# Direction-Aware Offline-to-Online Learning in Linear Contextual Bandits

**Zean Han, Ruihan Lin, Zezhen Ding*, Jiheng Zhang**
The Hong Kong University of Science and Technology
zhanax@connect.ust.hk, rlinah@connect.ust.hk
zdingah@connect.ust.hk, jiheng@ust.hk

## Abstract

Many bandit systems are deployed with offline historical data, such as past logs from earlier policies. Using these data can reduce early online exploration when they remain informative for the online problem. When the offline and online environments differ, such data can be biased for the online problem. For linear (contextual) bandits, this bias is directional: offline data may be informative in some feature directions and misleading in others. However, prior work typically controls this gap through a known Euclidean bound on the model parameters, which we prove is too coarse: even with the offline parameter known, bias in a single unknown direction can force dimension-dependent regret. To address this challenge, we introduce a directional bias certificate $(M_{\mathrm{bias}}, \rho)$ that measures the offline-to-online gap through an $M_{\mathrm{bias}}$-induced norm and assigns different bias budgets to different directions. Building on this certificate, we propose *Ellipsoidal-MINUCB*, which augments the online learning with an offline-pooled branch that safely exploits historical data. When the certificate is known, we show that the algorithm matches the standard SupLinUCB rate in the worst case and improves when offline coverage aligns with low-bias directions. When the certificate is unknown, we estimate it adaptively from offline and accumulated online data and establish a corresponding regret guarantee. Numerical experiments support the theory and show gains in aligned regimes.

## 1 Introduction

Many bandit systems are deployed with historical data already in hand, such as past logs or data from related environments. A natural goal is to use these data to reduce early online exploration. If the offline data remain informative in the live environment, they can give the learner a useful starting point and speed up online learning. This is the promise of offline-to-online learning [Zhang et al., 2019, Tennenholtz et al., 2021, Vijayan et al., 2025].

The difficulty is that the offline and online environments need not match. The offline data may have been collected under a different policy, population, or feature distribution, so the offline regression parameter $\theta^\dagger$ need not equal the online reward parameter $\theta_*$, and the offline estimate of rewards can be biased for the online problem. In linear contextual bandits, this mismatch is directional. Actions share a linear feature representation, so offline error is not equally important in all directions. The offline estimate may be helpful along some directions and misleading along others, and what matters is whether the bias lies in directions that affect online comparisons between actions.

Prior work often models offline-to-online mismatch through a single scalar bound, typically a Euclidean bound $\|\theta_* - \theta^\dagger\|_2 \le \rho$ with a known radius $\rho$ [Cheung and Lyu, 2024, Zhou et al., 2025,

*Corresponding author: zdingah@connect.ust.hk.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

He et al., 2025, Zhou et al., 2026]. We show that this is fundamentally too coarse. Such a bound controls only the total size of the mismatch; it does not say whether the mismatch lies in directions that matter for online decisions. Our impossibility result (Theorem 3) makes this precise: even an oracle given the offline parameter $\theta^\dagger$ in advance must suffer worst-case regret of order $\Omega(\min\{\rho T, \sqrt{dT}\})$, because the same scalar bound can describe an easy instance, where offline information already identifies the best action, and a hard instance, where substantial online exploration is still necessary. This raises the following question:

*How should one measure offline-to-online bias in a way that captures the directions that matter for online decisions, and how can such a measure guide algorithm design?*

We answer this question through four contributions:

**Directional bias certificate:** We introduce a directional way to control offline-to-online mismatch:

$$\|\theta_* - \theta^\dagger\|_{M_{\text{bias}}} \le \rho,$$

where $M_{\text{bias}} \succ 0$ determines how much mismatch is allowed across directions and $\rho \ge 0$ is the corresponding radius. This pair turns qualitative information about which directions are stable and which are unreliable into quantitative, direction-wise bias budgets. We call such a pair $(M_{\text{bias}}, \rho)$ a *directional bias certificate*, since it certifies how much offline-to-online bias is allowed along each direction. The intuition behind the certificate is given in Appendix A.

**Algorithm design:** Using this certificate, we design *Ellipsoidal-MINUCB*, which scores each action by the minimum of two upper confidence bounds: a purely online one and an offline-informed one built from the certificate. We use the offline-informed bound only when it is tighter, so misleading offline information cannot remove the usual online fallback. We prove regret bounds showing that the algorithm matches the standard SupLinUCB rate in the worst case, with the offline-informed term improving when offline coverage and directional bias are aligned with the directions that drive online decisions.

**Adaptive estimation of the directional bias certificate:** We also study the case where the directional bias certificate is not known before deployment. We propose an adaptive construction of the certificate which updates epoch by epoch from the offline samples together with the online observations collected so far, and we establish a corresponding regret guarantee. This extends the framework to settings where directional bias information is not available in advance, which largely enhances the practicality of our methods.

**Empirical validation:** We complement the theory with numerical simulations comparing our algorithms against online, warm-started, and offline-only baselines. The results confirm the predicted pattern: gains are largest when offline coverage aligns with low-bias directions, and the algorithm falls back gracefully to the online baseline otherwise.

## 2 Problem Setup

We specify the online bandit model, offline regression data, and a directional bias certificate, then record an offline confidence region for $\theta_*$ used in Section 3.

Let $\mathcal{F}_{t-1}$ be the $\sigma$-algebra generated by $(x_{s,\cdot}, a_s, r_s)_{s<t}$ and the current contexts $\{x_{t,a}\}_{a\in[K]}$. We consider a stochastic linear contextual bandit over horizon $T$.

**Assumption 1** (Linear reward with bounded features)**.** *At each round $t \in [T]$, the learner observes a finite action set $\{x_{t,a}\}_{a\in[K]} \subset \mathbb{R}^d$ with $\|x_{t,a}\|_2 \le 1$, selects $a_t$, and receives $r_t = x_{t,a_t}^\top \theta_* + \eta_t \in [0,1]$, where $\theta_* \in \mathbb{R}^d$ is unknown with $\|\theta_*\|_2 \le S$. The action set at round $t$ is revealed before the reward is observed, so $x_{t,a_t}$ and the SupLinUCB layer assignment (Section 3) are $\mathcal{F}_{t-1}$-measurable.*

**Assumption 2** (Conditionally sub-Gaussian online noise)**.** $\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}] = 0$ *and $\eta_t$ is conditionally $\sigma$-sub-Gaussian given $\mathcal{F}_{t-1}$.*

The cumulative regret is

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

Before the online phase, the learner observes offline regression samples $D_{\text{off}} = \{(z_i, y_i)\}_{i=1}^{n_{\text{off}}}$. The difficulty is that the offline parameter $\theta^\dagger$ need not equal the online parameter $\theta_*$.

**Assumption 3** (Offline regression model)**.** *$y_i = z_i^\top \theta^\dagger + \xi_i$ with $\|z_i\|_2 \le 1$, $\|\theta^\dagger\|_2 \le S$, and $\mathbb{E}[\xi_i \mid z_i] = 0$. The offline covariates $\{z_i\}$ are non-adaptive (fixed before the online phase), and $\xi_i$ is conditionally $\sigma$-sub-Gaussian given $z_i$.*

**Definition 1** (Directional bias certificate)**.** *Let $M_{\text{bias}} \in \mathbb{S}_{++}^d$ and $\rho \ge 0$ be known. The pair $(M_{\text{bias}}, \rho)$ is called a valid directional bias certificate for $(\theta_*, \theta^\dagger)$ if*

$$\|\theta_* - \theta^\dagger\|_{M_{\text{bias}}} \le \rho.$$

*Equivalently, every direction $x \in \mathbb{R}^d$ satisfies*

$$|x^\top(\theta_* - \theta^\dagger)| \le \rho \|x\|_{M_{\text{bias}}^{-1}}.$$

**Remark 1.** In linear contextual bandits, rewards depend on actions only through features, so the natural object to control is parameter error *along the directions* that enter UCB/LCB scores. A valid $(M_{\text{bias}}, \rho)$ is side information that encodes this anisotropy: it implies per-direction bounds with penalties $\rho\|x\|_{M_{\text{bias}}^{-1}}$ matching the linear optimistic template, and it is not identified from bandit feedback alone without additional structure. Intuitively, a large eigenvalue of $M_{\text{bias}}$ means that mismatch is tightly constrained along the corresponding eigenvector, while a small eigenvalue allows a larger offline-to-online parameter difference in that direction. A detailed explanation is given in Appendix A. By contrast, a Euclidean bound $\left\|\theta_* - \theta^\dagger\right\|_2 \le \rho$ controls only the *magnitude* of the bias and can hide whether the bias lies along a regret-relevant direction. Appendix B gives a finite-action Bernoulli specialization in which three instances share the same offline parameter and offline data law and satisfy the same scalar radius; in one case a fixed action is optimal throughout, while in a paired hard case the learner must identify the sign of a hidden bias and every policy suffers regret of order $\Omega(\rho^2 T/\rho_0)$ on one of the two instances.

To use this side information algorithmically, we combine it with the offline ridge estimate. The offline samples yield a confidence ellipsoid around $\theta^\dagger$. Together with the directional bias certificate, Proposition 1 fuses the two into a single confidence region for $\theta_*$.

The offline samples induce the usual ridge objects

$$G_{\text{off}} := I + \sum_{i=1}^{n_{\text{off}}} z_i z_i^\top, \qquad \hat{\theta}_{\text{off}} := G_{\text{off}}^{-1} \sum_{i=1}^{n_{\text{off}}} z_i y_i.$$

Here $G_{\text{off}}$ records offline coverage, while $\hat{\theta}_{\text{off}}$ estimates the offline parameter $\theta^\dagger$. The ridge concentration inequality controls $\hat{\theta}_{\text{off}} - \theta^\dagger$ in the $G_{\text{off}}$ metric. Then the directional bias certificate lifts this offline ellipsoid to a confidence region for $\theta_*$.

**Proposition 1** (Offline-to-online confidence region)**.** *For*

$$\beta_{\text{off}}(\delta) := \sigma\sqrt{2\log\frac{\det(G_{\text{off}})^{1/2}}{\delta}} + S,$$

*with probability at least $1 - \delta$,*

$$\theta_* \in \left\{\hat{\theta}_{\text{off}} + u + v : \|u\|_{G_{\text{off}}} \le \beta_{\text{off}}(\delta),\ \|v\|_{M_{\text{bias}}} \le \rho\right\} =: \mathcal{C}_{\text{off}}.$$

*Consequently, the support function of this region is*

$$\sup_{\theta \in \mathcal{C}_{\text{off}}} x^\top \theta = x^\top \hat{\theta}_{\text{off}} + \beta_{\text{off}}(\delta)\|x\|_{G_{\text{off}}^{-1}} + \rho\|x\|_{M_{\text{bias}}^{-1}}.$$

## 3 Ellipsoidal-MINUCB

So far, $\theta_*$ is described only through the offline-informed region above. Online play requires layerwise scores that update with streaming data. The next paragraphs define those scores. Fix an offline confidence budget $\delta_{\text{off}}$ and write $\beta_{\text{off}} := \beta_{\text{off}}(\delta_{\text{off}})$ throughout this section.

**Online branch.** For each layer $\ell$, define

$$A_{t,\ell} := \sum_{s \in \mathcal{H}_{t,\ell}} x_{s,a_s} x_{s,a_s}^\top, \qquad V_{t,\ell} := I + A_{t,\ell}, \qquad \hat{\theta}^{\mathrm{on}}_{t,\ell} := V_{t,\ell}^{-1} b_{t,\ell}, \qquad b_{t,\ell} := \sum_{s \in \mathcal{H}_{t,\ell}} x_{s,a_s} r_s. \tag{1}$$

The online UCB is

$$U^{\mathrm{on}}_{t,\ell}(a) = x_{t,a}^\top \hat{\theta}^{\mathrm{on}}_{t,\ell} + \mathrm{rad}^{\mathrm{on}}_{t,\ell}(x), \tag{2}$$

where $\mathrm{rad}^{\mathrm{on}}_{t,\ell}(x) := \beta^{\mathrm{on}}_{t,\ell} \|x\|_{V_{t,\ell}^{-1}}$ with the corresponding online LCB $L^{\mathrm{on}}_{t,\ell}(a)$ defined symmetrically, and

$$\beta^{\mathrm{on}}_{t,\ell} := \sigma \sqrt{2 \log \frac{\det(V_{t,\ell})^{1/2}}{\delta^{\mathrm{on}}_{t,\ell}}} + S$$

is the layerwise ridge confidence radius for budget $\delta^{\mathrm{on}}_{t,\ell}$.

**Pooled branch.** The offline-informed estimator is

$$\hat{\theta}^{\mathrm{pool}}_{t,\ell} := (A_{t,\ell} + G_{\mathrm{off}})^{-1} \big(b_{t,\ell} + G_{\mathrm{off}} \hat{\theta}_{\mathrm{off}}\big). \tag{3}$$

The confidence interval radius is

$$\mathrm{rad}^{\mathrm{pool}}_{t,\ell}(x) := \big(\gamma_{t,\ell} + \beta_{\mathrm{off}}\big) \|x\|_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}} + \rho\, \psi_{t,\ell}(x). \tag{4}$$

where

$$\psi_{t,\ell}(x) := \left\| M_{\mathrm{bias}}^{-1/2} G_{\mathrm{off}} (A_{t,\ell} + G_{\mathrm{off}})^{-1} x \right\|_2. \tag{5}$$

The pooled martingale radius is

$$\gamma_{t,\ell} := \sigma \sqrt{2 \log \frac{\det(A_{t,\ell} + G_{\mathrm{off}})^{1/2}}{\det(G_{\mathrm{off}})^{1/2} \delta^{\mathrm{pool}}_{t,\ell}}}. \tag{6}$$

Although $\gamma_{t,\ell}$ appears in the pooled branch, it controls only the online martingale noise measured in the pooled geometry $A_{t,\ell} + G_{\mathrm{off}}$; the offline ridge uncertainty is carried by the separate $\beta_{\mathrm{off}}$ term, and no additional $+S$ regularization slack is needed in this branch. The pooled UCB is

$$U^{\mathrm{pool}}_{t,\ell}(a) = x_{t,a}^\top \hat{\theta}^{\mathrm{pool}}_{t,\ell} + \mathrm{rad}^{\mathrm{pool}}_{t,\ell}(x_{t,a}), \tag{7}$$

with the corresponding pooled LCB $L^{\mathrm{pool}}_{t,\ell}(a)$ defined symmetrically.

