \le S \le dn_\star/64$,*

$$\mathfrak{R}_{\mathrm{fb}}(T, d, K, S) \ge c_1 \sqrt{dT(1 + S/d)} \exp\left(\frac{c_2 n_\star}{1 + S/d}\right)$$

with the universal constants $c_1 = 1/256$ and $c_2 = 1/32$.

Proof. Since $K \ge 2$, use Lemma 14 with $r = 1$, $m = \lfloor d/2 \rfloor$, and $J = \max\{1, \lceil 4S/m \rceil\}$. For every integer $d \ge 4$, one has $2d/5 \le m \le d/2$: for even d the upper bound is attained, and for odd $d \ge 5$, $(d-1)/2 \ge 2d/5$. Consequently

$$J \le 1 + \frac{4S}{m} \le 1 + \frac{10S}{d}, \qquad J - 1 \le \frac{5n_\star}{32}.$$

Set $\ell = \log(3T/(4m^2J))$. Since $m \le d/2$,

$$\ell \ge n_\star + \log(3/J).$$

We verify the horizon condition $\ell \ge (J+1)\log 16$ of Lemma 14, including the two smallest values of J. For $J = 1$ and $J = 2$, respectively, $n_\star \ge 8$ gives

$$8 + \log 3 > 2 \log 16, \qquad 8 + \log(3/2) > 3 \log 16.$$

For $J \ge 3$, the preceding bound on $J - 1$ gives $n_\star \ge (32/5)(J-1) > 6(J-1)$. Using $\log 16 < 3$,

$$\begin{aligned} \ell - (J+1)\log 16 &> 6(J-1) + \log(3/J) - 3(J+1) \\ &= 3(J-3) - \log(J/3) \ge 0. \end{aligned}$$

The last inequality follows from $\log(J/3) \le (J-3)/3$. Thus Lemma 14 applies throughout the asserted range.

Write $z = 1 + S/d$. The same bounds on J and $\log J \le J - 1$ give

$$J + 1 \le 2 + 10S/d \le 10z, \qquad \ell \ge n_\star + \log 3 - (J-1) \ge \frac{27n_\star}{32}.$$

It follows that

$$\frac{\ell}{2(J+1)} \ge \frac{27n_\star}{640z} \ge \frac{n_\star}{32z}.$$

Also $mJ \ge m$, $d \le 5m/2$, and $S \le mJ/4$, whence

$$dz = d + S \le \frac{5}{2}mJ + \frac{1}{4}mJ \le 3mJ.$$

Substituting these estimates into (70) yields

$$\mathfrak{R}_{\mathrm{fb}}(T, d, K, S) \ge \frac{1}{192}\sqrt{dTz}\, \exp\left(\frac{n_\star}{32z}\right) \ge \frac{1}{256}\sqrt{dTz}\, \exp\left(\frac{n_\star}{32z}\right),$$

which proves the stated constants. □

J.3 Feedback needed across independent coordinates

Lemma 15 (Coordinate-block feedback lower bound)**.** *Let $d \ge 1$, $K \ge 2$, and $T \ge d$ be integers, and set $N = \lfloor T/d \rfloor$. For every $S \ge 0$,*

$$\mathfrak{R}_{\mathrm{fb}}(T, d, K, S) \ge \frac{N}{\sqrt{d}}(d - S)_+. \tag{71}$$

If $T \ge 2d$, a learner with worst-case expected regret at most $T/(4\sqrt{d})$ must therefore have worst-case expected-query count at least $d/2$.

Proof. Draw independent uniform signs $\omega_1, \dots, \omega_d$ and take $\theta_\star = d^{-1/2}(\omega_1, \dots, \omega_d)$. Partition the first dN rounds into consecutive blocks of length N. In block i, offer the two indexed actions $(e_i, -e_i)$. Offer two identical zero vectors for any remaining rounds, and use independent $N(0,1)$ scalar reward noise. The parameter has norm one, all features have norm at most one, and the contexts do not depend on the signs. An incorrect sign choice in block i incurs regret $2/\sqrt{d}$.

Fix block i. All generated rewards before this block involve coordinates other than i. Its full pre-block history, including unrevealed earlier rewards, is therefore independent of ω_i. Include any reward delivery before the first action in that history; it reveals no block-i reward. Condition on this history and an independent seed array for the block; ω_i remains uniform. Simulate the learner until its first feedback query strictly between the block's first and last actions. If such a query occurs, suppress its reward delivery and play a fixed indexed action for the remaining block rounds. This defines a comparison action at every round using only the pre-block history, the fixed block contexts, and the private randomness.

Let F_i be the event that the real learner makes an internal query. The first internal query is chosen *before* any block-i reward has been revealed. The real and comparison executions agree up to that request, so its occurrence and time are functions of the same pre-block history and private randomness. In particular, the comparison actions and F_i are *jointly independent* of ω_i. On F_i^c, they agree with the real actions throughout the block. At every block round this gives

$$\begin{aligned}\mathbb{E}\,\mathbf{1}\{\text{real choice is wrong}\} &\geq \mathbb{E}[\mathbf{1}\{\text{comparison choice is wrong}\}\mathbf{1}\{F_i^c\}] \\ &= \tfrac{1}{2}\mathbb{P}(F_i^c).\end{aligned}$$

The equality is specific to this independent-coordinate construction. It does not assume that a no-query event is independent of a hidden sign in the general hierarchical test of Lemma 13. If $N = 1$, no internal query is possible and the same argument applies.

Consequently the expected regret in block i is at least $N\mathbb{P}(F_i^c)/\sqrt{d}$. Every query is internal to at most one block, so $\sum_i \mathbf{1}\{F_i\} \leq Q_T$. For a learner with uniform expected budget S, averaging over this finite prior and summing over blocks gives

$$\mathbb{E}R_T \geq \frac{N}{\sqrt{d}}\left(d - \sum_{i=1}^{d}\mathbb{P}(F_i)\right) \geq \frac{N}{\sqrt{d}}(d - S).$$

Nonnegative regret gives the positive part and hence the minimax bound. For the last assertion, $N \geq T/(2d)$ when $T \geq 2d$. If $S < d/2$, the bound is strictly larger than $N\sqrt{d}/2 \geq T/(4\sqrt{d})$, a contradiction. □

J.4 Inverting at a single minimax regret target

Fix $\delta > 0$ and $C_{\mathrm{reg}} > 0$. Let $d \geq 4$, $K \geq 2$, and $T \geq d^{2+\delta}$ be integers. Consider a learner with worst-case expected regret at most $C_{\mathrm{reg}}\sqrt{Td\log(1+T/d)\min\{d, \log K\}}$, and write $S = \sup_\nu \mathbb{E}_\nu Q_T$. Infinite S already satisfies the conclusion, so suppose $S < \infty$. Put

$$w = \log(2 + \min\{d, \log K\}), \qquad w_0 = \log(2 + \log 2) > 0,$$

so $w \geq w_0$. Also set

$$n_\star = \log(T/d^2), \qquad \alpha_\delta = \frac{\delta}{2+\delta}.$$

Since $\log d \leq \log T/(2+\delta)$, the following inequality holds throughout the assumed horizon range, without an asymptotic qualification:

$$n_\star \geq \alpha_\delta \log T > 0. \tag{72}$$

Define constants depending only on δ by

$$C_\delta = \frac{1}{\alpha_\delta}, \qquad M_\delta = \frac{8}{\alpha_\delta}.$$

Hierarchical inversion when $n_\star \geq 8$. Equation (72) gives

$$\log T \leq \frac{n_\star}{\alpha_\delta} = C_\delta n_\star.$$

Moreover, $T \geq d^{2+\delta} \geq 2d$ implies $1+T/d \leq T$ for $d \geq 4$, so $\log(1+T/d) \leq C_\delta n_\star$. This part of the inversion uses only the polynomial-gap part of the horizon condition in Theorem 2, not its additional comparison constraint. If $S \leq dn_\star/64$, put $z = 1 + S/d$, $x = n_\star/z$, and $A = C_{\mathrm{reg}}\sqrt{C_\delta}/c_1$. This learner is feasible in the infimum defining $\mathfrak{R}_{\mathrm{fb}}$, so $\mathfrak{R}_{\mathrm{fb}}(T, d, K, S) \leq \sup_\nu \mathbb{E}_\nu R_T$. Combining Proposition 3 with the learner's regret target gives

$$c_1\sqrt{dTz}\, e^{c_2 n_\star/z} \leq C_{\mathrm{reg}}\sqrt{dTC_\delta n_\star \min\{d, \log K\}}.$$

Dividing by $c_1\sqrt{dTn_\star}$ yields

$$\frac{e^{c_2 x}}{\sqrt{x}} \leq A\sqrt{\min\{d, \log K\}}.$$

The function $\frac{1}{2}\log x - (c_2/2)x$ is maximized at $x = 1/c_2$, so $B = \frac{1}{2}\log_+(1/(ec_2))$ satisfies $\frac{1}{2}\log x \leq (c_2/2)x + B$ for every $x > 0$. Taking logarithms and using $w \geq w_0$ and $\log_+ \min\{d, \log K\} \leq w$ gives

$$\frac{c_2 x}{2} \leq \log A + \tfrac{1}{2}\log\min\{d, \log K\} + B \leq \left(\frac{\log_+ A + B}{w_0} + \frac{1}{2}\right) w.$$