**Remark 2.** Pooling replaces $\|x\|_{V_{t,\ell}^{-1}}$ by $\|x\|_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}}$ when offline data is informative along $x$. The term $\rho\, \psi_{t,\ell}(x)$ depends on two factors: the pooling map $G_{\mathrm{off}}(A_{t,\ell} + G_{\mathrm{off}})^{-1}$ and the bias geometry $M_{\mathrm{bias}}$. Along $x$, offline data helps only if pooling makes the radius tighter than the purely online radius when $\rho\, \psi_{t,\ell}(x)$ is included.

**Branch aggregation.** The final aggregated UCB and LCB are

$$U^{\min}_{t,\ell}(a) := \min\{U^{\mathrm{on}}_{t,\ell}(a), U^{\mathrm{pool}}_{t,\ell}(a)\}, \qquad L^{\max}_{t,\ell}(a) := \max\{L^{\mathrm{on}}_{t,\ell}(a), L^{\mathrm{pool}}_{t,\ell}(a)\}. \tag{8}$$

This design makes the safety mechanism explicit: the pooled branch can only help if it is tighter, and the online branch remains available when the directional bias certificate is weak or the queried direction is poorly covered offline.

Let

$$w_{t,\ell}(a) := U^{\min}_{t,\ell}(a) - L^{\max}_{t,\ell}(a) \tag{9}$$

denote the aggregated interval width. The full round procedure is given in Algorithm 1.

The regret analysis of Algorithm 1 relies on two technical points. First, online data are partitioned across layers in SupLinUCB, so the ridge estimator at each layer uses a random subsequence of online rewards rather than the full trajectory. We handle these adaptive subsequences by viewing each layer's history through predictable masks and applying a masked self-normalized martingale bound. Second, the pooled estimator merges online Gram information $A_{t,\ell}$ with the offline design $G_{\mathrm{off}}$. For each direction $x$, we control $x^\top \hat{\theta}^{\mathrm{pool}}_{t,\ell} - x^\top \theta_*$ by summarizing online and offline estimation errors around $\theta^\dagger$ with the elliptic norm $\|x\|_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}}$ at radius scale $\gamma_{t,\ell} + \beta_{\mathrm{off}}$. The offline-to-online bias $\theta^\dagger - \theta_*$ is treated differently: under the directional bias certificate it contributes $\rho\, \psi_{t,\ell}(x)$, so extra uncertainty from bias depends on the direction instead of being treated uniformly.

**Algorithm 1** Ellipsoidal-MINUCB

**Input:** Offline data $D_{\mathrm{off}}$, directional bias certificate $(M_{\mathrm{bias}}, \rho)$, horizon $T$, offline confidence budget $\delta_{\mathrm{off}}$, online confidence schedule $\{\delta_{t,\ell}^{\mathrm{on}}\}_{t\le T,\ell\le\lceil\log_2 T\rceil}$, and pooled confidence schedule $\{\delta_{t,\ell}^{\mathrm{pool}}\}_{t\le T,\ell\le\lceil\log_2 T\rceil}$

Compute $G_{\mathrm{off}}$, $\hat{\theta}_{\mathrm{off}}$, and $\beta_{\mathrm{off}} := \beta_{\mathrm{off}}(\delta_{\mathrm{off}})$ from $D_{\mathrm{off}}$
$L \leftarrow \lceil \log_2 T \rceil$
Initialize $\mathcal{H}_{1,\ell} \leftarrow \varnothing$ for all $\ell = 0, 1, \ldots, L$
**for** $t = 1, 2, \ldots, T$ **do**
  Set $\mathcal{H}_{t+1,\ell} \leftarrow \mathcal{H}_{t,\ell}$ for every $\ell = 0, 1, \ldots, L$
  Observe the action set $\{x_{t,a}\}_{a\in[K]}$
  $\mathcal{A}_{t,0} \leftarrow [K]$
  **for** $\ell = 0, 1, \ldots, L$ **do**
    Build $A_{t,\ell}$, $V_{t,\ell}$, and $b_{t,\ell}$ from (1)
    Compute online scores by (2)
    Compute pooled scores by (3)–(7)
    Compute aggregated widths by (8)–(9)
    **for** each $a \in \mathcal{A}_{t,\ell}$ **do**
      Compute aggregated UCB $U_{t,\ell}^{\min}(a)$, aggregated LCB $L_{t,\ell}^{\max}(a)$, and $w_{t,\ell}(a)$
    **end for**
    **if** $\ell = L$ **then**
      Play $a_t \in \arg\max_{a\in\mathcal{A}_{t,\ell}} U_{t,\ell}^{\min}(a)$
      Observe reward $r_t$ and append $(x_{t,a_t}, r_t)$ to $\mathcal{H}_{t+1,\ell}$
      Record $\ell_t \leftarrow \ell$ and break the layer loop
    **else if** $\max_{a\in\mathcal{A}_{t,\ell}} w_{t,\ell}(a) > 2^{-\ell}$ **then**
      Play $a_t \in \arg\max_{a\in\mathcal{A}_{t,\ell}: w_{t,\ell}(a)>2^{-\ell}} U_{t,\ell}^{\min}(a)$
      Observe reward $r_t$ and append $(x_{t,a_t}, r_t)$ to $\mathcal{H}_{t+1,\ell}$
      Record $\ell_t \leftarrow \ell$ and break the layer loop
    **else**
      $m_{t,\ell} \leftarrow \max_{a\in\mathcal{A}_{t,\ell}} L_{t,\ell}^{\max}(a)$
      $\mathcal{A}_{t,\ell+1} \leftarrow \{a \in \mathcal{A}_{t,\ell} : U_{t,\ell}^{\min}(a) \ge m_{t,\ell}\}$
    **end if**
  **end for**
**end for**

## 4 Main Results

The next theorem states our main high-probability regret bound. Corollaries and structural comparisons follow in the same section.

**Theorem 1** (High-probability guarantee)**.** *Write $L := \lceil \log_2 T \rceil$. Run Ellipsoidal-MINUCB with confidence budgets $\delta_{\mathrm{off}}, \delta_{\mathrm{on}}, \{\delta_{t,\ell}^{\mathrm{on}}\}, \{\delta_{t,\ell}^{\mathrm{pool}}\}$ satisfying*

$$\sum_{t=1}^{T}\sum_{\ell=0}^{L} \delta_{t,\ell}^{\mathrm{on}} \le \delta_{\mathrm{on}}, \qquad \delta_{\mathrm{off}} + \delta_{\mathrm{on}} + \sum_{t=1}^{T}\sum_{\ell=0}^{L} \delta_{t,\ell}^{\mathrm{pool}} \le T^{-2}.$$

*Then, with probability at least $1 - T^{-2}$,*

$$R_T \le \min\Big\{ C_{\mathrm{SL}} \sqrt{T d\, L_T^3},\ 8\,\Gamma_{\mathrm{pool}} \sqrt{2T\,\widetilde{\Lambda}_T} + 8\rho\,\widetilde{\Psi}_T \Big\},$$

*where $\widetilde{\Lambda}_T := \sum_{\ell=0}^{L} \Lambda_\ell$ aggregates layerwise pooled log-determinant increments $(\Lambda_\ell)$ built from $(Q_\ell, A_\ell^{\mathrm{fin}})$ defined in Appendix H, $\widetilde{\Psi}_T := \sum_{t=1}^{T} \psi_{t,\ell_t}(x_{t,a_t})$, $L_T := \log(2KT\log T/\delta_{\mathrm{on}})$, $\beta_{\mathrm{off}} := \beta_{\mathrm{off}}(\delta_{\mathrm{off}})$, $\delta_{\min}^{\mathrm{pool}} := \min_{t,\ell} \delta_{t,\ell}^{\mathrm{pool}}$, $\Lambda_{\max} := \max_{0\le\ell\le L} \Lambda_\ell$, and $\Gamma_{\mathrm{pool}} := \beta_{\mathrm{off}} + \sigma\sqrt{\Lambda_{\max} + 2\log(1/\delta_{\min}^{\mathrm{pool}})}$.*

Theorem 1 reports whichever is smaller: the purely online regret, or the regret of the pooled branch. For the pooled branch regret, $\widetilde{\Lambda}_T$ tracks how much information comes from the pooled branch, while

$\widetilde{\Psi}_T$ tracks how much we must widen confidence intervals because offline estimates may be biased along directions that appear online. The prefactor $\Gamma_{\mathrm{pool}}$ uses the largest pooled log-det increment $\Lambda_{\max}$ instead of an ambient $\sqrt{d}$ term. We leave the proofs in Appendix H.

However, the above regret bound relies on the trajectory-dependent $\widetilde{\Lambda}_T$. To further simplify the bound, we introduce the following definition.

**Definition 2.** *Let $G \succ 0$ and $g_1, \ldots, g_d$ be its eigenvalues and let $B \geq 0$. Define*

$$\Phi_G(B) := \max_{\substack{a_j \geq 0 \\ \sum_{j=1}^d a_j \leq B}} \sum_{j=1}^d \log\left(1 + \frac{a_j}{g_j}\right).$$

*Equivalently, let $\tau$ solve $\sum_{j=1}^d (\tau - g_j)_+ = B$. Then $\Phi_G(B) = \sum_{j:\, g_j < \tau} \log(\tau/g_j)$.*

**Theorem 2** (Spectral description for pooled regret)**.** *Let $L_{\max} := \lceil \log_2 T \rceil + 1$. Along any realized trajectory,*

$$\widetilde{\Lambda}_T \leq L_{\max}\, \Phi_{G_{\mathrm{off}}}\left(\frac{T}{L_{\max}}\right).$$

*Denote $c_{\mathrm{align}} := \lambda_{\max}\Big(G_{\mathrm{off}}^{1/2} M_{\mathrm{bias}}^{-1} G_{\mathrm{off}}^{1/2}\Big)$ to be the largest eigenvalue, then on the event of Theorem 1, the pooled branch satisfies, for $\Gamma_{\mathrm{spec}} := \beta_{\mathrm{off}} + \sigma\sqrt{\Phi_{G_{\mathrm{off}}}(T) + 2\log(1/\delta_{\min}^{\mathrm{pool}})}$,*

$$R_T \leq 8\big(\Gamma_{\mathrm{spec}} + \rho\sqrt{c_{\mathrm{align}}}\big)\sqrt{2T\, L_{\max}\, \Phi_{G_{\mathrm{off}}}\left(\frac{T}{L_{\max}}\right)}.$$

*The full algorithm retains the online fallback $R_T \leq C_{\mathrm{SL}}\sqrt{Td\, L_T^3}$.*

Next, we give a specific regime in which the offline data are informative enough to reduce the online regret.

**Corollary 1** (Explicit pooled-better regime)**.** *On the event of Theorem 1, the pooled branch satisfies*

$$R_T \leq 8\big(\Gamma_{\mathrm{pool}} + \rho\sqrt{c_{\mathrm{align}}}\big)\sqrt{2T\,\widetilde{\Lambda}_T}.$$

*Consequently, if*

$$\widetilde{\Lambda}_T\big(\Gamma_{\mathrm{pool}} + \rho\sqrt{c_{\mathrm{align}}}\big)^2 \leq \frac{C_{\mathrm{SL}}^2}{128}\, d\, L_T^3,$$

*then the pooled term in Theorem 1 is no larger than the online fallback.*

**Remark 3.** Theorem 2 provides a deterministic regret bound. Offline gain enters only through the spectrum of $G_{\mathrm{off}}$ inside $\Phi_{G_{\mathrm{off}}}$, while the interaction of offline precision with the bias certificate is summarized by the single generalized-eigenvalue scale $c_{\mathrm{align}}$. Equivalently, $c_{\mathrm{align}}$ is the smallest $c \geq 0$ such that $G_{\mathrm{off}} M_{\mathrm{bias}}^{-1} G_{\mathrm{off}} \preceq c\, G_{\mathrm{off}}$. Corollary 1 makes the improvement regime explicit: transfer is favorable whenever the residual pooled complexity $\widetilde{\Lambda}_T$ and the pooled prefactor $\Gamma_{\mathrm{pool}} + \rho\sqrt{c_{\mathrm{align}}}$ are jointly much smaller than the ambient online benchmark $dL_T^3$. In the zero-bias case $\rho = 0$, if only $r$ directions remain weak after offline preconditioning in the sense that $\widetilde{\Lambda}_T = O(r)$, $\Lambda_{\max} = O(1)$, and $\beta_{\mathrm{off}} = O(1)$ along a family of instances, then the pooled branch scales as $O(\sqrt{rT})$ up to logarithmic factors, while the online fallback remains $O(\sqrt{dT})$ up to logarithmic factors. If the directional bias certificate is conservative but valid, the pooled term merely widens and the method falls back more often to the online branch. If the radius is undersized so that the true bias is excluded, then the pooled branch should be read as a heuristic warm start rather than as part of the theorem's guarantee.

**Theorem 3.** *Consider the diagonal Gaussian subclass with actions $\{e_1, \ldots, e_d\}$, arm-wise offline counts $n_i$, and directional bias budgets $v_i := \rho/\sqrt{b_i}$, equivalently $G_{\mathrm{off}} = \mathrm{diag}(1 + n_1, \ldots, 1 + n_d)$ and $M_{\mathrm{bias}} = \mathrm{diag}(b_1, \ldots, b_d)$. For $S \subset [d]$, write $N(S) := \sum_{i \in S} n_i$, $v(S) := \min_{i \in S} v_i$, and*

$$\Psi(S) := \sqrt{\frac{|S|T^2}{T + N(S)}} + \min\left\{\frac{N(S)T}{T + N(S)}\, v(S),\ \sqrt{\frac{|S|N(S)T}{T + N(S)}}\right\}.$$

*Then there exist universal constants $c > 0$ and $m_0 \in \mathbb{N}$ such that, for every $d \geq m_0$,*

$$\inf_{\pi} \sup_{I \in \mathcal{H}_{\mathrm{diag}}} \mathbb{E}_I[R_T(\pi)] \geq c \max_{\substack{S \subset [d] \\ |S| \geq m_0}} \min\left\{\sqrt{dT},\, \Psi(S)\right\}.$$

**Remark 4.** The lower bound carries the important information: transfer hardness is governed by subset-level offline coverage $N(S)$ and directional bias budgets $v(S)$, rather than by scalar summaries alone. Together, the upper and lower bounds show that offline-to-online transfer in linear contextual bandits is inherently directional. The proof is deferred to Appendix J.

## 5 Data-Driven Directional Bias Bounds

When a valid fixed directional bias certificate $(M_{\mathrm{bias}}, \rho)$ is unavailable, how can the learner safely use the offline data? Our answer is to learn a conservative epoch-wise surrogate certificate that can be plugged into the same algorithm while preserving the online learning safety.

The construction of such certificate combines an offline localization of $\theta^\dagger$ with an online localization of $\theta_*$ from historical data. Their estimate gap provides a data-dependent center, while the parallel-sum geometry records the precision shared by the two designs. Adding the corresponding ridge confidence widths then yields a conservative certificate for $\theta_* - \theta^\dagger$. Details and proofs are deferred to Appendices K.

**Epoch construction.** Fix $1 = \tau_1 < \cdots < \tau_{K_{\mathrm{ep}}} \leq T$ with $\tau_{K_{\mathrm{ep}}+1} := T + 1$. For each $k$, before round $\tau_k$, from past *online* rounds only, define

$$G_k^{\mathrm{on}} := I + \sum_{s < \tau_k} x_{s,a_s} x_{s,a_s}^\top, \qquad \hat{\theta}_k^{\mathrm{on}} := (G_k^{\mathrm{on}})^{-1} \sum_{s < \tau_k} x_{s,a_s} r_s.$$

The epoch directional bias certificate sets $\widehat{M}_{k,\mathrm{bias}}$ to the *parallel sum* of $G_k^{\mathrm{on}}$ and $G_{\mathrm{off}}$, which is $\widehat{M}_{k,\mathrm{bias}} := \left((G_k^{\mathrm{on}})^{-1} + G_{\mathrm{off}}^{-1}\right)^{-1}$, and the radius

$$\widehat{\rho}_k := \|\hat{\theta}_k^{\mathrm{on}} - \hat{\theta}_{\mathrm{off}}\|_{\widehat{M}_{k,\mathrm{bias}}} + \beta_k^{\mathrm{on}}(\delta_k) + \beta_{\mathrm{off}}(\delta_{\mathrm{off}}),$$

where $\beta_k^{\mathrm{on}}(\delta_k)$ is any valid ridge self-normalized bound for $\theta_*$ in the $G_k^{\mathrm{on}}$-norm.

**Remark 5.** The parallel sum is canonical here. For any vector $x$, the quadratic form induced by $\widehat{M}_{k,\mathrm{bias}}$ is the minimum joint precision cost of decomposing $x$ into an online-certifiable component and an offline-certifiable component:

$$\|x\|^2_{\widehat{M}_{k,\mathrm{bias}}} = \inf_{x=u+v} \left\{\|u\|^2_{G_k^{\mathrm{on}}} + \|v\|^2_{G_{\mathrm{off}}}\right\}.$$

Thus $\widehat{M}_{k,\mathrm{bias}}$ measures how cheaply a candidate bias vector can be certified simultaneously by the online and offline confidence geometries. The above variational identity is proved in Appendix K.1.

Rounds $t \in \{\tau_k, \ldots, \tau_{k+1} - 1\}$ use the layerwise branch aggregation of Section 3, replacing the fixed-bound map $\psi_{t,\ell}$ by $\widehat{\psi}_{k,t,\ell}$ in (11) and $\mathrm{rad}_{t,\ell}^{\mathrm{pool}}$ by $\widehat{\mathrm{rad}}_{t,\ell}^{\mathrm{pool,ep}}$ in (12) of Appendix K.1. The branch aggregation structure is otherwise unchanged, so the online branch remains a safe fallback whenever the learned pooled term is loose. The next theorem records the resulting regret guarantee. The full epoch-wise pseudocode is deferred to Appendix K.1 (Algorithm 2).

**Theorem 4** (Regret with epoch-wise learned directional bias certificates)**.** *Run Algorithm 2. Let $k(t)$ be the epoch index satisfying $t \in \{\tau_{k(t)}, \ldots, \tau_{k(t)+1} - 1\}$, and define the routed path sum*

$$\widetilde{\Psi}_T^{\mathrm{ep}} := \sum_{t=1}^{T} \widehat{\rho}_{k(t)}\, \widehat{\psi}_{k(t),t,\ell_t}(x_{t,a_t}).$$

*Then, with probability at least $1 - T^{-2} - \delta_{\mathrm{bias}}$,*

$$R_T \leq \min\left\{C_{\mathrm{SL}}\sqrt{Td\,L_T^3},\; 8 \sum_{\ell=0}^{\lceil \log_2 T \rceil} \left(\beta_{\mathrm{off}} + \gamma_\ell^{\mathrm{fin}}\right)\sqrt{2|Q_\ell|\,\Lambda_\ell} + 8\,\widetilde{\Psi}_T^{\mathrm{ep}}\right\},$$

*where $Q_\ell$, $\Lambda_\ell$, and $\gamma_\ell^{\mathrm{fin}}$ are as in Appendix H. In particular,*

$$R_T \le \min\left\{ C_{\mathrm{SL}}\sqrt{Td\,L_T^3},\ 8\Gamma_{\mathrm{pool}}\sqrt{2T\,\widetilde{\Lambda}_T} + 8\,\widetilde{\Psi}_T^{\mathrm{ep}} \right\},$$

*where $\Gamma_{\mathrm{pool}}$ is as in Theorem 1.*

Theorem 4 preserves the same min-structure as the fixed-certificate result. When the learned certificate is informative, the pooled branch improves on the online branch. When it is loose, the pooled term widens and the guarantee falls back toward the purely online rate. To obtain a closed-form refinement, we specialize to a standard doubling schedule. Since this schedule is only a default instantiation, we defer its precise definition and the associated confidence split to Appendix K.1 and state here only the resulting corollary.

**Corollary 2** (Adaptivity penalty under doubling epochs)**.** *Under the doubling schedule and confidence split of Theorem 7, define*

$$\widehat{c}_k := \lambda_{\max}\big(G_{\mathrm{off}}^{1/2}\widehat{M}_{k,\mathrm{bias}}^{-1}G_{\mathrm{off}}^{1/2}\big), \qquad \mathcal{S}_{\mathrm{ep}} := \sum_{k=1}^{K_{\mathrm{ep}}}(\tau_{k+1}-\tau_k)\,\widehat{\rho}_k^2\,\widehat{c}_k.$$

*Then, with probability at least $1 - T^{-2} - \delta_{\mathrm{bias}}$,*

$$R_T \le \min\left\{ C_{\mathrm{SL}}\sqrt{Td\,L_T^3},\ 8\Big(\Gamma_{\mathrm{pool}} + \sqrt{\mathcal{S}_{\mathrm{ep}}/T}\Big)\sqrt{2T\,\widetilde{\Lambda}_T} \right\}.$$

**Remark 6.** Corollary 2 isolates the price of learning the certificate from data. The fixed-certificate pooled branch pays the deterministic prefactor $\Gamma_{\mathrm{pool}} + \rho\sqrt{c_{\mathrm{align}}}$, whereas the learned version pays $\Gamma_{\mathrm{pool}} + \sqrt{\mathcal{S}_{\mathrm{ep}}/T}$. Thus the learned rule approaches fixed-certificate behavior when the epoch-wise estimator gap stabilizes quickly and the learned geometries $\widehat{M}_{k,\mathrm{bias}}$ stay well aligned with $G_{\mathrm{off}}$, so that the average penalty $\mathcal{S}_{\mathrm{ep}}/T$ is small. Conversely, persistent disagreement between online and offline estimates in highly covered directions enlarges $\widehat{\rho}_k$ or $\widehat{c}_k$, widens the learned pooled term, and drives the method back toward the online fallback.

## 6 Experiments

We visualize how regret responds to increasing coordinate-split mismatch between $\theta_*$ and $\theta^\dagger$ under randomized contexts. The main experiment uses $K{=}5$ arms and $d{=}5$ features with normalized random contexts. We fix $\theta_* = (s, 1, 1, 1, 1)$ and $\theta^\dagger = (1, s, 1, 1, 1)$ with $s \in \{1.1, 2, 10\}$ in the three panels of Figure 1. These regimes move from mild to severe coordinate-split mismatch: the online parameter puts extra mass on the first coordinate while the offline parameter puts matching mass on the second.

We compare a standard SupLinUCB baseline, Ellipsoidal-MINUCB, 3D Ellipsoidal-MINUCB(Algorithm 2), a warm-started SupLinUCB, and a non-adaptive policy that uses only the offline estimate. Full hyperparameters and implementation details are in Appendix L.

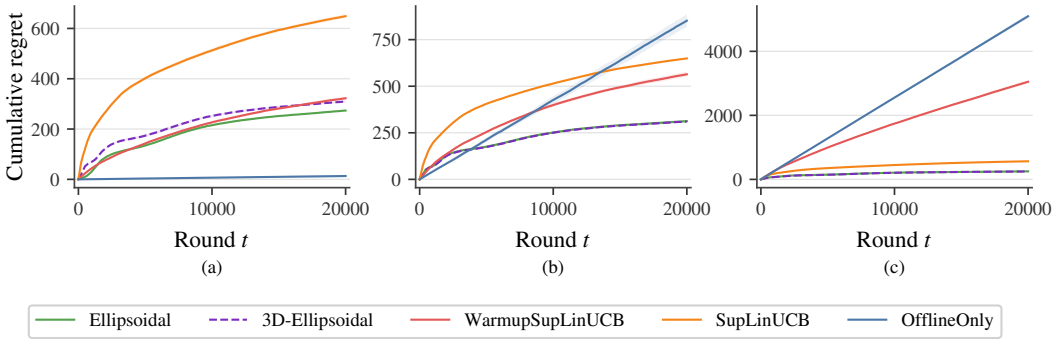



Figure 1: Cumulative regret curves

Figure 1 shows that larger $s$ increases mismatch between the highlighted coordinates. The ordering of curves is consistent with the min-structure in Theorem 1, and the adaptive variant tracks the same phases with additional conservatism.

**Additional experiments.** Appendix L reports three additional experiments: randomized bias generation that trades a multiplicative component against an additive offset; fixed $\theta_*$ with varying alignment of the offline mean; and *offline* covariate support held strictly inside the *online* support, so that the offline design matrix is rank-deficient in some directions when the two parameters align.

## 7 Related Work

A broad offline-to-online literature studies how historical data, hints, or source models can reduce the cost of online learning. Warm-start and transfer approaches combine supervised and bandit feedback [Zhang et al., 2019], use initial hints [Cutkosky et al., 2022], handle deficient support in logged data [Tran-The et al., 2021], and transfer across contextual bandit tasks [Cai et al., 2024]. Other work considers weighted hypothesis transfer [Bilaj et al., 2023], partially observed confounded offline data [Tennenholtz et al., 2021], offline logs in larger sequential settings [Wagenmaker and Pacchiano, 2023, Kausik et al., 2025], and offline-to-online interaction in pricing [Zhang et al., 2025, Bu et al., 2020, 2025]. A related offline contextual bandit and off-policy learning line learns or evaluates policies from logged feedback under support, variance, or pessimism constraints [Dudík et al., 2011, Swaminathan and Joachims, 2015, Wang et al., 2024, Gabbianelli et al., 2024]. These works establish many ways in which past data can help decision-making, but their main object is not the online regret value of biased offline evidence after deployment.

The closest comparisons come from two directions. In linear bandits with offline data, prior observations can lower exploration cost through offline Gram matrices and design-based exploration when the offline and online parameters coincide [Vijayan et al., 2025]. In multi-armed bandits with biased offline data, MIN-UCB shows that a learner should exploit offline samples only when the mismatch is controlled, and otherwise fall back toward online learning [Cheung and Lyu, 2024]; related biased offline-to-online questions have also been studied in combinatorial and dueling bandit variants [Zhou et al., 2025, He et al., 2025]. Zhou et al. [2026] studied the biased offline to online setting under noncontextual linear bandit model where the bias is modeled through a known Euclidean bound which fails to capture the directional information in the offline data. These lines provide the two ingredients we use: coverage in a linear model and controlled use of biased offline information. Our focus is matrix-shaped bias budgets together with offline design geometry in a shared-feature contextual model, rather than arm-level transfer alone.

At the algorithmic level, our online branch builds on the optimism-based foundations of contextual and linear bandits: LinUCB and SupLinUCB [Li et al., 2010, Chu et al., 2011], the OFUL analysis [Abbasi-Yadkori et al., 2011], and linear Thompson-style approaches [Agrawal and Goyal, 2013]. We also relate to misspecification-robust bandits, where the online reward model may deviate from the assumed linear class. Classic and follow-up work includes Ghosh et al. [2017] and Takemura et al. [2021]; more recent developments include Foster et al. [2020] and Krishnamurthy et al. [2021]. Here the online model remains linear; the difficulty is that the *offline* parameter may differ from the online one in a structured, direction-dependent way. Our contribution is to encode that mismatch through a directional bias certificate and use it inside a hybrid pooled/online optimistic rule. For the adaptive part, we are not aware of prior regret guarantees that learn a directional bias certificate from biased offline data in linear contextual bandits.