With $c_2 = 1/32$ from Proposition 3, define

$$D_{\delta, C_{\mathrm{reg}}} = \frac{\log_+ A + B}{w_0} + \frac{1}{2}, \qquad b_{\delta, C_{\mathrm{reg}}} = \min\left\{\frac{1}{64 D_{\delta, C_{\mathrm{reg}}}}, \frac{w_0}{64}\right\} > 0.$$

The preceding inequality gives $x \leq 64 D_{\delta, C_{\mathrm{reg}}} w$. Since $S = d(n_\star/x - 1)$, it follows that

$$S \geq b_{\delta, C_{\mathrm{reg}}} \frac{dn_\star}{\log(2 + \min\{d, \log K\})} - d. \tag{73}$$

By construction $b_{\delta, C_{\mathrm{reg}}} \leq w_0/64$. Hence (73) also holds if $S > dn_\star/64$, since its positive term is at most $dn_\star/64$. Thus (73) holds for every finite S when $n_\star \geq 8$, without the additional comparison constraint.

Covering every horizon under the direct comparison condition. Now impose the horizon condition in Theorem 2. Its second constraint gives $C_{\mathrm{reg}}^2 dT \log(1 + T/d)\min\{d, \log K\} \leq T^2/(16d)$; taking square roots yields

$$C_{\mathrm{reg}}\sqrt{dT \log(1 + T/d)\min\{d, \log K\}} \leq \frac{T}{4\sqrt{d}}.$$

Since $T \geq d^{2+\delta} \geq 2d$, Lemma 15 gives $S \geq d/2$ throughout this range. For $n_\star \geq 8$, $d \leq 2S$ and (73) give $3S \geq S + d \geq b_{\delta, C_{\mathrm{reg}}} dn_\star/w$. Using $n_\star \geq \log T/C_\delta$ yields

$$S \geq \frac{b_{\delta, C_{\mathrm{reg}}}}{3C_\delta} \frac{d \log T}{\log(2 + \min\{d, \log K\})}.$$

For $0 < n_\star < 8$, (72) instead gives $\log T < M_\delta$. Hence the coordinate bound alone yields

$$S \geq \frac{d}{2} \geq \frac{w_0}{2M_\delta} \frac{d \log T}{\log(2 + \min\{d, \log K\})}.$$

Taking

$$c_{\delta, C_{\mathrm{reg}}} = \min\left\{\frac{b_{\delta, C_{\mathrm{reg}}}}{3C_\delta}, \frac{w_0}{2M_\delta}\right\} > 0$$

proves (4) for every stated integer horizon. No lower threshold on T beyond the combined horizon condition in Theorem 2 remains. The coordinate construction and the hierarchy may be different instances: the uniform worst-case budget and the same regret target apply to each. The result holds for the larger feedback class, and hence for fixed within-batch policies.

For the upper comparison in the discussion of Theorem 2, choose C_{reg} at least the universal moment constant in Theorem 1 and use its single horizon-free policy. Its pathwise update guarantee also bounds worst-case expected feedback. The same numerical target constant is used for both classes. No attainment assumption is needed: the lower bound applies to every learner satisfying this fixed regret target, so taking the infimum of their worst-case expected query counts preserves the same bound. Bounded $\min\{d, \log K\}$ makes both sides $\Theta(d \log T)$. For unbounded $\min\{d, \log K\}$, the comparison in the discussion of Theorem 2 retains a genuine iterated-logarithmic gap; this proof does not establish its necessity.

K Strong sequential benchmarks under the same noise model

Sequential benchmarks are not included as additional rows in the fixed sublinear-policy-update comparison. They nevertheless determine which parts of the regret guarantee can be claimed as new. Li et al. (2024) obtain the finite-action rate $\sqrt{dT \log T \log K}$ with a residual $\mathrm{poly}(\log\log(KT))$ multiplier. For the large-action result of Li et al. (2021), taking the confidence parameter in Theorem 2 to be a fixed constant gives $\mathcal{O}(d\sqrt{T \log T} \log\log T)$ expected regret. Indeed, if B_T is that theorem's deterministic threshold, then $\mathbb{E}R_T \le B_T + \mathbb{E}(R_T - B_T)_+ \le B_T + C_2 \delta d \sqrt{T}$; take $\delta = 1/2$. That sequential comparison should not use $1/T$ merely because the table fixes that calibration for some limited-update methods.

`SquareCB.Lin`, the log-determinant-barrier reduction in the expanded arXiv version of Foster et al. (2021), supplies an important additional benchmark; the NeurIPS 2020 version calls it `SquareCB.Inf`. In its bounded-loss, realizable setting, Theorem 2 combined with online linear regression gives $\mathcal{O}(d\sqrt{T \log T})$ expected regret. For precision, the following derivation uses its *deterministic* decision inequality rather than substituting unbounded sub-Gaussian rewards into a bounded-loss theorem. It is a specialization of the existing sequential reduction, not a new `GATE` result.

Write $x_t = x_{t,a_t}$ and use immediate-feedback ridge regression with $V_0 = I$, $\widehat{\theta}_0 = 0$, $V_t = I + \sum_{i=1}^t x_i x_i^\top$, and $\widehat{\theta}_t = V_t^{-1} \sum_{i=1}^t x_i r_i$. For analysis, define

$$e_t = x_t^\top (\widehat{\theta}_{t-1} - \theta_\star), \qquad \ell_t = x_t^\top V_{t-1}^{-1} x_t \le 1, \qquad P_t = \|\widehat{\theta}_t - \theta_\star\|_{V_t}^2.$$

Here e_t is the mean-prediction error before r_t, whereas P_t measures parameter error after the update; $\ell_t \le 1$ follows from $V_{t-1} \succeq I$ and $\|x_t\|_2 \le 1$. Sherman–Morrison gives

$$\widehat{\theta}_t - \theta_\star = \widehat{\theta}_{t-1} - \theta_\star + \frac{V_{t-1}^{-1} x_t}{1 + \ell_t} (\eta_t - e_t).$$

Substituting into P_t and using $V_t = V_{t-1} + x_t x_t^\top$ gives

$$\begin{aligned} P_t - P_{t-1} &= e_t^2 + 2e_t(\eta_t - e_t) + \frac{\ell_t}{1+\ell_t}(\eta_t - e_t)^2 \\ &= -\frac{e_t^2}{1+\ell_t} + \frac{2e_t\eta_t}{1+\ell_t} + \frac{\ell_t \eta_t^2}{1+\ell_t}. \end{aligned}$$

Conditional on the past and selected feature, e_t, ℓ_t are fixed, and the noise has mean zero and second moment at most one. Thus the cross term has zero expectation. For $u \ge 0$, $\log(1+u) = \int_0^u (1+v)^{-1}\, dv \ge u/(1+u)$; the determinant identity in Section A.3 therefore gives $\sum_t \ell_t/(1+\ell_t) \le \sum_t \log(1+\ell_t) = \log\det V_T \le d\log(1 + T/d)$. Taking expectations and summing telescopes the potential differences; using $P_0 = \|\theta_\star\|_2^2 \le 1$ and $P_T \ge 0$, we obtain

$$\begin{aligned} \sum_{t=1}^T \mathbb{E} e_t^2 &\le 2 \sum_{t=1}^T \mathbb{E} \frac{e_t^2}{1+\ell_t} \\ &\le 2\mathbb{E}\left[P_0 - P_T + \sum_{t=1}^T \frac{\ell_t}{1+\ell_t} \right] \le 2[1 + d\log(1 + T/d)]. \end{aligned} \tag{74}$$

Use the action distribution p_t in Lemma 2 of Foster et al. (2021), with $\gamma > 0$ and predicted loss parameter $-\widehat{\theta}_{t-1}$. Its dimension term is $q_t = \dim \mathrm{span}\{x_{t,a} - x_{t,1} : a \in [K_t]\}$, the affine dimension.