## 8 Conclusion

This paper investigates how to use biased offline data in linear contextual bandits. We propose a novel direction-aware framework to measure the bias of the offline data and use them safely with comprehensive theoretical analysis and empirical validation. It makes bandit algorithms easier to apply in practice by relaxing how historical data must relate to the live environment. Future directions include extensions to other online learning settings, such as generalized linear models and online Markov decision processes. Also, it would also be interesting to develop refined lower bounds in the non-diagonal settings which are more challenging.

## Acknowledgments and Disclosure of Funding

This work was supported by the Theme-based Research Project T32-615/24-R and the General Research Fund 16500023 from the Hong Kong Research Grants Council. The authors declare no competing interests.

## References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, 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*, pages 127–135, 2013.

Steven Bilaj, Sofien Dhouib, and Setareh Maghsudi. Hypothesis transfer in bandits by weighted models. In *Machine Learning and Knowledge Discovery in Databases*, volume 13716 of *Lecture Notes in Computer Science*, pages 284–299, 2023. ECML PKDD 2022.

Jinzhi Bu, David Simchi-Levi, and Yunzong Xu. Online pricing with offline data: Phase transition and inverse square law. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1202–1210, 2020.

Jinzhi Bu, David Simchi-Levi, Yunzong Xu, and Chen Wen Zhai. Feature-based dynamic pricing with online learning and offline data. Working paper 5261068, SSRN, 2025.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207–232, 2024.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8286–8309, 2024.

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*, pages 208–214, 2011.

Ashok Cutkosky, Chris Dann, Abhimanyu Das, and Qiuyi Zhang. Leveraging initial hints for free in stochastic linear bandits. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 282–318, 2022.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems 33*, 2020.

Germano Gabbianelli, Gergely Neu, and Matteo Papini. Importance-weighted offline learning done right. In *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 614–634, 2024.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017.

Qijia He, Minghan Wang, Xutong Liu, Zhiyong Wang, and Fang Kong. Learning across the gap: Hybrid multi-armed bandits with heterogeneous offline and online data. In *Advances in Neural Information Processing Systems*, 2025.

Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. Leveraging offline data in linear latent contextual bandits. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 29345–29389, 2025.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5805–5814, 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*, pages 661–670. ACM, 2010.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 814–823, 2015.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 2021.

Guy Tennenholtz, Uri Shalit, Shie Mannor, and Yonathan Efroni. Bandits with partially observable confounded data. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 430–439, 2021.

Hung Tran-The, Sunil Gupta, Thanh Nguyen-Tang, Santu Rana, Dinh Phung, and Svetha Venkatesh. Combining online learning and offline learning for contextual bandits with deficient support. *arXiv preprint arXiv:2107.11533*, 2021.

Sushant Vijayan, Arun Suggala, Karthikeyan Shanmugam, and Soumyabrata Pal. Regret minimization in linear bandits with offline data via extended D-optimal exploration. *arXiv preprint arXiv:2508.08420*, 2025.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35300–35338, 2023.

Lequn Wang, Akshay Krishnamurthy, and Alex Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 766–774, 2024.

Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand N. Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7335–7344, 2019.

Yixuan Zhang, Ruihao Zhu, and Qiaomin Xie. Contextual online pricing with (biased) offline data. In *Advances in Neural Information Processing Systems*, 2025.

Kongchang Zhou, Tingyu Zhang, Wei Chen, and Fang Kong. Hybrid combinatorial multi-armed bandits with probabilistically triggered arms. *arXiv preprint arXiv:2512.21925*, 2025.

Kongchang Zhou, Zheshun Wu, Wei Chen, and Fang Kong. Regret analysis of hybrid linear bandits with biased offline data, 2026. URL `https://openreview.net/forum?id=NA4seLpkPd`.

## A Intuition behind the directional bias certificate

In this section, we discuss the intuition behind the directional bias certificate in detail. The core idea is to express the error $\Delta := \theta_* - \theta^\dagger$ in the linear combination of the eigenvectors of $M_{\text{bias}}$ and get a clear view of the bias along each specific eigenvector.

Let

$$\Delta := \theta_* - \theta^\dagger.$$

The directional bias certificate

$$\|\Delta\|_{M_{\text{bias}}} \le \rho$$

is equivalent to

$$\Delta^\top M_{\text{bias}} \Delta \le \rho^2.$$

We can write the eigen-decomposition of $M_{\text{bias}}$ as

$$M_{\text{bias}} = Q\Lambda Q^\top, \qquad Q = [q_1, \dots, q_d], \qquad Q^\top Q = I, \qquad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_d),$$

where $\lambda_j > 0$ and $q_j$ are the eigenvectors of $M_{\text{bias}}$. Define

$$\alpha := Q^\top \Delta, \qquad \alpha_j = q_j^\top \Delta.$$

Here, $\alpha_j$ can be viewed as the error projected to the $j$-th eigenvector. Then we have the following expression:

$$\Delta = Q\alpha = \sum_{j=1}^{d} \alpha_j q_j.$$

By definition, we have

$$\Delta^\top M_{\text{bias}} \Delta = \Delta^\top Q\Lambda Q^\top \Delta = (Q^\top \Delta)^\top \Lambda (Q^\top \Delta) = \alpha^\top \Lambda \alpha = \sum_{j=1}^{d} \lambda_j \alpha_j^2.$$

Therefore

$$\|\Delta\|_{M_{\text{bias}}} \le \rho \quad \Longleftrightarrow \quad \sum_{j=1}^{d} \lambda_j \alpha_j^2 \le \rho^2.$$

In particular, for each eigen-direction $q_j$,

$$|\alpha_j| = |q_j^\top \Delta| \le \frac{\rho}{\sqrt{\lambda_j}}.$$

Till now, we can see that a larger $\lambda_j$ typically means a small error along $q_j$. Under such circumstance, the offline data should be informative along $q_j$. Otherwise, the offline data is misleading along $q_j$.

For any feature direction $x$, we have

$$|x^\top(\theta_* - \theta^\dagger)| \le \rho \|x\|_{M_{\text{bias}}^{-1}}.$$

The right hand side in the above inequality can be written as:

$$\|x\|^2_{M_{\text{bias}}^{-1}} = \sum_{j=1}^{d} \frac{(q_j^\top x)^2}{\lambda_j}.$$

It means that directions aligned with small eigenvalues are amplified by $M_{\text{bias}}^{-1}$, while directions aligned with large eigenvalues are shrunk. Thus a direction with most of its energy on small-$\lambda_j$ eigendirections makes the above bound looser leading to a large worst-case bias width in the analysis of the UCB radius.

## B An Impossibility Result

The main text models bias through a directional bias certificate $\|\theta_* - \theta^\dagger\|_{M_{\text{bias}}} \le \rho$. The next theorem explains why replacing this by a single Euclidean summary $\|\theta_* - \theta^\dagger\|_2 \le \rho$ is too coarse. Under a finite-action Bernoulli linear bandit with dimension $d = 3$, parameter norm bound $S = 1$, and $|\mathcal{A}| = 3$, we have the following theorem:

**Theorem 5** (Isotropic bounds do not identify transferable directions)**.** *There exist universal constants $c_0, c_1 > 0$ such that the following holds. Fix $T \ge 1$, $\rho_0 \in (0, 1/4]$, and*

$$0 < \rho \le \min\Big\{\frac{\rho_0}{4},\, c_0 T^{-1/2}\Big\}.$$

*Then there exist three Bernoulli linear bandit instances $I_+, I_-, I_{\text{good}}$ sharing a common offline parameter $\theta^\dagger$ and a common action set $\mathcal{A}_t \equiv \mathcal{A}$, such that, writing $\theta_*^I$ for the online parameter of instance $I$, the offline data $D_{\text{off}}$ has the same distribution under $I_+, I_-, I_{\text{good}}$, and*

$$\big\|\theta_*^I - \theta^\dagger\big\|_2 \le \rho \quad \forall I \in \{I_+, I_-, I_{\text{good}}\}.$$

*Nevertheless, the corresponding online problems have different hardness:*

$$\inf_{\pi} \sup_{I \in \{I_+, I_-\}} \mathbb{E}_I[R_T(\pi)] \ge c_1 \frac{\rho^2 T}{\rho_0};$$

*where the infimum is over all non-anticipatory policies. In contrast, there exists a deterministic fixed-action policy $\pi_{\text{good}}$ such that*

$$\mathbb{E}_{I_{\text{good}}}[R_T(\pi_{\text{good}})] = 0.$$

*Consequently, a $\rho$-only isotropic summary cannot determine which offline directions transfer.*

*Proof.* Let $e_0, e_1, e_2$ denote the canonical basis of $\mathbb{R}^3$, and set

$$\bar\rho := \frac{\rho}{\sqrt{2}}.$$

Define the three actions

$$a_0 := e_1, \qquad a_+ := \frac{\rho_0 e_1 + \bar\rho e_2}{\sqrt{\rho_0^2 + \bar\rho^2}}, \qquad a_- := \frac{\rho_0 e_1 - \bar\rho e_2}{\sqrt{\rho_0^2 + \bar\rho^2}},$$

and lift them to

$$x^{(0)} := \frac{e_0 + a_0}{\sqrt{2}}, \qquad x^{(+)} := \frac{e_0 + a_+}{\sqrt{2}}, \qquad x^{(-)} := \frac{e_0 + a_-}{\sqrt{2}}.$$

Each lifted action has norm 1, so the finite action set

$$\mathcal{A}_t := \{x^{(0)}, x^{(+)}, x^{(-)}\} \qquad \forall t$$

fits the main setup with $K = 3$.

Now define the common offline parameter and the three online parameters by

$$\tilde\theta^\dagger := \frac{1}{\sqrt{2}} e_0 + \sqrt{2}\rho_0 e_1,$$

$$\tilde\theta_+ := \frac{1}{\sqrt{2}} e_0 + \sqrt{2}\big(\rho_0 e_1 + \bar\rho e_2\big), \qquad \tilde\theta_- := \frac{1}{\sqrt{2}} e_0 + \sqrt{2}\big(\rho_0 e_1 - \bar\rho e_2\big),$$

$$\tilde\theta_{\text{good}} := \frac{1}{\sqrt{2}} e_0 + \sqrt{2}(\rho_0 + \bar\rho) e_1.$$

Write these parameters as $\theta_+, \theta_-, \theta_{\text{good}}$, with common offline parameter $\theta^\dagger$. Take any deterministic offline covariates $\{z_i\}$ with $\|z_i\|_2 \le 1$ and $z_i^\top \theta^\dagger \in [0, 1]$, and let the offline rewards be Bernoulli with means $z_i^\top \theta^\dagger$. Online rewards are also Bernoulli, with mean $x^\top \theta$ when action $x \in \mathcal{A}_t$ is played under parameter $\theta$.

All displayed online means lie in $[1/2, 13/16]$: the smallest is $1/2 + (\rho_0^2 - \bar{\rho}^2)/\sqrt{\rho_0^2 + \bar{\rho}^2} \geq 1/2$, and the largest is $1/2 + \rho_0 + \bar{\rho} \leq 13/16$ because $\rho_0 \leq 1/4$ and $0 < \rho \leq \rho_0/4$. Hence rewards are in $[0, 1]$, and the centered Bernoulli noise is conditionally sub-Gaussian.

The three instances share the same offline parameter $\theta^\dagger$ and the same deterministic offline covariates, so the law of $D_{\text{off}}$ is identical across all three instances. Also,

$$\left\|\theta_+ - \theta^\dagger\right\|_2 = \left\|\theta_- - \theta^\dagger\right\|_2 = \left\|\theta_{\text{good}} - \theta^\dagger\right\|_2 = \rho,$$

so the same isotropic directional bias certificate is valid throughout.

For the hard family, let $P_+$ and $P_-$ denote the laws of the full online interaction under $\theta_+$ and $\theta_-$, respectively. Let $v := e_2$, and define the event

$$E := \left\{ \sum_{t=1}^{T} \mathbf{1}\{\langle v, x_t \rangle \geq 0\} \geq \frac{T}{2} \right\}.$$

Under $\theta_-$, the unique optimal action in $\mathcal{A}_t$ is $x^{(-)}$, whose mean reward equals

$$(x^{(-)})^\top \theta_- = \frac{1}{2} + \sqrt{\rho_0^2 + \bar{\rho}^2}.$$

For any chosen action $x_t \in \mathcal{A}_t$, write

$$x_t = \frac{e_0 + a_t}{\sqrt{2}} \qquad \text{with} \qquad a_t \in \{a_0, a_+, a_-\}.$$

Any action with $\langle v, x_t \rangle \geq 0$ is either $x^{(0)}$ or $x^{(+)}$, and each satisfies

$$x_t^\top \theta_- = \frac{1}{2} + \rho_0 \langle e_1, a_t \rangle - \bar{\rho} \langle v, a_t \rangle \leq \frac{1}{2} + \rho_0.$$

Hence every round in $E$ incurs regret at least

$$\sqrt{\rho_0^2 + \bar{\rho}^2} - \rho_0 \geq \frac{\bar{\rho}^2}{4\rho_0} = \frac{\rho^2}{8\rho_0},$$

where the last inequality uses $\bar{\rho} \leq \rho_0/4$. Therefore,

$$R_T(\theta_-) \geq \frac{\rho^2 T}{16\rho_0} \qquad \text{on } E.$$

By symmetry,

$$R_T(\theta_+) \geq \frac{\rho^2 T}{16\rho_0} \qquad \text{on } E^c.$$

The offline KL divergence between the two hard instances is zero because the offline law is the same. For the online interaction, if the chosen action at round $t$ has Bernoulli means $p_t := x_t^\top \theta_+$ and $q_t := x_t^\top \theta_-$, then

$$\mathrm{kl}(p_t, q_t) \leq \frac{(p_t - q_t)^2}{q_t(1 - q_t)} \leq \frac{256}{39}(p_t - q_t)^2,$$

because $q_t \in [1/2, 13/16]$. Moreover, $p_t - q_t = 0$ for $x_t = x^{(0)}$, while for $x_t \in \{x^{(+)}, x^{(-)}\}$,

$$|p_t - q_t| = \left|x_t^\top (\theta_+ - \theta_-)\right| = \frac{2\bar{\rho}^2}{\sqrt{\rho_0^2 + \bar{\rho}^2}} \leq \frac{2\bar{\rho}^2}{\rho_0}.$$

Hence every round contributes at most

$$\mathrm{KL}(p_t, q_t) \leq \frac{1024}{39} \frac{\bar{\rho}^4}{\rho_0^2} \leq \frac{64}{39}\bar{\rho}^2 = \frac{32}{39}\rho^2,$$

where the second inequality again uses $\bar{\rho} \leq \rho_0/4$. By the chain rule,

$$\mathrm{KL}(P_+ \,\|\, P_-) \leq \frac{32}{39}\rho^2 T.$$

Thus $\mathrm{KL}(P_+ \,\|\, P_-) \le 1/8$ when $c_0$ is chosen small enough. Bretagnolle–Huber then gives

$$P_+(E^c) + P_-(E) \ge \frac{1}{2} e^{-1/8}.$$

Combining the last three displays yields

$$\max\{\mathbb{E}_+[R_T],\, \mathbb{E}_-[R_T]\} \ge \frac{e^{-1/8}}{64} \frac{\rho^2 T}{\rho_0}.$$

This yields the first claim after absorbing constants into $c_1$.