There are at most $K_t - 1$ nonzero differences, so $q_t \le \min\{d, K-1\}$; if $q_t = 0$, all features coincide and round regret is zero. For a best action $a_t^\star$, applying that lemma to $-\theta_\star$ gives

$$\sum_{a=1}^{K_t} p_t(a)(x_{t,a_t^\star} - x_{t,a})^\top \theta_\star \le \frac{q_t}{\gamma} + \frac{\gamma}{4}\sum_{a=1}^{K_t} p_t(a)[x_{t,a}^\top(\widehat{\theta}_{t-1} - \theta_\star)]^2.$$

The minus signs convert rewards to losses. This deterministic inequality holds for arbitrary real prediction parameters and involves no observed rewards; no clipping or bounded-reward assumption is used here. Given the past and current action set, before drawing $a_t \sim p_t$, the left side is expected round regret and the last sum is expected e_t^2. Taking expectations and summing gives

$$\begin{aligned} \mathbb{E}R_T &\le \frac{T\min\{d, K-1\}}{\gamma} + \frac{\gamma}{4}\sum_{t=1}^{T}\mathbb{E}e_t^2 \\ &\le \frac{T\min\{d, K-1\}}{\gamma} + \frac{\gamma}{2}[1 + d\log(1 + T/d)] \qquad (\gamma > 0). \end{aligned} \tag{75}$$

Balancing the two terms, take $\gamma = \sqrt{2T\min\{d, K-1\}/[1 + d\log(1+T/d)]}$ to obtain

$$\mathbb{E}R_T \le \sqrt{2T\min\{d, K-1\}[1 + d\log(1+T/d)]}.$$

Since $d\log(1+T/d) \ge \log 2$ for $T, d \ge 1$, this implies the familiar $\mathcal{O}(d\sqrt{T\log(1+T/d)})$ bound. The sharper displayed bound matches our expected-regret order: $\min\{d, K-1\} = d$ in the large-action regime, and $\min\{d, K-1\} = \Theta(\min\{d, \log K\})$ for bounded K. The $\min\{d, K-1\}$ dependence, however, does not recover the optimal K-dependent rate uniformly throughout the full small-action regime. This calculation is a specialization of the existing sequential reduction, not a new `GATE` result, and it does not establish a sublinear policy-update, fixed-within-batch-policy, or second-moment guarantee for `SquareCB.Lin`. `GATE`'s statistical contribution is constant-factor minimax optimality uniformly across the full small- and large-action regimes in Proposition 1's range. It additionally attains the stated policy-update and horizon-free guarantees at the same regret level.

L Synthetic Experiments and Shared Protocol

L.1 Evaluation protocol

We compare all six complete policies using their documented native parameter rules and feedback schedules, fixed independently of evaluation outcomes. The same rules apply across synthetic and real-covariate settings, with their prescribed dependence on public inputs (Section L.5). Known-horizon methods receive T; horizon-free `GATE` uses only (d, K) and its evolving public bound for its schedule and decisions. Every method completes the declared horizon, and all replications enter aggregation without outcome-based selection.

Within each paired replication, methods share precommitted contexts, action-index permutations, and a common sequence of independent $N(0,1)$ scalar noises, but receive only their own selected-action rewards under their native feedback schedules. Only selected observations enter an estimator. Environment generation and learner randomness are independent; pairing changes the coupling between methods, not their marginal noise laws. We measure cumulative pseudo-regret using true linear means and count native nonterminal policy updates, excluding initialization and terminal feedback.

These are descriptive comparisons at native settings, not a matched-feedback-budget study or an optimized regret–update frontier. Validation in Section M selects a shared covariate construction, not bandit exploration coefficients or update schedules. Branch activity describes execution of the complete policy rather than the causal contribution of individual components.

L.2 Synthetic environments

The original-scale panel uses $T = 10{,}000$ and $(K, d) = (1000, 5), (100, 30)$ under each generator, with 50 paired replications per setting. The long-horizon panel uses $T = 2 \times 10^7$ and $(K, d) =$

$(16, 4), (128, 4)$, with 20 paired replications per setting. Both panels use the same generator formulas and independent $N(0, 1)$ reward noise. Each parameter remains fixed within a replication; parameters and contexts are generated independently of learner actions and reward observations.

I.i.d. unit-ball instances. Normalize an independent standard Gaussian vector to obtain the unit-norm parameter. For each round and indexed action, draw a Gaussian direction, normalize it, and multiply by $U^{1/d}$ for independent $U \sim \mathrm{Unif}[0, 1]$. This gives the uniform distribution in the unit Euclidean ball. Apply a fresh, learner-independent permutation of the K indices. The entire random sequence can equivalently be generated before interaction, so the model is a distribution over oblivious instances.

Oblivious rotating-spectrum instances. Take $\theta_\star$ to have independent Rademacher coordinates divided by $\sqrt{d}$. There are eight phases. The first seven lengths are $\lfloor T2^{b-1}/(2^8 - 1) \rfloor$, $b = 1, \ldots, 7$, and the last phase uses the remainder. A pool of 32 Haar orthogonal matrices is drawn independently of learner behavior. In odd and even phases the covariance condition-number parameters are 10^2 and 10^4, respectively; the linear transformation has eigenvalues $\kappa^{-j/(2(d-1))}$, $j = 0, \ldots, d - 1$.

For every candidate rotation, pre-generate the phase's unit-sphere features and calculate the median top-two mean-reward gap over the entire phase. Choose an unused rotation with smallest median gap, preferring rotations whose leading axis makes an absolute principal angle of at least $\pi/6$ with the previous leading axis. If no unused rotation meets this angle condition, drop only the angle restriction. Ties use the candidate index. For phase length n_b and median unscaled gap g_b, multiply all phase features by $\min\{1, [4\sqrt{n_b} \max\{g_b, 10^{-12}\}]^{-1}\}$. The transformation norm and scale are at most one, so features remain in the unit ball. Rotation choice can depend on the fixed parameter and the pre-generated features, but not on rewards or any learner trajectory. It is therefore parameter-aware and oblivious.

L.3 Original-scale results: root-dominated execution

Figures 1 and 2 show the zoomed and full-range original-scale results. Both GATE versions have lower sample-mean terminal regret and fewer mean updates than the four baselines. Neither version makes a later-layer assignment: these are complete-policy runs whose layer-assigned training observations enter layer 0, not a later-layer ablation. Tables 5 and 8 give terminal statistics and paired differences.

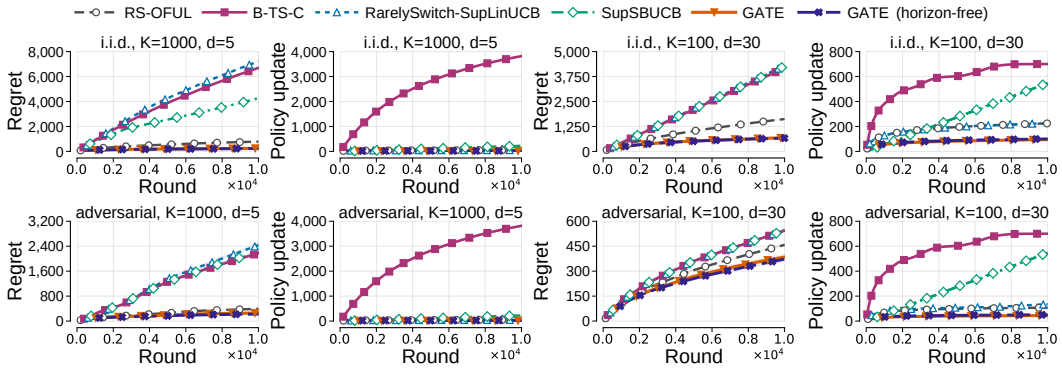

Figure 2: Full-range original-scale results ($T = 10^4$, $n = 50$ paired replications per setting). Arrangement and methods are as in Figure 1. Curves show means with pointwise ± 1.96 SE bands, clipped below at zero. Regret and update scales vary across settings.

L.4 Long-horizon results

The four long-horizon settings use the generators above at their full horizons, with K generated indexed feature vectors in every round. The action counts span both regimes: $\log 16 \simeq 2.77259 < d = 4 < \log 128 \simeq 4.85203$. Table 2 gives the known-horizon parameter values. Their final-layer

thresholds are uncapped, since $\Phi/(4T) < 1$; this alone implies neither a numerical regret bound below $2T$ nor a later-layer assignment in every replication.

Table 2: Long-horizon synthetic configurations, each evaluated under both generators with $n = 20$ paired replications. The last three columns are deterministic known-horizon parameter values.

T	K	d	L	$\Phi/(4T)$	ε_{L-1}
20,000,000	16	4	2	0.682265499	0.341132749
20,000,000	128	4	2	0.775149294	0.387574647

Figures 3 and 4 show all four settings; Tables 6 and 9 report terminal outcomes and paired intervals. Both `GATE` versions have lower sample-mean terminal regret than all four baselines. On the i.i.d. instances, both also use fewer mean updates; known-horizon `GATE` has the smallest sample-mean update count in all four settings. For rotating contexts at $K = 16$, `GATE`'s regret is more variable than in the i.i.d. counterpart, and the paired regret intervals against `RS-OFUL` include zero for both versions. Horizon-free `GATE` also uses more mean updates there (31.85) than `RS-OFUL` (26.55) and `RarelySwitch-SupLinUCB` (27.70).