For the good instance, the action $x^{(0)}$ has mean

$$(x^{(0)})^\top \theta_{\mathrm{good}} = \frac{1}{2} + \rho_0 + \bar{\rho},$$

whereas

$$(x^{(\pm)})^\top \theta_{\mathrm{good}} = \frac{1}{2} + \frac{\rho_0(\rho_0 + \bar{\rho})}{\sqrt{\rho_0^2 + \bar{\rho}^2}} < \frac{1}{2} + \rho_0 + \bar{\rho}.$$

Thus $x^{(0)}$ is optimal at every round, and the deterministic policy that always plays $x^{(0)}$ has zero regret, yielding the second claim.

$D_{\mathrm{off}}$ has the same law across the good and hard instances, so any measurable summary $S = \Phi(D_{\mathrm{off}})$ also has the same distribution. Hence no offline-only summary can distinguish the easy instance from the hard family. □

# C Notation

Table 1 collects recurrent symbols. The layerwise procedure is stated as Algorithm 1 in the main text.

Table 1: Main notation.

| Symbol | Meaning |
|---|---|
| $T, K, d$ | Horizon, number of actions per round, feature dimension. |
| $\theta_*$ | Unknown online reward parameter; $\|\theta_*\|_2 \le S$. |
| $\theta^\dagger$ | Offline regression parameter in $y_i = z_i^\top \theta^\dagger + \xi_i$. |
| $D_{\mathrm{off}}, n_{\mathrm{off}}$ | Offline sample $\{(z_i, y_i)\}$ and its size. |
| $G_{\mathrm{off}}, \hat{\theta}_{\mathrm{off}}, \beta_{\mathrm{off}}(\delta)$ | Offline ridge design $I + \sum_i z_i z_i^\top$, ridge estimator, and offline concentration radius. |
| $(M_{\mathrm{bias}}, \rho)$ | Directional bias certificate: $\|\theta_* - \theta^\dagger\|_{M_{\mathrm{bias}}} \le \rho$. |
| $\mathcal{F}_{t-1}$ | Filtration generated by past data and current contexts (Section 2). |
| $\ell, L$ | SupLinUCB layer index; $L = \lceil \log_2 T \rceil$. |
| $\mathcal{H}_{t,\ell}$ | History routed to layer $\ell$ before round $t$. |
| $A_{t,\ell}, V_{t,\ell}, b_{t,\ell}$ | Layerwise online Gram matrix, $V_{t,\ell} = I + A_{t,\ell}$, and response vector $b_{t,\ell}$ on layer $\ell$ (Section 3). |
| $\hat{\theta}^{\mathrm{on}}_{t,\ell}, \hat{\theta}^{\mathrm{pool}}_{t,\ell}$ | Online and pooled ridge estimators on layer $\ell$. |
| $\beta^{\mathrm{on}}_{t,\ell}, \gamma_{t,\ell}$ | Online SupLin radius and pooled martingale radius. |
| $\psi_{t,\ell}(x)$ | Bias routing term in the pooled radius (Section 3). |
| $U^{\mathrm{on}}, U^{\mathrm{pool}}, U^{\min}, L^{\max}, w$ | Branch UCB/LCB scores and aggregated width. |
| $\delta_{\mathrm{off}}, \delta_{\mathrm{on}}, \delta^{\mathrm{on}}_{t,\ell}, \delta^{\mathrm{pool}}_{t,\ell}$ | Offline and online union budgets; per-$(t,\ell)$ online and pooled budgets (Theorem 1). |
| $\widetilde{\Lambda}_T, \widetilde{\Psi}_T$ | Offline design log-det sum and path sum of $\psi$ (Theorem 1). |
| $A^{\mathrm{fin}}_\ell$ | Final Gram matrix on layer $\ell$ at horizon $T$. |
| $\Lambda_{\max}, \Gamma_{\mathrm{pool}}, \Gamma_{\mathrm{spec}}$ | Pooled log-det and spectral shorthands in Theorems 1 and 2. |

# D Proof of Proposition 1

*Proof.* Condition on the offline covariates $\{z_i\}_{i=1}^{n_{\mathrm{off}}}$. The imported self-normalized ridge inequality states that, with probability at least $1 - \delta$,

$$\|\hat{\theta}_{\mathrm{off}} - \theta^\dagger\|_{G_{\mathrm{off}}} \le \beta_{\mathrm{off}}(\delta).$$

On the same event, the directional bias certificate gives

$$\|\theta_* - \theta^\dagger\|_{M_{\text{bias}}} \le \rho.$$

Write

$$\theta_* - \hat{\theta}_{\text{off}} = (\theta^\dagger - \hat{\theta}_{\text{off}}) + (\theta_* - \theta^\dagger) =: u + v.$$

Then

$$\|u\|_{G_{\text{off}}} = \|\hat{\theta}_{\text{off}} - \theta^\dagger\|_{G_{\text{off}}} \le \beta_{\text{off}}(\delta), \qquad \|v\|_{M_{\text{bias}}} = \|\theta_* - \theta^\dagger\|_{M_{\text{bias}}} \le \rho.$$

Hence

$$\theta_* = \hat{\theta}_{\text{off}} + u + v$$

with $u$ and $v$ belonging to the two ellipsoids claimed in the proposition, which proves the region inclusion. Let

$$\theta = \hat{\theta}_{\text{off}} + u + v$$

with $\|u\|_{G_{\text{off}}} \le \beta_{\text{off}}$ and $\|v\|_{M_{\text{bias}}} \le \rho$. Then

$$x^\top \theta = x^\top \hat{\theta}_{\text{off}} + x^\top u + x^\top v.$$

By Cauchy–Schwarz in the $G_{\text{off}}$ metric,

$$x^\top u = (G_{\text{off}}^{-1/2} x)^\top (G_{\text{off}}^{1/2} u) \le \|x\|_{G_{\text{off}}^{-1}} \|u\|_{G_{\text{off}}} \le \beta_{\text{off}} \|x\|_{G_{\text{off}}^{-1}}.$$

Likewise,

$$x^\top v = (M_{\text{bias}}^{-1/2} x)^\top (M_{\text{bias}}^{1/2} v) \le \|x\|_{M_{\text{bias}}^{-1}} \|v\|_{M_{\text{bias}}} \le \rho \|x\|_{M_{\text{bias}}^{-1}}.$$

This yields the upper bound

$$x^\top \theta \le x^\top \hat{\theta}_{\text{off}} + \beta_{\text{off}} \|x\|_{G_{\text{off}}^{-1}} + \rho \|x\|_{M_{\text{bias}}^{-1}}.$$

Equality holds by choosing $u$ and $v$ aligned with the dual directions $G_{\text{off}}^{-1} x$ and $M_{\text{bias}}^{-1} x$, respectively; hence the display gives the exact support function. □

**Lemma 1** (Support function of the offline region)**.** *Under the event of Proposition 1, every $x \in \mathbb{R}^d$ satisfies*

$$\sup_{\theta \in \mathcal{C}_{\text{off}}} x^\top \theta = x^\top \hat{\theta}_{\text{off}} + \beta_{\text{off}} \|x\|_{G_{\text{off}}^{-1}} + \rho \|x\|_{M_{\text{bias}}^{-1}}.$$

*Proof.* This is exactly the support-function computation established in the proof of Proposition 1. □

# E Technical Lemmas

**Lemma 2** (Self-normalized martingale term)**.** *Let $(\mathcal{F}_s)_{s=0}^n$ be a filtration. Suppose $\phi_s \in \mathbb{R}^d$ is $\mathcal{F}_{s-1}$-measurable with $\|\phi_s\|_2 \le 1$, and $\eta_s$ is conditionally $\sigma$-sub-Gaussian given $\mathcal{F}_{s-1}$. Define*

$$Z_t := \sum_{s=1}^t \phi_s \eta_s, \qquad V_t := I + \sum_{s=1}^t \phi_s \phi_s^\top.$$

*Then, for every $\delta \in (0, 1)$,*

$$\mathbb{P}\left( \|Z_n\|_{V_n^{-1}} \le \sigma \sqrt{2 \log \frac{\det(V_n)^{1/2}}{\delta}} \right) \ge 1 - \delta.$$

*Proof.* Fix $\nu \in \mathbb{R}^d$ and define

$$M_t(\nu) := \exp\left( \sigma^{-1} \langle \nu, Z_t \rangle - \frac{1}{2} \nu^\top (V_t - I) \nu \right).$$

Since $\phi_t$ is $\mathcal{F}_{t-1}$-measurable and $\eta_t$ is conditionally $\sigma$-sub-Gaussian,

$$\mathbb{E}\left[ \exp\left( \sigma^{-1} \eta_t \langle \nu, \phi_t \rangle \right) \middle| \mathcal{F}_{t-1} \right] \le \exp\left( \frac{1}{2} \langle \nu, \phi_t \rangle^2 \right).$$

Therefore

$$\mathbb{E}\left[\frac{M_t(\nu)}{M_{t-1}(\nu)}\middle|\mathcal{F}_{t-1}\right] \le 1,$$

so $(M_t(\nu))_{t=0}^n$ is a nonnegative supermartingale and $\mathbb{E}[M_n(\nu)] \le 1$.

Now mix over $\nu$ with the Gaussian density

$$f(\nu) := \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\nu^\top \nu\right),$$

and define

$$\widetilde{M}_n := \int_{\mathbb{R}^d} M_n(\nu) f(\nu)\, d\nu.$$

By Fubini,

$$\mathbb{E}[\widetilde{M}_n] = \int_{\mathbb{R}^d} \mathbb{E}[M_n(\nu)] f(\nu)\, d\nu \le \int_{\mathbb{R}^d} f(\nu)\, d\nu = 1.$$

Combining the definitions of $M_n(\nu)$ and $f(\nu)$ gives

$$\widetilde{M}_n = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\left(\sigma^{-1}\langle \nu, Z_n\rangle - \frac{1}{2}\nu^\top V_n \nu\right) d\nu.$$

Completing the square,

$$\sigma^{-1}\langle \nu, Z_n\rangle - \frac{1}{2}\nu^\top V_n \nu = -\frac{1}{2}\left\|\nu - \sigma^{-1}V_n^{-1}Z_n\right\|_{V_n}^2 + \frac{1}{2\sigma^2}\|Z_n\|_{V_n^{-1}}^2.$$

Hence

$$\widetilde{M}_n = \frac{1}{(2\pi)^{d/2}} \exp\left(\frac{1}{2\sigma^2}\|Z_n\|_{V_n^{-1}}^2\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left\|\nu - \sigma^{-1}V_n^{-1}Z_n\right\|_{V_n}^2\right) d\nu.$$

The remaining Gaussian integral equals $(2\pi)^{d/2}\det(V_n)^{-1/2}$, so

$$\widetilde{M}_n = \frac{1}{\det(V_n)^{1/2}} \exp\left(\frac{1}{2\sigma^2}\|Z_n\|_{V_n^{-1}}^2\right).$$

Markov's inequality and $\mathbb{E}[\widetilde{M}_n] \le 1$ imply

$$\mathbb{P}\left(\widetilde{M}_n \ge \delta^{-1}\right) \le \delta.$$

Rearranging the event $\widetilde{M}_n \le \delta^{-1}$ yields the claim. □

**Lemma 3** (Self-normalized martingale with an SPD base matrix)**.** *Let $(\mathcal{F}_s)_{s=0}^n$ be a filtration. Suppose $\phi_s \in \mathbb{R}^d$ is $\mathcal{F}_{s-1}$-measurable with $\|\phi_s\|_2 \le 1$, and $\eta_s$ is conditionally $\sigma$-sub-Gaussian given $\mathcal{F}_{s-1}$. Let $H \in \mathbb{S}_{++}^d$ be deterministic or, more generally, $\mathcal{F}_0$-measurable. Define*

$$Z_t := \sum_{s=1}^t \phi_s \eta_s, \qquad V_t := H + \sum_{s=1}^t \phi_s \phi_s^\top.$$

*Then, for every $\delta \in (0,1)$,*

$$\mathbb{P}\left(\|Z_n\|_{V_n^{-1}} \le \sigma\sqrt{2\log\frac{\det(V_n)^{1/2}}{\det(H)^{1/2}\delta}}\right) \ge 1-\delta.$$

*Proof.* Condition on $H$ and fix $\nu \in \mathbb{R}^d$. Define

$$M_t(\nu) := \exp\left(\sigma^{-1}\langle \nu, Z_t\rangle - \frac{1}{2}\nu^\top (V_t - H)\nu\right).$$