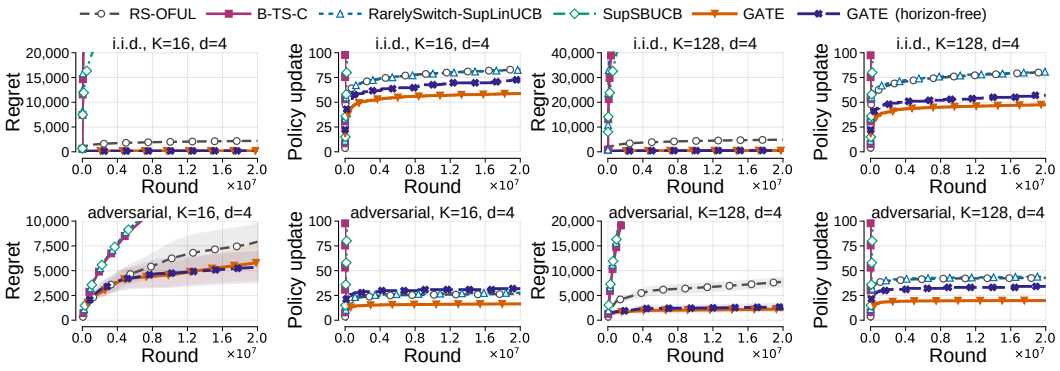

Figure 3: Long-horizon results ($T = 2 \times 10^7$, $d = 4$, $n = 20$ paired replications per setting). Rows show i.i.d. and oblivious rotating contexts; column pairs use $K = 16, 128$. Curves show all six methods' means with pointwise ± 1.96 SE bands. Axes are zoomed; full ranges and terminal statistics are in Figure 4 and Table 6.

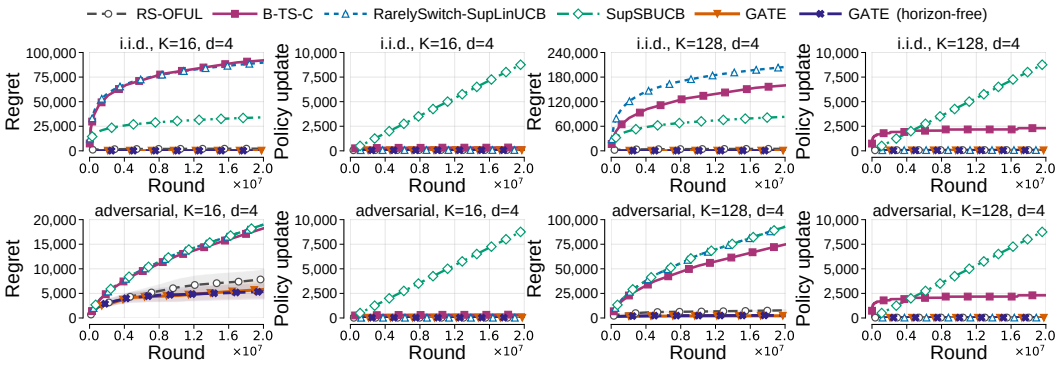

Figure 4: Full-range companion to Figure 3, with the same six methods, $n = 20$ paired replications, and pointwise ± 1.96 SE bands. Only the vertical display ranges differ.

Horizon-free schedule. For both action counts, the schedule starts at $H_0 = 2$ ((53)), and the first uncapped epoch begins before round 8,388,608, when H becomes 16,777,216. Each run includes 24 deterministic public updates and retains all observations from earlier capped epochs. Thus the known-horizon cap status does not label the entire horizon-free trajectory.

L.5 SIX-METHOD IMPLEMENTATION

The bandit ridge parameter is one. `RS-OFUL` uses determinant factor two; `RarelySwitch-SupLinUCB` uses $C = 2$; `B-TS-C` uses native per-arm doubling; and `SupSBUCB` uses $\lceil\sqrt{dT}\rceil$ uniform batches. Native probability calibrations are $1/T$. Empty native queries still count as nonterminal updates; their contents are not modified to include the triggering observation. These conventions are fixed across panels.

`GATE` uses the public parameter map in Section G. Its two versions share corresponding Gaussian random variables indexed by global round, layer, and draw type, with independent test and selection variables. This coupling does not force equal actions or update times. The horizon-free constructor receives the simulation horizon only for accounting; its public schedule and decisions do not use the final T.

L.6 NUMERICAL IMPLEMENTATION AND STATISTICAL REPORTING

`GATE` uses the incremental Cholesky maintenance and snapshot copying in Section B, including layer-authorized prefix updates and unchanged prefixes at pool-only endpoints. Calculations use float64, positive-diagonal lower Cholesky factors, log-domain threshold comparisons, fixed input-index ties, and disabled `fastmath`. Near a determinant threshold, the log-determinant ratio is recomputed from factor diagonals; invalid factors raise an error rather than invoking dense refactorization or jitter. Arithmetic and storage accounting, including pending observations, are given in Section B.

Statistical reporting. The replication, not the round, is the sampling unit. For a metric $Z_i(t)$, we report its sample mean $\bar{Z}(t)$, sample standard deviation $\mathrm{SD}(t)$, and standard error $\mathrm{SE}(t) = \mathrm{SD}(t)/\sqrt{n}$, using the $n-1$ denominator for sample variances. Marginal metric tables give means $\pm$ one SE and separate SDs. Curve bands are pointwise normal-approximation bands $\bar{Z}(t) \pm 1.96\,\mathrm{SE}(t)$, clipped below at zero, not simultaneous bands.

For baseline m, `GATE` version g, and $Z \in \{R_T, S_T\}$, define the paired differences and their intervals by

$$D_i^Z = Z_{m,i} - Z_{g,i}, \qquad \widehat{\Delta}_{m,g}^Z = \frac{1}{n}\sum_{i=1}^{n} D_i^Z, \qquad I_{m,g}^Z = \widehat{\Delta}_{m,g}^Z \pm t_{n-1,0.975}\frac{s_D}{\sqrt{n}},$$

where s_D is the sample SD of the paired differences and $t_{n-1,0.975}$ is the Student t quantile. Positive differences favor the indicated `GATE` version. These nominal two-sided 95% intervals are descriptive, unadjusted per-comparison summaries, not a prespecified confirmatory family or simultaneous intervals for the 160 contrasts. Coverage is approximate for non-Gaussian differences; zero sample variance does not establish zero population variance. Marginal band overlap does not determine paired significance, and an interval containing zero does not establish equivalence.

Synthetic replications vary the instance, noise, and learner randomness. Real-covariate replications condition on the shared fitted representation and retained blocks, varying block order, action indices, noise, and learner randomness. These descriptive averages neither establish a population ranking nor empirically verify the fixed-instance moment guarantee in Theorem 1.

Every round contributes to cumulative regret and terminal statistics; sparse plotting affects display resolution only. Update curves use the exact union of native nonterminal endpoints as post-step curves. Zoomed and full-range plots use the same replications, mean curves, and uncertainty bands; only their vertical display ranges differ.

L.7 BRANCH ACTIVITY

Table 3 reports long-horizon branch activity. Both `GATE` versions make later-layer assignments in every i.i.d. replication. For rotating contexts at $K = 16$, known-horizon `GATE` makes six such assignments in one of 20 replications, while the horizon-free version makes none; at $K = 128$, 16 of 20 known-horizon and 17 of 20 horizon-free runs are active. A reached Gaussian test may instead be followed by a response read and terminal selection, so test counts and assignment counts measure different operations. Each reached exploration branch uses an independent selection draw. Every experimental `GATE` epoch has $L = 2$.

Without later-layer assignments, those layers remain at $(I, 0)$: their zero predicted means and common radius yield identical intervals, so response reads retain the incoming action set. Terminal ranking still uses the pooled estimator. This explains the root-dominated original-scale and real-covariate runs without identifying the causal effect of higher-layer exploration; conditional independence follows from Lemma 7, not branch frequencies.

Table 3: Branch activity in long-horizon synthetic experiments ($T = 2 \times 10^7$, $n = 20$ per setting). Root, test, and terminal percentages use all T rounds and are averaged over replications. Noninitial is the mean number of assignments to layers $s \geq 1$; Active counts replications with at least one. Terminal combines both terminal branches; Pool-only gives mean nonterminal pool-only endpoints. Simultaneous requests count once.

Version	Root (%)	Noninitial	Tests (%)	Terminal (%)	Active	Pool-only
I.i.d., $(K, d) = (16, 4)$						
Known horizon	7.59	4,600.40	91.20	92.39	20/20	7.80
Horizon-free	9.04	4,251.00	88.52	90.94	20/20	8.10
I.i.d., $(K, d) = (128, 4)$						
Known horizon	6.42	4,300.85	86.75	93.56	20/20	7.10
Horizon-free	7.06	3,366.30	83.36	92.93	20/20	4.00
Rotating, $(K, d) = (16, 4)$						
Known horizon	3.78	0.30	95.56	96.22	1/20	5.00
Horizon-free	2.19	0.00	97.36	97.81	0/20	2.30
Rotating, $(K, d) = (128, 4)$						
Known horizon	4.16	588.60	92.98	95.84	16/20	3.35
Horizon-free	2.75	604.60	95.56	97.24	17/20	1.25

M Real-Covariate Experiments: Jester and KuaiRec

M.1 Purpose and data sources

We use recommendation data to construct contexts and simulate realizable linear rewards. These *real-covariate, semi-synthetic controls* preserve the norm and independent-noise assumptions; they are not online evaluations using observed user rewards.