Exactly as in the proof of Lemma 2,

$$\mathbb{E}\left[\frac{M_t(\nu)}{M_{t-1}(\nu)}\middle|\mathcal{F}_{t-1}\right] \le 1,$$

so $(M_t(\nu))_{t=0}^n$ is a nonnegative supermartingale and $\mathbb{E}[M_n(\nu) \mid H] \le 1$.
Now mix over $\nu$ with the Gaussian density having covariance $H^{-1}$,

$$f_H(\nu) := \frac{\det(H)^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\nu^\top H \nu\right),$$

and define

$$\widetilde{M}_n := \int_{\mathbb{R}^d} M_n(\nu) f_H(\nu)\, d\nu.$$

By Fubini,

$$\mathbb{E}[\widetilde{M}_n \mid H] = \int_{\mathbb{R}^d} \mathbb{E}[M_n(\nu) \mid H] f_H(\nu)\, d\nu \le 1.$$

Combining the definitions gives

$$\widetilde{M}_n = \frac{\det(H)^{1/2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\left(\sigma^{-1} \langle \nu, Z_n \rangle - \frac{1}{2}\nu^\top V_n \nu\right) d\nu.$$

Completing the square,

$$\sigma^{-1} \langle \nu, Z_n \rangle - \frac{1}{2}\nu^\top V_n \nu = -\frac{1}{2} \left\| \nu - \sigma^{-1} V_n^{-1} Z_n \right\|_{V_n}^2 + \frac{1}{2\sigma^2} \|Z_n\|_{V_n^{-1}}^2.$$

The remaining Gaussian integral equals $(2\pi)^{d/2} \det(V_n)^{-1/2}$, so

$$\widetilde{M}_n = \frac{\det(H)^{1/2}}{\det(V_n)^{1/2}} \exp\left(\frac{1}{2\sigma^2} \|Z_n\|_{V_n^{-1}}^2\right).$$

Markov's inequality conditional on $H$ implies

$$\mathbb{P}\Big(\widetilde{M}_n \ge \delta^{-1} \Big| H\Big) \le \delta.$$

Rearranging the event $\widetilde{M}_n \le \delta^{-1}$ yields the claim. □

**Proposition 2** (Ridge confidence ellipsoid)**.** *Under the assumptions of Lemma 2, let*

$$Y_s = \phi_s^\top \theta + \eta_s, \qquad \hat{\theta}_n := V_n^{-1} \sum_{s=1}^n \phi_s Y_s, \qquad \|\theta\|_2 \le S.$$

*Then, with probability at least* $1 - \delta$*,*

$$\|\hat{\theta}_n - \theta\|_{V_n} \le \sigma \sqrt{2 \log \frac{\det(V_n)^{1/2}}{\delta}} + S.$$

*Proof.* Using $Y_s = \phi_s^\top \theta + \eta_s$,

$$V_n \hat{\theta}_n = \sum_{s=1}^n \phi_s Y_s = \sum_{s=1}^n \phi_s \phi_s^\top \theta + \sum_{s=1}^n \phi_s \eta_s = (V_n - I)\theta + Z_n.$$

Therefore

$$V_n(\hat{\theta}_n - \theta) = Z_n - \theta.$$

Taking the $V_n^{-1}$ norm and applying the triangle inequality,

$$\|\hat{\theta}_n - \theta\|_{V_n} = \|Z_n - \theta\|_{V_n^{-1}} \le \|Z_n\|_{V_n^{-1}} + \|\theta\|_{V_n^{-1}}.$$

Lemma 2 controls the first term. Since $V_n \succeq I$, we have $V_n^{-1} \preceq I$, hence

$$\|\theta\|_{V_n^{-1}} \le \|\theta\|_2 \le S.$$

Combining the two bounds proves the result. □

**Lemma 4** (Predictable masks)**.** *Let $(\mathcal{F}_s)_{s=0}^T$ be a filtration. Suppose $\phi_s$ is $\mathcal{F}_{s-1}$-measurable and $\eta_s$ is conditionally $\sigma$-sub-Gaussian given $\mathcal{F}_{s-1}$, and $\|\phi_s\|_2 \le 1$. Let $H \in \mathbb{S}_{++}^d$ be deterministic or, more generally, $\mathcal{F}_0$-measurable. Let $m_s \in [0,1]$ be $\mathcal{F}_{s-1}$-measurable. Define*

$$\widetilde{Z}_t := \sum_{s=1}^{t} m_s \phi_s \eta_s, \qquad \widetilde{V}_t := H + \sum_{s=1}^{t} m_s^2 \phi_s \phi_s^\top.$$

*Then, with probability at least $1-\delta$,*

$$\|\widetilde{Z}_T\|_{\widetilde{V}_T^{-1}} \le \sigma \sqrt{2 \log \frac{\det(\widetilde{V}_T)^{1/2}}{\det(H)^{1/2}\delta}}.$$

*Proof.* Set $\widetilde{\phi}_s := m_s \phi_s$. Since $m_s$ and $\phi_s$ are both $\mathcal{F}_{s-1}$-measurable, $\widetilde{\phi}_s$ is predictable. Moreover $\|\widetilde{\phi}_s\|_2 \le \|\phi_s\|_2 \le 1$ because $m_s \in [0,1]$. Applying Lemma 3 to the predictable pair $(\widetilde{\phi}_s, \eta_s)$ with base matrix $H$ yields the claim. □

# F Confidence Events and Elimination Lemmas

This section supplies the entire logical chain behind the main theorem: simultaneous confidence, combined UCB and LCB validity, survival of the optimal arm, and the stopping-layer regret bound.

For each round let

$$a_t^\star \in \arg\max_{a \in [K]} x_{t,a}^\top \theta_*, \qquad \Delta_t(a) := x_{t,a_t^\star}^\top \theta_* - x_{t,a}^\top \theta_*.$$

Let $\mathcal{A}_{t,\ell}$ denote the set of arms surviving through layer $\ell$ on round $t$, and let $\ell_t$ be the stopping layer on round $t$. Define the widths

$$w_{t,\ell}(a) := U_{t,\ell}^{\min}(a) - L_{t,\ell}^{\max}(a), \qquad w_{t,\ell}^{\mathrm{on}}(a) := U_{t,\ell}^{\mathrm{on}}(a) - L_{t,\ell}^{\mathrm{on}}(a),$$

and similarly $w_{t,\ell}^{\mathrm{pool}}(a) := U_{t,\ell}^{\mathrm{pool}}(a) - L_{t,\ell}^{\mathrm{pool}}(a)$.

**Lemma 5** (Width comparison)**.** *For every round, layer, and action,*

$$w_{t,\ell}(a) \le \min\{w_{t,\ell}^{\mathrm{on}}(a), w_{t,\ell}^{\mathrm{pool}}(a)\} \le w_{t,\ell}^{\mathrm{on}}(a),$$

*and*

$$w_{t,\ell}(a) \le w_{t,\ell}^{\mathrm{pool}}(a) = 2\,\mathrm{rad}_{t,\ell}^{\mathrm{pool}}(x_{t,a}).$$

*Proof.* By definition,

$$w_{t,\ell}(a) = \min\{U_{t,\ell}^{\mathrm{on}}(a), U_{t,\ell}^{\mathrm{pool}}(a)\} - \max\{L_{t,\ell}^{\mathrm{on}}(a), L_{t,\ell}^{\mathrm{pool}}(a)\}.$$

For any numbers $u_1, u_2, \ell_1, \ell_2$,

$$\min\{u_1, u_2\} - \max\{\ell_1, \ell_2\} \le \min\{u_1 - \ell_1, u_2 - \ell_2\}.$$

Applying this with the online and pooled endpoints proves the first display. The second follows from the explicit form of the pooled UCB–LCB pair. □

Choose confidence levels so that

$$\sum_{t=1}^{T} \sum_{\ell=0}^{\lceil \log_2 T \rceil} \delta_{t,\ell}^{\mathrm{on}} \le \delta_{\mathrm{on}}, \qquad \delta_{\mathrm{off}} + \delta_{\mathrm{on}} + \sum_{t=1}^{T} \sum_{\ell=0}^{\lceil \log_2 T \rceil} \delta_{t,\ell}^{\mathrm{pool}} \le T^{-2},$$

abbreviate $\beta_{\mathrm{off}} := \beta_{\mathrm{off}}(\delta_{\mathrm{off}})$, set

$$L_T := \log\Big(\frac{2KT \log T}{\delta_{\mathrm{on}}}\Big).$$

Also define

$$\mathcal{E}_{t,\ell}^{\mathrm{on}} := \Big\{ \|b_{t,\ell} - A_{t,\ell}\theta_*\|_{V_{t,\ell}^{-1}} \le \beta_{t,\ell}^{\mathrm{on}} - S \Big\}.$$

Define the events

$$\mathcal{E}_{\text{off}} := \left\{ \|\hat{\theta}_{\text{off}} - \theta^\dagger\|_{G_{\text{off}}} \le \beta_{\text{off}} \right\},$$

$$\mathcal{E}_{t,\ell}^{\text{pool}} := \left\{ \|b_{t,\ell} - A_{t,\ell}\theta_*\|_{(A_{t,\ell}+G_{\text{off}})^{-1}} \le \gamma_{t,\ell} \right\},$$

$$\mathcal{E}_{\text{on}} := \bigcap_{t=1}^{T} \bigcap_{\ell=0}^{\lceil \log_2 T \rceil} \mathcal{E}_{t,\ell}^{\text{on}},$$

and finally set

$$\mathcal{E} := \mathcal{E}_{\text{on}} \cap \mathcal{E}_{\text{off}} \cap \bigcap_{t=1}^{T} \bigcap_{\ell=0}^{\lceil \log_2 T \rceil} \mathcal{E}_{t,\ell}^{\text{pool}}.$$

**Lemma 6** (Simultaneous confidence event)**.** *We have* $\mathbb{P}(\mathcal{E}) \ge 1 - T^{-2}$.

*Proof.* By the offline ridge bound,

$$\mathbb{P}(\mathcal{E}_{\text{off}}) \ge 1 - \delta_{\text{off}}.$$

Fix $(t, \ell)$ and define the predictable mask

$$m_s^{(t,\ell)} := \mathbf{1}\{s < t,\ s \in \mathcal{H}_{t,\ell}\}, \qquad s = 1, \dots, T.$$

Because the action set at round $s$ is revealed before the layer assignment and action choice, the event $\{s \in \mathcal{H}_{t,\ell}\}$ is determined by information available at time $s-1$ together with the current contexts; hence $m_s^{(t,\ell)}$ is $\mathcal{F}_{s-1}$-measurable. Therefore

$$b_{t,\ell} - A_{t,\ell}\theta_* = \sum_{s=1}^{T} m_s^{(t,\ell)} x_{s,a_s} \eta_s, \qquad \eta_s := r_s - x_{s,a_s}^\top \theta_*,$$

and

$$A_{t,\ell} + G_{\text{off}} = G_{\text{off}} + \sum_{s=1}^{T} m_s^{(t,\ell)} x_{s,a_s} x_{s,a_s}^\top.$$

Applying Lemma 4 with $\phi_s = x_{s,a_s}$, mask $m_s^{(t,\ell)}$, and base matrix $H = G_{\text{off}}$ gives

$$\|b_{t,\ell} - A_{t,\ell}\theta_*\|_{(A_{t,\ell}+G_{\text{off}})^{-1}} \le \sigma \sqrt{2 \log \frac{\det(A_{t,\ell} + G_{\text{off}})^{1/2}}{\det(G_{\text{off}})^{1/2} \delta_{t,\ell}^{\text{pool}}}}$$

with probability at least $1 - \delta_{t,\ell}^{\text{pool}}$, which is exactly the event $\mathcal{E}_{t,\ell}^{\text{pool}}$. Thus

$$\mathbb{P}(\mathcal{E}_{t,\ell}^{\text{pool}}) \ge 1 - \delta_{t,\ell}^{\text{pool}}.$$

The same masked bound with confidence level $\delta_{t,\ell}^{\text{on}}$ and base matrix $H = I$ yields

$$\mathbb{P}(\mathcal{E}_{t,\ell}^{\text{on}}) \ge 1 - \delta_{t,\ell}^{\text{on}}.$$

Applying the union bound over all $(t, \ell)$ and using $\sum_{t,\ell} \delta_{t,\ell}^{\text{on}} \le \delta_{\text{on}}$ gives

$$\mathbb{P}(\mathcal{E}_{\text{on}}) \ge 1 - \delta_{\text{on}}.$$

Applying the union bound over $\mathcal{E}_{\text{off}}^c$, $\mathcal{E}_{\text{on}}^c$, and all pooled events gives

$$\mathbb{P}(\mathcal{E}^c) \le \delta_{\text{off}} + \delta_{\text{on}} + \sum_{t,\ell} \delta_{t,\ell}^{\text{pool}} \le T^{-2}.$$

□

**Lemma 7** (Validity of the pooled branch)**.** *Fix a round-layer pair* $(t, \ell)$*. On the event* $\mathcal{E}_{\text{off}} \cap \mathcal{E}_{t,\ell}^{\text{pool}}$*, every action* $a$ *satisfies*

$$\left| x_{t,a}^\top (\hat{\theta}_{t,\ell}^{\text{pool}} - \theta_*) \right| \le (\gamma_{t,\ell} + \beta_{\text{off}}) \|x_{t,a}\|_{(A_{t,\ell}+G_{\text{off}})^{-1}} + \rho\, \psi_{t,\ell}(x_{t,a}).$$

*Proof.* Write $A := A_{t,\ell}$, $G := G_{\text{off}}$, and $x := x_{t,a}$. Expanding the pooled estimator,

$$\hat{\theta}^{\text{pool}}_{t,\ell} - \theta_* = (A+G)^{-1}\Big[(b_{t,\ell} - A\theta_*) + G(\hat{\theta}_{\text{off}} - \theta^\dagger) + G(\theta^\dagger - \theta_*)\Big].$$

Therefore

$$\Big|x^\top(\hat{\theta}^{\text{pool}}_{t,\ell} - \theta_*)\Big| \le |T_1| + |T_2| + |T_3|,$$

where

$$T_1 := x^\top(A+G)^{-1}(b_{t,\ell} - A\theta_*), \quad T_2 := x^\top(A+G)^{-1}G(\hat{\theta}_{\text{off}} - \theta^\dagger),$$
$$T_3 := x^\top(A+G)^{-1}G(\theta^\dagger - \theta_*).$$