For Jester, we use all three archives of the original 100-joke release (Goldberg et al., 2001), excluding the missing-rating sentinel 99 and invalid ratings and dividing valid ratings by 10 for representation learning. For KuaiRec, we use `small_matrix.csv` from the raw-feature release at Zenodo record 18164998 (Gao et al., 2022; Gao, 2026). We retain format-valid nonnegative user/video identifiers, finite timestamps, positive durations, and finite nonnegative watch ratios. The watch ratio w is mapped to $(w-1)/(w+1)$ for representation learning. Repeated user–video pairs are reduced to the earliest timestamp, with original row number breaking ties. These targets are not the online bandit rewards.

M.2 User splitting and representation learning

Users are sorted by a fixed identifier-only SHA-256 hash, with exact ties broken by the canonical user key. The first $\lfloor 0.6n \rfloor$ users train the representation, the next $\lfloor 0.2n \rfloor$ validate it, and the remainder form test contexts. Only items observed in the representation split receive factors, ordered by numeric identifier.

We fit rank-5 regularized matrix factorization by alternating ridge least squares, updating user factors and then item factors for 20 complete sweeps at each regularization value. User factors start at zero and item factors at independent $N(0, 0.1^2/5)$ entries. Representation and profile regularization values both come from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. The selected pair minimizes validation nonprofile RMSE, with exact ties preferring the larger representation regularizer and then the larger

profile regularizer. Bandit regret and test-candidate targets are not used for selection. The selected representation is shared by all six methods and fixed across replications.

Within each Jester user, entries are ordered by a fixed identifier-only item hash, with joke identifier breaking ties. KuaiRec entries are ordered by timestamp, video identifier, and original row number. The first ten usable entries infer a five-dimensional user profile by ridge regression on frozen item factors. Subsequent entries form consecutive complete blocks of 20 candidate items; only the incomplete tail is discarded. Profile and candidate items are disjoint. Candidate targets are not read to construct test features or score bandit rewards.

M.3 Unit-norm bandit construction and fixed streams

Let $p_u \in \mathbb{R}^5$ denote the fitted profile and $v_i \in \mathbb{R}^5$ a frozen item factor. For each retained candidate pair, use

$$x_{u,i} = \frac{\mathrm{vec}(p_u v_i^\top)}{c_D}, \qquad c_D = \max\{1, \max_{(u,i)} \|p_u\|_2 \|v_i\|_2\}, \qquad \theta_\star = \frac{\mathrm{vec}(I_5)}{\sqrt{5}}. \tag{76}$$

Thus $d = 25$, $K = 20$, $\|\theta_\star\|_2 = 1$, and feature norms are at most one. The mean is $p_u^\top v_i/(\sqrt{5}c_D)$. The maximum defining c_D uses features, not candidate targets; vectorization is consistent between features and the parameter.

All 27,246 retained Jester blocks and 44,782 retained KuaiRec blocks are used, giving the respective horizons. Each of the 50 paired replications per dataset permutes all blocks and independently permutes the 20 action indices within each block, using the shared-context and independent $N(0, 1)$ noise protocol in Section L.1. The blocks cover 10,880 eligible Jester test users and 283 eligible KuaiRec test users, with zero profile/candidate overlaps and zero repeated candidate pairs.

Table 4: Preprocessing summary ($K = 20$, $d = 25$, $n = 50$ paired replications per dataset). Validation RMSE uses nonprofile representation-learning targets, not online bandit rewards.

Dataset	T	λ_{MF}	$\lambda_{\mathrm{profile}}$	c_D	Validation RMSE
Jester	27,246	100	10	1.742981	0.450828
KuaiRec	44,782	0.1	1	4.014734	0.268408

M.4 Regret and policy-update results

Figures 5 and 6 show the results; Tables 7 and 10 give terminal statistics and paired intervals. Both `GATE` versions have lower sample-mean terminal regret and fewer mean updates than all four baselines on each dataset. These comparisons are conditional on the constructed covariates and simulated rewards, not observed gains in recommendation engagement or ratings.

Neither version makes a later-layer assignment. Mean root-assignment counts for known-horizon / horizon-free `GATE` are $27{,}223.98/27{,}223.70$ of $27{,}246$ on Jester, with the remaining decisions split between root-certified and complete-cascade terminal branches. On KuaiRec, the counts are $44{,}781.80/44{,}781.72$ of $44{,}782$; the remaining decisions are root-certified and no later-layer test is reached. Thus these complete-policy runs are root-dominated, as discussed in Section L.7.

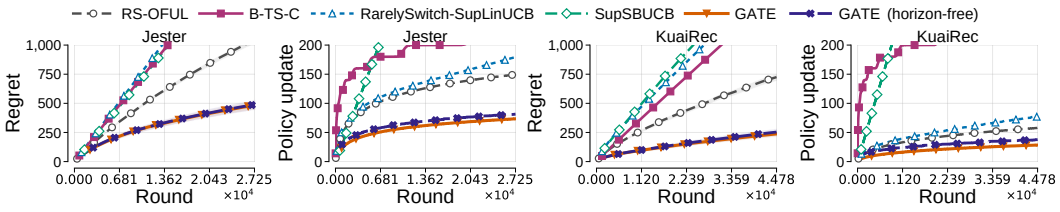

Figure 5: Real-covariate, semi-synthetic controls: Jester at $T = 27{,}246$ and KuaiRec at $T = 44{,}782$, with $K = 20$, $d = 25$, and $n = 50$ paired replications. Each pair shows cumulative pseudo-regret and nonterminal policy updates for all six methods, with means and pointwise ± 1.96 SE bands. Axes are zoomed; full ranges and terminal statistics are in Figure 6 and Table 7. Round axes are linear with $\times 10^4$ multipliers.

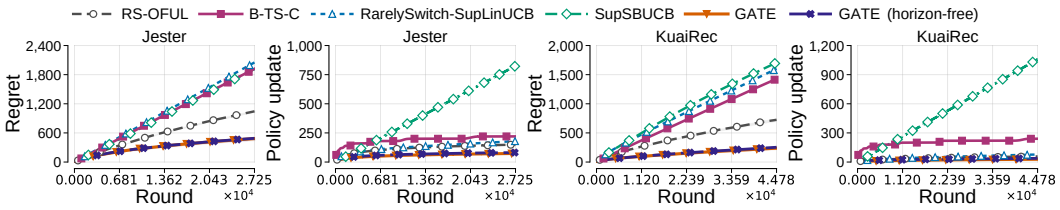

Figure 6: Full-range companion to Figure 5, with the same six methods, $n = 50$ paired replications, and pointwise ± 1.96 SE bands. Only the vertical display ranges differ.

N Terminal Regret, Policy Updates, and Runtime

N.1 Terminal regret and policy updates

Tables 5 to 7 report all six methods at the native settings in Sections L.1 and L.5. Entries are sample means $\pm$ one SE, with sample SDs in adjacent columns; Section L.6 defines the uncertainty summaries. Values are rounded only after aggregation. Boldface marks the smallest unrounded sample mean within a setting and metric, not significance. Update counts follow the nonterminal endpoint convention; in particular, SupSBUCB has exactly $\lceil \sqrt{dT} \rceil - 1$ updates.

Table 5: Terminal results for the original-scale synthetic panel ($T = 10{,}000$, $n = 50$ paired replications per setting). Entries are means $\pm$ one SE with adjacent sample SDs. Boldface marks the smallest unrounded sample mean, not significance.