For $T_1$,

$$|T_1| = |x^\top(A+G)^{-1}(b_{t,\ell} - A\theta_*)| \le \|x\|_{(A+G)^{-1}}\|b_{t,\ell} - A\theta_*\|_{(A+G)^{-1}}.$$

Using $\mathcal{E}^{\text{pool}}_{t,\ell}$ gives

$$|T_1| \le \gamma_{t,\ell}\|x\|_{(A+G)^{-1}}.$$

For $T_2$, write

$$|T_2| = \Big|\big(G^{1/2}(A+G)^{-1}x\big)^\top\big(G^{1/2}(\hat{\theta}_{\text{off}} - \theta^\dagger)\big)\Big|$$

and apply Cauchy–Schwarz:

$$|T_2| \le \|x\|_{(A+G)^{-1}G(A+G)^{-1}}\|\hat{\theta}_{\text{off}} - \theta^\dagger\|_G.$$

Because $G \preceq A + G$,

$$(A+G)^{-1}G(A+G)^{-1} \preceq (A+G)^{-1},$$

so on $\mathcal{E}_{\text{off}}$,

$$|T_2| \le \beta_{\text{off}}\|x\|_{(A+G)^{-1}}.$$

For $T_3$, let $w := G(A+G)^{-1}x$. Then

$$|T_3| = |w^\top(\theta^\dagger - \theta_*)| \le \rho\|w\|_{M^{-1}_{\text{bias}}} = \rho\,\psi_{t,\ell}(x)$$

by the directional bias certificate. Summing the three bounds proves the claim. □

**Lemma 8** (Combined UCB and LCB validity)**.** *On $\mathcal{E}$, every round-layer-action triple satisfies*

$$x^\top_{t,a}\theta_* \in [L^{\max}_{t,\ell}(a), U^{\min}_{t,\ell}(a)].$$

*Proof.* On $\mathcal{E}$, the pooled UCB–LCB pair is valid by Lemma 7. For the online branch,

$$\hat{\theta}^{\text{on}}_{t,\ell} - \theta_* = V^{-1}_{t,\ell}(b_{t,\ell} - A_{t,\ell}\theta_*) - V^{-1}_{t,\ell}\theta_*.$$

Thus

$$\|\hat{\theta}^{\text{on}}_{t,\ell} - \theta_*\|_{V_{t,\ell}} \le \|b_{t,\ell} - A_{t,\ell}\theta_*\|_{V^{-1}_{t,\ell}} + \|\theta_*\|_{V^{-1}_{t,\ell}} \le (\beta^{\text{on}}_{t,\ell} - S) + S = \beta^{\text{on}}_{t,\ell}$$

on $\mathcal{E}^{\text{on}}_{t,\ell}$, since $V_{t,\ell} \succeq I$ and $\|\theta_*\|_2 \le S$. Therefore $|x^\top_{t,a}(\hat{\theta}^{\text{on}}_{t,\ell} - \theta_*)| \le \beta^{\text{on}}_{t,\ell}\|x_{t,a}\|_{V^{-1}_{t,\ell}}$ on $\mathcal{E}$, so the online pair is also valid. Therefore

$$x^\top_{t,a}\theta_* \le \min\{U^{\text{on}}_{t,\ell}(a), U^{\text{pool}}_{t,\ell}(a)\} = U^{\min}_{t,\ell}(a)$$

and similarly

$$x^\top_{t,a}\theta_* \ge \max\{L^{\text{on}}_{t,\ell}(a), L^{\text{pool}}_{t,\ell}(a)\} = L^{\max}_{t,\ell}(a).$$

□

**Lemma 9** (Optimal action survives elimination)**.** *On $\mathcal{E}$, for every round $t$ and layer $\ell$,*

$$a^\star_t \in \mathcal{A}_{t,\ell}.$$

*Proof.* The base case $\ell = 0$ is immediate since $\mathcal{A}_{t,0} = [K]$. Suppose $a_t^\star \in \mathcal{A}_{t,\ell}$ and the algorithm descends to layer $\ell + 1$. By definition,

$$m_{t,\ell} := \max_{a \in \mathcal{A}_{t,\ell}} L_{t,\ell}^{\max}(a).$$

Lemma 8 implies

$$L_{t,\ell}^{\max}(a) \le x_{t,a}^\top \theta_* \le x_{t,a_t^\star}^\top \theta_* \le U_{t,\ell}^{\min}(a_t^\star) \qquad \text{for all } a \in \mathcal{A}_{t,\ell}.$$

Therefore

$$U_{t,\ell}^{\min}(a_t^\star) \ge m_{t,\ell},$$

so the optimal action is not eliminated and remains in $\mathcal{A}_{t,\ell+1}$. □

**Lemma 10** (Survivor quality after a narrow layer)**.** *On $\mathcal{E}$, fix $t$ and $\ell < \lceil \log_2 T \rceil$. If every arm in $\mathcal{A}_{t,\ell}$ satisfies $w_{t,\ell}(a) \le 2^{-\ell}$ and the algorithm descends to layer $\ell + 1$, then every $a \in \mathcal{A}_{t,\ell+1}$ obeys*

$$\Delta_t(a) \le 2^{1-\ell}.$$

*Proof.* Take $a \in \mathcal{A}_{t,\ell+1}$. By the elimination rule,

$$U_{t,\ell}^{\min}(a) \ge m_{t,\ell} \ge L_{t,\ell}^{\max}(a_t^\star).$$

Hence

$$\begin{aligned}\Delta_t(a) &= x_{t,a_t^\star}^\top \theta_* - x_{t,a}^\top \theta_* \\ &= \big(x_{t,a_t^\star}^\top \theta_* - L_{t,\ell}^{\max}(a_t^\star)\big) + \big(L_{t,\ell}^{\max}(a_t^\star) - U_{t,\ell}^{\min}(a)\big) + \big(U_{t,\ell}^{\min}(a) - x_{t,a}^\top \theta_*\big).\end{aligned}$$

The middle term is nonpositive, and the first and third terms are each at most $2^{-\ell}$ by Lemma 8 and the width assumption. Thus

$$\Delta_t(a) \le 2^{-\ell} + 2^{-\ell} = 2^{1-\ell}.$$

□

**Corollary 3** (Layer invariant)**.** *On $\mathcal{E}$, every round $t$ and every layer $\ell \ge 1$ satisfy*

$$\Delta_t(a) \le 2^{2-\ell} \qquad \textit{for all } a \in \mathcal{A}_{t,\ell}.$$

*Proof.* On $\mathcal{E}$, apply Lemma 10 at layer $\ell - 1$. □

**Lemma 11** (Stopping-layer regret bound)**.** *On $\mathcal{E}$, every round $t$ satisfies*

$$\Delta_t(a_t) \le 4 \min\{1, w_{t,\ell_t}(a_t)\} \le 8 \min\{1, \mathrm{rad}_{t,\ell_t}^{\mathrm{pool}}(x_{t,a_t})\}.$$

*Proof.* We consider the stopping layer $\ell_t$.

**Case 1: $\ell_t = \lceil \log_2 T \rceil$.** The played arm maximizes $U_{t,\ell_t}^{\min}$ over $\mathcal{A}_{t,\ell_t}$, and $a_t^\star \in \mathcal{A}_{t,\ell_t}$ by Lemma 9. Therefore

$$U_{t,\ell_t}^{\min}(a_t) \ge U_{t,\ell_t}^{\min}(a_t^\star) \ge x_{t,a_t^\star}^\top \theta_*.$$

Using Lemma 8,

$$\Delta_t(a_t) \le U_{t,\ell_t}^{\min}(a_t) - L_{t,\ell_t}^{\max}(a_t) = w_{t,\ell_t}(a_t).$$

Also, rewards lie in $[0, 1]$, so $\Delta_t(a_t) \le 1$. Hence

$$\Delta_t(a_t) \le \min\{1, w_{t,\ell_t}(a_t)\} \le 4 \min\{1, w_{t,\ell_t}(a_t)\}.$$

**Case 2:** $\ell_t < \lceil \log_2 T \rceil$**.** If $\ell_t = 0$, then $\Delta_t(a_t) \le 1$ because rewards are bounded in $[0,1]$, and $1 \le 4\min\{1, w_{t,0}(a_t)\}$ since the algorithm stops only when $w_{t,0}(a_t) > 1$ or the terminal layer is reached.

If $\ell_t \ge 1$, Corollary 3 gives

$$\Delta_t(a_t) \le 2^{2-\ell_t}.$$

But the stopping rule at a nonterminal layer implies $w_{t,\ell_t}(a_t) > 2^{-\ell_t}$, so

$$4\, w_{t,\ell_t}(a_t) > 4 \cdot 2^{-\ell_t} = 2^{2-\ell_t}.$$

Hence $\Delta_t(a_t) \le 4\, w_{t,\ell_t}(a_t)$.

Finally, Lemma 5 implies

$$w_{t,\ell_t}(a_t) \le 2\,\mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t}),$$

so

$$4\min\{1, w_{t,\ell_t}(a_t)\} \le 4\min\{1, 2\,\mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\} \le 8\min\{1, \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\},$$

which yields the second bound. □

**Corollary 4** (Mixed-branch master bound)**.** *On $\mathcal{E}$, every round $t$ satisfies*

$$\Delta_t(a_t) \le 8\min\left\{1,\ \mathrm{rad}^{\mathrm{on}}_{t,\ell_t}(x_{t,a_t}),\ \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\right\}.$$

*Consequently,*

$$R_T \le 8\sum_{t=1}^{T} \min\left\{1,\ \mathrm{rad}^{\mathrm{on}}_{t,\ell_t}(x_{t,a_t}),\ \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\right\}.$$

*Proof.* By Lemma 11,

$$\Delta_t(a_t) \le 4\min\{1, w_{t,\ell_t}(a_t)\}.$$

Lemma 5 and the explicit online and pooled widths give

$$w_{t,\ell_t}(a_t) \le \min\{w^{\mathrm{on}}_{t,\ell_t}(a_t), w^{\mathrm{pool}}_{t,\ell_t}(a_t)\} = 2\min\left\{\mathrm{rad}^{\mathrm{on}}_{t,\ell_t}(x_{t,a_t}), \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\right\}.$$

Therefore, if

$$m_t := \min\left\{\mathrm{rad}^{\mathrm{on}}_{t,\ell_t}(x_{t,a_t}), \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\right\},$$

then

$$\Delta_t(a_t) \le 4\min\{1, 2m_t\} \le 8\min\{1, m_t\} = 8\min\left\{1,\ \mathrm{rad}^{\mathrm{on}}_{t,\ell_t}(x_{t,a_t}),\ \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\right\}.$$

This proves the first display, and summing over $t$ yields the second. □

## G Roadmap for the Online Fallback

The online branch follows the classical SupLinUCB logic. We spell out the two standard ingredients needed to transfer the fallback bound to our combined UCB/LCB wrapper.

**Lemma 12** (Online width trigger)**.** *If round $t$ stops at a nonterminal layer $\ell_t$, then*

$$w^{\mathrm{on}}_{t,\ell_t}(a_t) > 2^{-\ell_t}.$$

*Proof.* The stopping rule gives $w_{t,\ell_t}(a_t) > 2^{-\ell_t}$, and $w_{t,\ell_t}(a_t) \le w^{\mathrm{on}}_{t,\ell_t}(a_t)$ by Lemma 5. □

**Lemma 13** (Layerwise online potential)**.** *Let*

$$Q_\ell := \{t \in [T] : \ell_t = \ell\}$$

*and*

$$V^{\mathrm{fin}}_\ell := I + \sum_{t \in Q_\ell} x_{t,a_t} x_{t,a_t}^\top.$$

*Then*

$$\sum_{t \in Q_\ell} \min\left\{1, \|x_{t,a_t}\|^2_{V^{-1}_{t,\ell}}\right\} \le 2\log\frac{\det(V^{\mathrm{fin}}_\ell)}{\det(I)} \le 2d\log(1+T).$$

*Proof.* Enumerate $Q_\ell$ as $t_1 < \cdots < t_{|Q_\ell|}$ and define

$$W_{\ell,0} := I, \qquad W_{\ell,j} := I + \sum_{i=1}^{j} x_{t_i,a_{t_i}} x_{t_i,a_{t_i}}^\top.$$

Then $W_{\ell,j-1} = V_{t_j,\ell}$. By the matrix-determinant lemma,

$$\det(W_{\ell,j}) = \det(W_{\ell,j-1}) \left(1 + \|x_{t_j,a_{t_j}}\|^2_{W_{\ell,j-1}^{-1}}\right).$$

Summing logarithms gives

$$\log \frac{\det(V_\ell^{\mathrm{fin}})}{\det(I)} = \sum_{j=1}^{|Q_\ell|} \log \left(1 + \|x_{t_j,a_{t_j}}\|^2_{W_{\ell,j-1}^{-1}}\right).$$

Using $\log(1+u) \geq \frac{1}{2}\min\{u,1\}$ proves the first inequality. The second follows from AM–GM:

$$\det(V_\ell^{\mathrm{fin}}) \leq \left(1 + \frac{T}{d}\right)^d.$$

□

**Proposition 3** (Classical online fallback)**.** *On $\mathcal{E}$, the online branch yields*

$$R_T \leq C_{\mathrm{SL}} \sqrt{T d\, L_T^3}.$$

*Proof.* The standard SupLinUCB counting argument uses only the following ingredients:

1. the optimal action survives elimination at every layer;
2. every arm surviving a narrow layer is near-optimal;
3. stopping at a nonterminal layer shows a large online width;
4. the layerwise online widths satisfy an elliptical-potential bound.

In our wrapper these are supplied, respectively, by Lemmas 9, 10, 12, and 13. The proof of Theorem 1 in Chu et al. [2011] is purely pathwise once those four ingredients hold, so the same dyadic summation yields $R_T \leq C_{\mathrm{SL}} \sqrt{T d\, L_T^3}$ on $\mathcal{E}$. □

## H Proof of Theorem 1

We now prove the pooled part of the theorem in full detail.