Method	R_T: mean $\pm$ SE	SD(R_T)	S_T: mean $\pm$ SE	SD(S_T)
I.i.d., $(K, d) = (1000, 5)$				
RS-OFUL	783.77 ± 9.05	64.01	45.10 ± 0.04	0.30
B-TS-C	$6{,}694.97 \pm 13.15$	93.00	$3{,}817.36 \pm 1.06$	7.52
RarelySwitch-SupLinUCB	$7{,}218.98 \pm 24.35$	172.21	50.00 ± 0.00	0.00
SupSBUCB	$4{,}269.66 \pm 20.69$	146.32	223.00 ± 0.00	0.00
GATE	246.47 ± 12.53	88.57	$\mathbf{17.00 \pm 0.00}$	0.00
GATE (horizon-free)	$\mathbf{225.69 \pm 9.75}$	68.97	22.96 ± 0.06	0.40
I.i.d., $(K, d) = (100, 30)$				
RS-OFUL	$1{,}621.36 \pm 12.91$	91.29	224.44 ± 0.10	0.67
B-TS-C	$4{,}176.95 \pm 3.07$	21.70	700.38 ± 0.09	0.64
RarelySwitch-SupLinUCB	$4{,}298.41 \pm 2.93$	20.70	225.76 ± 0.06	0.43
SupSBUCB	$4{,}270.46 \pm 4.70$	33.20	547.00 ± 0.00	0.00
GATE	682.16 ± 20.56	145.40	$\mathbf{94.42 \pm 0.07}$	0.50
GATE (horizon-free)	$\mathbf{663.48 \pm 15.05}$	106.45	100.02 ± 0.11	0.80
Rotating, $(K, d) = (1000, 5)$				
RS-OFUL	391.88 ± 13.62	96.28	43.80 ± 0.11	0.81
B-TS-C	$2{,}176.29 \pm 29.19$	206.42	$3{,}817.72 \pm 1.28$	9.04
RarelySwitch-SupLinUCB	$2{,}436.20 \pm 37.59$	265.80	44.58 ± 0.14	0.97
SupSBUCB	$2{,}163.90 \pm 26.10$	184.56	223.00 ± 0.00	0.00
GATE	257.48 ± 22.76	160.94	$\mathbf{16.38 \pm 0.07}$	0.53
GATE (horizon-free)	$\mathbf{220.47 \pm 17.00}$	120.24	23.06 ± 0.14	0.96
Rotating, $(K, d) = (100, 30)$				
RS-OFUL	458.06 ± 5.40	38.20	107.32 ± 0.41	2.89
B-TS-C	543.82 ± 1.19	8.41	700.32 ± 0.08	0.55
RarelySwitch-SupLinUCB	545.56 ± 1.27	8.99	134.00 ± 0.48	3.43
SupSBUCB	545.29 ± 1.30	9.17	547.00 ± 0.00	0.00
GATE	384.62 ± 9.46	66.87	$\mathbf{44.06 \pm 0.17}$	1.22
GATE (horizon-free)	$\mathbf{372.25 \pm 9.76}$	69.03	50.20 ± 0.21	1.50

Table 6: Terminal results for the long-horizon synthetic panel ($T = 2 \times 10^7$, $n = 20$ paired replications per setting). Entries are means $\pm$ one SE with adjacent sample SDs. Boldface marks the smallest unrounded sample mean, not significance.

Method	R_T: mean $\pm$ SE	SD(R_T)	S_T: mean $\pm$ SE	SD(S_T)
I.i.d., $(K, d) = (16, 4)$				
RS-OFUL	2,211.67 $\pm$ 31.01	138.70	83.05 $\pm$ 0.05	0.22
B-TS-C	92,027.03 $\pm$ 91.05	407.21	336.00 $\pm$ 0.00	0.00
RarelySwitch-SupLinUCB	89,758.21 $\pm$ 74.06	331.20	82.85 $\pm$ 0.08	0.37
SupSBUCB	34,078.62 $\pm$ 41.43	185.26	8,944.00 $\pm$ 0.00	0.00
GATE	**183.17 $\pm$ 7.82**	34.96	**58.80 $\pm$ 0.22**	1.01
GATE (horizon-free)	201.59 $\pm$ 8.64	38.65	72.65 $\pm$ 0.23	1.04
I.i.d., $(K, d) = (128, 4)$				
RS-OFUL	4,898.75 $\pm$ 58.37	261.05	80.20 $\pm$ 0.09	0.41
B-TS-C	159,698.85 $\pm$ 59.95	268.09	2,304.00 $\pm$ 0.00	0.00
RarelySwitch-SupLinUCB	204,699.18 $\pm$ 92.10	411.88	80.95 $\pm$ 0.05	0.22
SupSBUCB	82,955.82 $\pm$ 51.44	230.05	8,944.00 $\pm$ 0.00	0.00
GATE	**469.27 $\pm$ 27.45**	122.75	**47.50 $\pm$ 0.14**	0.61
GATE (horizon-free)	542.76 $\pm$ 30.00	134.16	57.00 $\pm$ 0.00	0.00
Rotating, $(K, d) = (16, 4)$				
RS-OFUL	7,916.50 $\pm$ 995.93	4,453.94	26.55 $\pm$ 0.32	1.43
B-TS-C	18,224.61 $\pm$ 56.36	252.05	336.00 $\pm$ 0.00	0.00
RarelySwitch-SupLinUCB	19,000.25 $\pm$ 2.29	10.22	27.70 $\pm$ 0.31	1.38
SupSBUCB	19,001.58 $\pm$ 1.94	8.65	8,944.00 $\pm$ 0.00	0.00
GATE	5,801.40 $\pm$ 947.73	4,238.39	**16.45 $\pm$ 0.26**	1.15
GATE (horizon-free)	**5,349.63 $\pm$ 831.81**	3,719.96	31.85 $\pm$ 0.38	1.69
Rotating, $(K, d) = (128, 4)$				
RS-OFUL	7,756.58 $\pm$ 578.72	2,588.13	42.65 $\pm$ 0.36	1.60
B-TS-C	74,759.53 $\pm$ 469.06	2,097.70	2,304.00 $\pm$ 0.00	0.00
RarelySwitch-SupLinUCB	92,943.97 $\pm$ 8.19	36.61	43.90 $\pm$ 0.37	1.65
SupSBUCB	92,720.82 $\pm$ 116.39	520.53	8,944.00 $\pm$ 0.00	0.00
GATE	**2,235.13 $\pm$ 168.71**	754.49	**19.80 $\pm$ 0.33**	1.47
GATE (horizon-free)	2,657.18 $\pm$ 466.00	2,084.01	34.10 $\pm$ 0.31	1.37

Table 7: Terminal results for the real-covariate controls ($n = 50$ paired replications per dataset). Entries are means $\pm$ one SE with adjacent sample SDs. Boldface marks the smallest unrounded sample mean, not significance.

Method	R_T: mean $\pm$ SE	SD(R_T)	S_T: mean $\pm$ SE	SD(S_T)
Jester, $T = 27{,}246$*,* $(K, d) = (20, 25)$				
RS-OFUL	1,044.90 $\pm$ 12.97	91.73	149.30 $\pm$ 0.11	0.79
B-TS-C	1,920.19 $\pm$ 1.61	11.40	220.00 $\pm$ 0.00	0.00
RarelySwitch-SupLinUCB	2,051.33 $\pm$ 0.46	3.25	179.52 $\pm$ 0.09	0.65
SupSBUCB	1,914.04 $\pm$ 0.92	6.53	825.00 $\pm$ 0.00	0.00
GATE	**483.31 $\pm$ 16.55**	117.04	**73.42 $\pm$ 0.11**	0.81
GATE (horizon-free)	488.91 $\pm$ 21.44	151.62	81.44 $\pm$ 0.18	1.30
KuaiRec, $T = 44{,}782$*,* $(K, d) = (20, 25)$				
RS-OFUL	725.31 $\pm$ 14.63	103.48	57.72 $\pm$ 0.09	0.61
B-TS-C	1,428.77 $\pm$ 1.44	10.15	240.00 $\pm$ 0.00	0.00
RarelySwitch-SupLinUCB	1,601.24 $\pm$ 0.40	2.82	77.04 $\pm$ 0.03	0.20
SupSBUCB	1,707.25 $\pm$ 0.56	3.98	1,058.00 $\pm$ 0.00	0.00
GATE	**237.76 $\pm$ 11.22**	79.35	**28.36 $\pm$ 0.09**	0.66
GATE (horizon-free)	253.45 $\pm$ 15.53	109.79	37.76 $\pm$ 0.16	1.13

N.2 PAIRED TERMINAL COMPARISONS

Tables 8 to 10 give terminal regret and update differences for all four baselines against each GATE version in all ten settings. Each cell contains the mean paired difference and its nominal 95% Student interval, as defined in Section L.6; positive values favor the indicated GATE version. KH and HF denote known horizon and horizon free, and RS-SupLin abbreviates RarelySwitch-SupLinUCB in these tables. The intervals are unadjusted per-comparison summaries.

Table 8: Paired terminal differences for original-scale synthetic settings ($T = 10^4$, $n = 50$). Cells give baseline minus the indicated GATE version: mean [nominal 95% interval]. Intervals are unadjusted; abbreviations are defined in Section N.2.

Baseline	Ref.	ΔR_T [95% interval]	ΔS_T [95% interval]
I.i.d., $(K, d) = (1000, 5)$			
RS-OFUL	KH	537.30 [508.36, 566.25]	28.10 [28.01, 28.19]
RS-OFUL	HF	558.08 [533.76, 582.40]	22.14 [22.00, 22.28]
B-TS-C	KH	6,448.50 [6,410.04, 6,486.97]	3,800.36 [3,798.22, 3,802.50]
B-TS-C	HF	6,469.28 [6,434.73, 6,503.84]	3,794.40 [3,792.25, 3,796.55]
RS-SupLin	KH	6,972.52 [6,916.20, 7,028.83]	33.00 [33.00, 33.00]
RS-SupLin	HF	6,993.29 [6,936.46, 7,050.13]	27.04 [26.93, 27.15]
SupSBUCB	KH	4,023.19 [3,973.66, 4,072.73]	206.00 [206.00, 206.00]
SupSBUCB	HF	4,043.97 [3,995.33, 4,092.62]	200.04 [199.93, 200.15]
I.i.d., $(K, d) = (100, 30)$			
RS-OFUL	KH	939.20 [890.98, 987.41]	130.02 [129.76, 130.28]
RS-OFUL	HF	957.88 [917.77, 997.98]	124.42 [124.12, 124.72]
B-TS-C	KH	3,494.79 [3,454.76, 3,534.82]	605.96 [605.70, 606.22]
B-TS-C	HF	3,513.47 [3,481.11, 3,545.83]	600.36 [600.04, 600.68]
RS-SupLin	KH	3,616.25 [3,574.81, 3,657.70]	131.34 [131.15, 131.53]
RS-SupLin	HF	3,634.93 [3,604.76, 3,665.11]	125.74 [125.46, 126.02]
SupSBUCB	KH	3,588.30 [3,546.36, 3,630.23]	452.58 [452.44, 452.72]
SupSBUCB	HF	3,606.97 [3,574.41, 3,639.54]	446.98 [446.75, 447.21]
Rotating, $(K, d) = (1000, 5)$			
RS-OFUL	KH	134.40 [90.66, 178.14]	27.42 [27.22, 27.62]
RS-OFUL	HF	171.41 [142.31, 200.51]	20.74 [20.46, 21.02]
B-TS-C	KH	1,918.81 [1,834.26, 2,003.36]	3,801.34 [3,798.79, 3,803.89]
B-TS-C	HF	1,955.82 [1,876.71, 2,034.92]	3,794.66 [3,792.07, 3,797.25]
RS-SupLin	KH	2,178.72 [2,080.50, 2,276.94]	28.20 [27.94, 28.46]
RS-SupLin	HF	2,215.72 [2,120.51, 2,310.94]	21.52 [21.19, 21.85]
SupSBUCB	KH	1,906.42 [1,832.12, 1,980.73]	206.62 [206.47, 206.77]
SupSBUCB	HF	1,943.43 [1,874.78, 2,012.08]	199.94 [199.67, 200.21]
Rotating, $(K, d) = (100, 30)$			
RS-OFUL	KH	73.45 [50.58, 96.32]	63.26 [62.76, 63.76]
RS-OFUL	HF	85.81 [63.52, 108.10]	57.12 [56.59, 57.65]
B-TS-C	KH	159.20 [140.37, 178.04]	656.26 [655.87, 656.65]
B-TS-C	HF	171.57 [152.67, 190.47]	650.12 [649.65, 650.59]
RS-SupLin	KH	160.94 [142.24, 179.65]	89.94 [89.27, 90.61]
RS-SupLin	HF	173.30 [154.38, 192.22]	83.80 [83.11, 84.49]
SupSBUCB	KH	160.68 [141.91, 179.44]	502.94 [502.59, 503.29]
SupSBUCB	HF	173.04 [154.11, 191.97]	496.80 [496.37, 497.23]

Table 9: Paired terminal differences for long-horizon synthetic settings ($T = 2 \times 10^7$, $n = 20$). Cells give baseline minus the indicated GATE version: mean [nominal 95% interval]. Intervals are unadjusted; abbreviations are defined in Section N.2.

Baseline	Ref.	ΔR_T [95% interval]	ΔS_T [95% interval]
I.i.d., $(K, d) = (16, 4)$			
RS-OFUL	KH	2,028.50 [1,961.65, 2,095.34]	24.25 [23.75, 24.75]
RS-OFUL	HF	2,010.07 [1,946.39, 2,073.75]	10.40 [9.93, 10.87]
B-TS-C	KH	91,843.86 [91,659.34, 92,028.39]	277.20 [276.73, 277.67]
B-TS-C	HF	91,825.44 [91,641.49, 92,009.40]	263.35 [262.86, 263.84]
RS-SupLin	KH	89,575.04 [89,422.71, 89,727.38]	24.05 [23.56, 24.54]
RS-SupLin	HF	89,556.62 [89,403.90, 89,709.34]	10.20 [9.75, 10.65]
SupSBUCB	KH	33,895.45 [33,807.31, 33,983.59]	8,885.20 [8,884.73, 8,885.67]
SupSBUCB	HF	33,877.03 [33,786.00, 33,968.06]	8,871.35 [8,870.86, 8,871.84]
I.i.d., $(K, d) = (128, 4)$			
RS-OFUL	KH	4,429.48 [4,313.39, 4,545.57]	32.70 [32.32, 33.08]
RS-OFUL	HF	4,355.98 [4,244.01, 4,467.96]	23.20 [23.01, 23.39]
B-TS-C	KH	159,229.58 [159,074.06, 159,385.10]	2,256.50 [2,256.22, 2,256.78]
B-TS-C	HF	159,156.09 [159,010.58, 159,301.59]	2,247.00 [2,247.00, 2,247.00]
RS-SupLin	KH	204,229.91 [204,029.32, 204,430.50]	33.45 [33.09, 33.81]
RS-SupLin	HF	204,156.41 [203,946.11, 204,366.72]	23.95 [23.85, 24.05]
SupSBUCB	KH	82,486.55 [82,373.62, 82,599.48]	8,896.50 [8,896.22, 8,896.78]
SupSBUCB	HF	82,413.06 [82,304.39, 82,521.72]	8,887.00 [8,887.00, 8,887.00]
Rotating, $(K, d) = (16, 4)$			
RS-OFUL	KH	2,115.10 [−159.48, 4,389.68]	10.10 [9.37, 10.83]
RS-OFUL	HF	2,566.86 [−447.19, 5,580.91]	−5.30 [−6.17, −4.43]
B-TS-C	KH	12,423.21 [10,387.71, 14,458.71]	319.55 [319.01, 320.09]
B-TS-C	HF	12,874.98 [11,155.84, 14,594.12]	304.15 [303.36, 304.94]
RS-SupLin	KH	13,198.85 [11,216.51, 15,181.19]	11.25 [10.59, 11.91]
RS-SupLin	HF	13,650.62 [11,908.16, 15,393.07]	−4.15 [−4.97, −3.33]
SupSBUCB	KH	13,200.19 [11,217.42, 15,182.95]	8,927.55 [8,927.01, 8,928.09]
SupSBUCB	HF	13,651.95 [11,910.15, 15,393.75]	8,912.15 [8,911.36, 8,912.94]
Rotating, $(K, d) = (128, 4)$			
RS-OFUL	KH	5,521.45 [4,523.11, 6,519.78]	22.85 [22.22, 23.48]
RS-OFUL	HF	5,099.40 [3,983.24, 6,215.56]	8.55 [7.85, 9.25]
B-TS-C	KH	72,524.40 [71,462.95, 73,585.84]	2,284.20 [2,283.51, 2,284.89]
B-TS-C	HF	72,102.35 [70,775.23, 73,429.47]	2,269.90 [2,269.26, 2,270.54]
RS-SupLin	KH	90,708.84 [90,356.79, 91,060.89]	24.10 [23.42, 24.78]
RS-SupLin	HF	90,286.80 [89,313.03, 91,260.56]	9.80 [9.03, 10.57]
SupSBUCB	KH	90,485.69 [90,083.54, 90,887.84]	8,924.20 [8,923.51, 8,924.89]
SupSBUCB	HF	90,063.64 [89,083.57, 91,043.72]	8,909.90 [8,909.26, 8,910.54]

Table 10: Paired terminal differences for real-covariate controls ($n = 50$). Cells give baseline minus the indicated GATE version: mean [nominal 95% interval]. Intervals are unadjusted; abbreviations are defined in Section N.2.

Baseline	Ref.	ΔR_T [95% interval]	ΔS_T [95% interval]
Jester, $T = 27{,}246$, $(K, d) = (20, 25)$			
RS-OFUL	KH	561.59 [526.93, 596.24]	75.88 [75.60, 76.16]
RS-OFUL	HF	555.99 [511.37, 600.62]	67.86 [67.48, 68.24]
B-TS-C	KH	1,436.87 [1,403.91, 1,469.84]	146.58 [146.35, 146.81]
B-TS-C	HF	1,431.28 [1,388.50, 1,474.06]	138.56 [138.19, 138.93]
RS-SupLin	KH	1,568.02 [1,534.92, 1,601.12]	106.10 [105.80, 106.40]
RS-SupLin	HF	1,562.43 [1,519.47, 1,605.38]	98.08 [97.68, 98.48]
SupSBUCB	KH	1,430.73 [1,397.53, 1,463.93]	751.58 [751.35, 751.81]
SupSBUCB	HF	1,425.14 [1,382.13, 1,468.15]	743.56 [743.19, 743.93]
KuaiRec, $T = 44{,}782$, $(K, d) = (20, 25)$			
RS-OFUL	KH	487.55 [455.62, 519.48]	29.36 [29.12, 29.60]
RS-OFUL	HF	471.86 [431.72, 511.99]	19.96 [19.63, 20.29]
B-TS-C	KH	1,191.01 [1,168.81, 1,213.21]	211.64 [211.45, 211.83]
B-TS-C	HF	1,175.32 [1,144.32, 1,206.31]	202.24 [201.92, 202.56]
RS-SupLin	KH	1,363.49 [1,341.00, 1,385.97]	48.68 [48.47, 48.89]
RS-SupLin	HF	1,347.79 [1,316.68, 1,378.91]	39.28 [38.95, 39.61]
SupSBUCB	KH	1,469.49 [1,446.95, 1,492.03]	1,029.64 [1,029.45, 1,029.83]
SupSBUCB	HF	1,453.79 [1,422.56, 1,485.03]	1,020.24 [1,019.92, 1,020.56]

N.3 Runtime and measurement scope

Tables 11 to 13 report online-loop wall seconds (mean $\pm$ SE and sample SD), mean process CPU seconds, and mean wall microseconds per round, using the same replications as the regret tables. Each method's compiled online loop is timed separately after warm-up over the full evaluation horizon.

The CPU-only environment has eight logical CPUs (four reported physical cores), x86_64 architecture, and approximately 51 GiB of RAM. Eight process workers handle independent paired streams, with one BLAS/OpenMP thread per worker. The numerical environment is Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, Numba 0.65.1, and llvmlite 0.47.0. All methods share each generated context chunk and oracle means within a stream, with float64 and disabled `fastmath`.

Timing includes action selection, authorized snapshot and prefix updates, private Gaussian probes, evaluator regret accumulation, and event-array bookkeeping. It excludes shared context/true-mean generation, representation fitting and other preprocessing, full-phase rotation search, JIT warm-up, initialization, compression, I/O, plotting, queue waiting, and disconnected time. Wall time includes concurrent scheduling effects and is distinct from process CPU time; this is neither an isolated single-core microbenchmark nor an end-to-end deployment benchmark.

RS-OFUL and B-TS-C have the two lowest mean wall times in every setting, followed by RarelySwitch-SupLinUCB, the two GATE versions, and SupSBUCB. The order between the GATE versions varies. Fewer policy updates therefore do not imply lower online runtime: GATE still evaluates frozen geometry and reached Gaussian tests. These timings describe this implementation and workload, not an asymptotic arithmetic bound or universal runtime ordering.

Table 11: Online-loop runtime for the original-scale synthetic panel ($T = 10{,}000$, $n = 50$). Wall seconds are means $\pm$ one SE, with sample SDs; CPU seconds and wall microseconds per round are means. Measurement scope is defined in Section N.3.

Method	Wall time (s) mean $\pm$ SE	SD (s)	CPU time (s) mean	Wall time μs/round
I.i.d., $(K, d) = (1000, 5)$*;* $n = 50$				
RS-OFUL	0.0855 ± 0.0009	0.0063	0.0834	8.55
B-TS-C	0.0871 ± 0.0010	0.0073	0.0835	8.71
RarelySwitch-SupLinUCB	1.1840 ± 0.0080	0.0563	1.1395	118.40
SupSBUCB	2.9927 ± 0.0157	0.1108	2.8659	299.27
GATE	1.6038 ± 0.0097	0.0684	1.5289	160.38
GATE (horizon-free)	1.5960 ± 0.0069	0.0490	1.5264	159.60
I.i.d., $(K, d) = (100, 30)$*;* $n = 50$				
RS-OFUL	0.0779 ± 0.0011	0.0076	0.0744	7.79
B-TS-C	0.0839 ± 0.0010	0.0070	0.0800	8.39
RarelySwitch-SupLinUCB	0.6183 ± 0.0039	0.0279	0.5957	61.83
SupSBUCB	1.0246 ± 0.0051	0.0363	0.9927	102.46
GATE	0.7618 ± 0.0036	0.0253	0.7418	76.18
GATE (horizon-free)	0.7616 ± 0.0039	0.0277	0.7403	76.16
Rotating, $(K, d) = (1000, 5)$*;* $n = 50$				
RS-OFUL	0.0888 ± 0.0011	0.0076	0.0855	8.88
B-TS-C	0.0913 ± 0.0011	0.0080	0.0859	9.13
RarelySwitch-SupLinUCB	1.1420 ± 0.0056	0.0397	1.0958	114.20
SupSBUCB	2.8705 ± 0.0141	0.0997	2.7459	287.05
GATE	1.5376 ± 0.0093	0.0661	1.4712	153.76
GATE (horizon-free)	1.5253 ± 0.0072	0.0510	1.4714	152.53
Rotating, $(K, d) = (100, 30)$*;* $n = 50$				
RS-OFUL	0.0709 ± 0.0007	0.0053	0.0689	7.09
B-TS-C	0.0845 ± 0.0010	0.0071	0.0813	8.45
RarelySwitch-SupLinUCB	0.6968 ± 0.0093	0.0657	0.6733	69.68
SupSBUCB	1.1907 ± 0.0137	0.0970	1.1568	119.07
GATE	0.7382 ± 0.0085	0.0603	0.7172	73.82
GATE (horizon-free)	0.7358 ± 0.0085	0.0601	0.7169	73.58

Table 12: Online-loop runtime for the long-horizon synthetic panel ($T = 2 \times 10^7$, $n = 20$). Wall seconds are means $\pm$ one SE, with sample SDs; CPU seconds and wall microseconds per round are means. Measurement scope is defined in Section N.3.

Method	Wall time (s) mean $\pm$ SE	SD (s)	CPU time (s) mean	Wall time μs/round
I.i.d., $(K, d) = (16, 4)$*;* $n = 20$				
RS-OFUL	17.801 ± 0.025	0.112	16.966	0.89
B-TS-C	22.069 ± 0.030	0.135	21.026	1.10
RarelySwitch-SupLinUCB	56.886 ± 0.058	0.261	54.282	2.84
SupSBUCB	332.959 ± 0.256	1.146	318.310	16.65
GATE	122.149 ± 0.156	0.699	116.675	6.11
GATE (horizon-free)	121.946 ± 0.135	0.603	116.575	6.10
I.i.d., $(K, d) = (128, 4)$*;* $n = 20$				
RS-OFUL	39.144 ± 0.307	1.373	37.290	1.96
B-TS-C	43.879 ± 0.368	1.645	41.767	2.19
RarelySwitch-SupLinUCB	296.683 ± 1.011	4.520	283.049	14.83
SupSBUCB	$2{,}211.005 \pm 7.026$	31.419	2,110.643	110.55
GATE	610.858 ± 2.172	9.713	582.992	30.54
GATE (horizon-free)	610.180 ± 2.212	9.891	582.335	30.51
Rotating, $(K, d) = (16, 4)$*;* $n = 20$				
RS-OFUL	19.366 ± 0.070	0.315	18.188	0.97
B-TS-C	24.028 ± 0.075	0.334	22.554	1.20
RarelySwitch-SupLinUCB	61.653 ± 0.102	0.456	58.020	3.08
SupSBUCB	351.133 ± 0.641	2.865	330.628	17.56
GATE	194.850 ± 0.325	1.454	183.340	9.74
GATE (horizon-free)	196.034 ± 0.381	1.704	184.727	9.80
Rotating, $(K, d) = (128, 4)$*;* $n = 20$				
RS-OFUL	38.096 ± 1.171	5.236	36.422	1.90
B-TS-C	42.438 ± 1.299	5.808	40.553	2.12
RarelySwitch-SupLinUCB	284.421 ± 10.308	46.101	272.268	14.22
SupSBUCB	$2{,}076.556 \pm 77.473$	346.468	1,987.909	103.83
GATE	$1{,}052.886 \pm 37.618$	168.232	1,008.065	52.64
GATE (horizon-free)	$1{,}069.610 \pm 38.611$	172.675	1,023.874	53.48

Table 13: Online-loop runtime for the real-covariate controls ($n = 50$). Wall seconds are means $\pm$ one SE, with sample SDs; CPU seconds and wall microseconds per round are means. Measurement scope is defined in Section N.3.

Method	Wall time (s) mean $\pm$ SE	SD (s)	CPU time (s) mean	Wall time μs/round
Jester, $T = 27{,}246$*,* $(K, d) = (20, 25)$*;* $n = 50$				
RS-OFUL	0.0861 ± 0.0009	0.0066	0.0824	3.16
B-TS-C	0.1176 ± 0.0013	0.0094	0.1119	4.32
RarelySwitch-SupLinUCB	0.3522 ± 0.0024	0.0167	0.3359	12.93
SupSBUCB	0.6502 ± 0.0046	0.0324	0.6218	23.86
GATE	0.5376 ± 0.0049	0.0344	0.5123	19.73
GATE (horizon-free)	0.5358 ± 0.0041	0.0288	0.5117	19.67
KuaiRec, $T = 44{,}782$*,* $(K, d) = (20, 25)$*;* $n = 50$				
RS-OFUL	0.1385 ± 0.0016	0.0112	0.1322	3.09
B-TS-C	0.1900 ± 0.0019	0.0132	0.1821	4.24
RarelySwitch-SupLinUCB	0.5824 ± 0.0052	0.0366	0.5614	13.01
SupSBUCB	1.4213 ± 0.0118	0.0836	1.3744	31.74
GATE	0.8625 ± 0.0079	0.0558	0.8277	19.26
GATE (horizon-free)	0.8653 ± 0.0075	0.0529	0.8338	19.32