For each layer let

$$Q_\ell := \{t \in [T] : \ell_t = \ell\}, \qquad A_\ell^{\mathrm{fin}} := \sum_{t \in Q_\ell} x_{t,a_t} x_{t,a_t}^\top.$$

Define

$$\Lambda_\ell := \log \frac{\det(A_\ell^{\mathrm{fin}} + G_{\mathrm{off}})}{\det(G_{\mathrm{off}})}, \qquad \widetilde{\Lambda}_T := \sum_{\ell=0}^{\lceil \log_2 T \rceil} \Lambda_\ell, \qquad \widetilde{\Psi}_T := \sum_{t=1}^{T} \psi_{t,\ell_t}(x_{t,a_t}).$$

Also set

$$\delta_{\ell,\min}^{\mathrm{pool}} := \min_{t \leq T} \delta_{t,\ell}^{\mathrm{pool}}, \qquad \gamma_\ell^{\mathrm{fin}} := \sigma \sqrt{\Lambda_\ell + 2\log(1/\delta_{\ell,\min}^{\mathrm{pool}})},$$

$$\Lambda_{\max} := \max_{0 \leq \ell \leq \lceil \log_2 T \rceil} \Lambda_\ell, \qquad \delta_{\min}^{\mathrm{pool}} := \min_{t,\ell} \delta_{t,\ell}^{\mathrm{pool}}, \qquad \Gamma_{\mathrm{pool}} := \beta_{\mathrm{off}} + \sigma \sqrt{\Lambda_{\max} + 2\log(1/\delta_{\min}^{\mathrm{pool}})},$$

and

$$\Gamma_{\mathrm{spec}} := \beta_{\mathrm{off}} + \sigma \sqrt{\Phi_{G_{\mathrm{off}}}(T) + 2\log(1/\delta_{\min}^{\mathrm{pool}})}.$$

**Lemma 14** (Layerwise pooled elliptical potential)**.** *For every layer $\ell$,*

$$\sum_{t\in Q_\ell} \|x_{t,a_t}\|^2_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}} \le 2\log\frac{\det(A^{\mathrm{fin}}_\ell + G_{\mathrm{off}})}{\det(G_{\mathrm{off}})}.$$

*Proof.* Enumerate $Q_\ell$ as $t_1 < \cdots < t_{|Q_\ell|}$ and define

$$W_{\ell,0} := G_{\mathrm{off}}, \qquad W_{\ell,j} := G_{\mathrm{off}} + \sum_{i=1}^{j} x_{t_i,a_{t_i}} x_{t_i,a_{t_i}}^\top.$$

Then $W_{\ell,j-1} = A_{t_j,\ell} + G_{\mathrm{off}}$ and $W_{\ell,|Q_\ell|} = A^{\mathrm{fin}}_\ell + G_{\mathrm{off}}$. The matrix-determinant lemma yields

$$\det(W_{\ell,j}) = \det(W_{\ell,j-1})\left(1 + \|x_{t_j,a_{t_j}}\|^2_{W^{-1}_{\ell,j-1}}\right).$$

Taking logarithms and summing,

$$\log\frac{\det(A^{\mathrm{fin}}_\ell + G_{\mathrm{off}})}{\det(G_{\mathrm{off}})} = \sum_{j=1}^{|Q_\ell|} \log\left(1 + \|x_{t_j,a_{t_j}}\|^2_{W^{-1}_{\ell,j-1}}\right).$$

Since $G_{\mathrm{off}} \succeq I$ and $W_{\ell,j-1} \succeq G_{\mathrm{off}}$, we have $W^{-1}_{\ell,j-1} \preceq I$, hence for $\|x_{t_j,a_{t_j}}\|_2 \le 1$,

$$\|x_{t_j,a_{t_j}}\|^2_{W^{-1}_{\ell,j-1}} \le 1.$$

Therefore $\log(1+u) \ge u/2$ on this range, and the claim follows. □

**Corollary 5** (Pooled variance accumulation)**.** *We have*

$$\sum_{t=1}^{T} \|x_{t,a_t}\|_{(A_{t,\ell_t}+G_{\mathrm{off}})^{-1}} \le \sqrt{2T\,\widetilde{\Lambda}_T}.$$

*Proof.* For each layer,

$$\sum_{t\in Q_\ell} \|x_{t,a_t}\|_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}} \le \sqrt{|Q_\ell|}\left(\sum_{t\in Q_\ell} \|x_{t,a_t}\|^2_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}}\right)^{1/2}$$

by Cauchy–Schwarz. Apply Lemma 14, then sum over $\ell$ and apply Cauchy–Schwarz again across layers. □

*Proof of Theorem 1.* On $\mathcal{E}$, Proposition 3 gives the fallback term $C_{\mathrm{SL}}\sqrt{Td\,L_T^3}$. It therefore remains to prove the pooled term.

By Lemma 11,

$$R_T \le 8\sum_{t=1}^{T} \min\left\{1, \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t})\right\} \le 8\sum_{t=1}^{T} \mathrm{rad}^{\mathrm{pool}}_{t,\ell_t}(x_{t,a_t}),$$

since $\min\{1,r\} \le r$ for $r \ge 0$. Expanding the pooled radius,

$$R_T \le 8\sum_{\ell=0}^{\lceil\log_2 T\rceil} \sum_{t\in Q_\ell} \big(\gamma_{t,\ell_t} + \beta_{\mathrm{off}}\big)\|x_{t,a_t}\|_{(A_{t,\ell_t}+G_{\mathrm{off}})^{-1}} + 8\rho\sum_{t=1}^{T} \psi_{t,\ell_t}(x_{t,a_t}).$$

For every $t \in Q_\ell$, we have $\delta^{\mathrm{pool}}_{t,\ell} \ge \delta^{\mathrm{pool}}_{\ell,\min}$ and $A_{t,\ell} \preceq A^{\mathrm{fin}}_\ell$, so

$$\det(A_{t,\ell} + G_{\mathrm{off}}) \le \det(A^{\mathrm{fin}}_\ell + G_{\mathrm{off}})$$

and hence $\gamma_{t,\ell} \le \gamma^{\mathrm{fin}}_\ell$. Therefore

$$R_T \le 8\sum_{\ell=0}^{\lceil\log_2 T\rceil} \big(\beta_{\mathrm{off}} + \gamma^{\mathrm{fin}}_\ell\big) \sum_{t\in Q_\ell} \|x_{t,a_t}\|_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}} + 8\rho\,\widetilde{\Psi}_T.$$

For each layer, Cauchy–Schwarz and Lemma 14 give

$$\sum_{t\in Q_\ell} \|x_{t,a_t}\|_{(A_{t,\ell}+G_{\text{off}})^{-1}} \le \sqrt{|Q_\ell|}\left(\sum_{t\in Q_\ell} \|x_{t,a_t}\|^2_{(A_{t,\ell}+G_{\text{off}})^{-1}}\right)^{1/2} \le \sqrt{2|Q_\ell|\Lambda_\ell}.$$

Substituting this into the previous display proves the theorem's first pooled bound:

$$R_T \le 8 \sum_{\ell=0}^{\lceil \log_2 T\rceil} \left(\beta_{\text{off}} + \gamma_\ell^{\text{fin}}\right)\sqrt{2|Q_\ell|\Lambda_\ell} + 8\rho\,\widetilde{\Psi}_T.$$

For the coarser simplification, note that $\delta^{\text{pool}}_{\ell,\min} \ge \delta^{\text{pool}}_{\min}$ and $\Lambda_\ell \le \Lambda_{\max}$, so every $\gamma_\ell^{\text{fin}}$ is at most $\Gamma_{\text{pool}} - \beta_{\text{off}}$. Hence

$$\sum_{\ell=0}^{\lceil \log_2 T\rceil} \left(\beta_{\text{off}} + \gamma_\ell^{\text{fin}}\right)\sqrt{2|Q_\ell|\Lambda_\ell} \le \Gamma_{\text{pool}} \sum_{\ell=0}^{\lceil \log_2 T\rceil} \sqrt{2|Q_\ell|\Lambda_\ell}.$$

Applying Cauchy–Schwarz across layers and using $\sum_\ell |Q_\ell| = T$ yields

$$\sum_{\ell=0}^{\lceil \log_2 T\rceil} \sqrt{2|Q_\ell|\Lambda_\ell} \le \sqrt{2T\,\widetilde{\Lambda}_T},$$

which proves the theorem's second pooled bound. Both the fallback term and the pooled terms hold on the same event $\mathcal{E}$, so taking the minimum of the two completes the proof. □

# I Proof of Theorem 2

*Proof.* Write $L_{\max} = \lceil \log_2 T\rceil + 1$, matching the number of SupLinUCB layers indexed by $\ell \in \{0,\dots,\lceil \log_2 T\rceil\}$. Fix $\ell$ and diagonalize $G_{\text{off}} = U\,\mathrm{diag}(g_1,\dots,g_d)\,U^\top$ with $U$ orthogonal. Let $\widetilde{A}_\ell := U^\top A_\ell^{\text{fin}} U$. Hadamard's inequality gives

$$\det(A_\ell^{\text{fin}} + G_{\text{off}}) = \det(\widetilde{A}_\ell + \mathrm{diag}(g_1,\dots,g_d)) \le \prod_{j=1}^{d} (g_j + (\widetilde{A}_\ell)_{jj}).$$

Therefore

$$\Lambda_\ell = \log \frac{\det(A_\ell^{\text{fin}} + G_{\text{off}})}{\det(G_{\text{off}})} \le \sum_{j=1}^{d} \log\left(1 + \frac{(\widetilde{A}_\ell)_{jj}}{g_j}\right).$$

Because $\sum_j (\widetilde{A}_\ell)_{jj} = \mathrm{tr}(A_\ell^{\text{fin}}) \le |Q_\ell|$ and each played feature has $\|x_{t,a_t}\|_2 \le 1$, Definition 2 yields $\Lambda_\ell \le \Phi_{G_{\text{off}}}(|Q_\ell|)$. Summing over layers,

$$\widetilde{\Lambda}_T \le \sum_{\ell=0}^{\lceil \log_2 T\rceil} \Phi_{G_{\text{off}}}(|Q_\ell|).$$

The map $B \mapsto \Phi_G(B)$ is nondecreasing and concave in $B$: it is the value function of maximizing a sum of concave coordinate functions $\log(1 + a_j/g_j)$ subject to $a_j \ge 0$ and $\sum_j a_j \le B$. Hence Jensen's inequality implies

$$\frac{1}{L_{\max}} \sum_{\ell=0}^{\lceil \log_2 T\rceil} \Phi_{G_{\text{off}}}(|Q_\ell|) \le \Phi_{G_{\text{off}}}\left(\frac{1}{L_{\max}} \sum_{\ell=0}^{\lceil \log_2 T\rceil} |Q_\ell|\right) = \Phi_{G_{\text{off}}}\left(\frac{T}{L_{\max}}\right),$$

since $\sum_\ell |Q_\ell| = T$. Multiplying by $L_{\max}$ proves the spectral envelope bound on $\widetilde{\Lambda}_T$.

Define

$$c_{\text{align}} := \lambda_{\max}\left(G_{\text{off}}^{1/2} M_{\text{bias}}^{-1} G_{\text{off}}^{1/2}\right).$$

Then $G_{\mathrm{off}}^{1/2} M_{\mathrm{bias}}^{-1} G_{\mathrm{off}}^{1/2} \preceq c_{\mathrm{align}} I$, hence, conjugating by $G_{\mathrm{off}}^{1/2}$,

$$G_{\mathrm{off}} M_{\mathrm{bias}}^{-1} G_{\mathrm{off}} \preceq c_{\mathrm{align}} G_{\mathrm{off}}.$$

For the alignment bound on $\widetilde{\Psi}_T$, repeat the Cauchy–Schwarz argument used previously for the pooled regime: write $A := A_{t,\ell}$, $G := G_{\mathrm{off}}$, and $M := M_{\mathrm{bias}}$. Under $GM^{-1}G \preceq c_{\mathrm{align}} G$,

$$\psi_{t,\ell}(x)^2 \le c_{\mathrm{align}} \, \|x\|^2_{(A+G)^{-1}},$$

then sum along the trajectory and invoke Lemma 14 layer by layer to obtain $\sum_{t=1}^{T} \psi_{t,\ell_t}(x_{t,a_t})^2 \le 2c_{\mathrm{align}} \widetilde{\Lambda}_T$, hence $\widetilde{\Psi}_T \le \sqrt{2c_{\mathrm{align}} T \, \widetilde{\Lambda}_T}$.

Finally, substitute these inequalities into the pooled branch bound from Theorem 1:

$$R_T \le 8\Gamma_{\mathrm{pool}} \sqrt{2T \, \widetilde{\Lambda}_T} + 8\rho \sqrt{2c_{\mathrm{align}} T \, \widetilde{\Lambda}_T},$$

so

$$R_T \le 8\big(\Gamma_{\mathrm{pool}} + \rho\sqrt{c_{\mathrm{align}}}\big) \sqrt{2T \, \widetilde{\Lambda}_T}.$$

Since $\Lambda_\ell \le \Phi_{G_{\mathrm{off}}}(|Q_\ell|) \le \Phi_{G_{\mathrm{off}}}(T)$ for every layer, we also have $\Lambda_{\max} \le \Phi_{G_{\mathrm{off}}}(T)$ and therefore $\Gamma_{\mathrm{pool}} \le \Gamma_{\mathrm{spec}}$. Substituting this and the bound $\widetilde{\Lambda}_T \le L_{\max} \Phi_{G_{\mathrm{off}}}(T/L_{\max})$ proves the displayed theorem. The online fallback is unchanged from Theorem 1. □

*Proof of Corollary 1.* The first display is exactly the intermediate inequality obtained at the end of the proof of Theorem 2. Squaring the pooled term yields

$$\left(8\big(\Gamma_{\mathrm{pool}} + \rho\sqrt{c_{\mathrm{align}}}\big) \sqrt{2T \, \widetilde{\Lambda}_T}\right)^2 = 128 \, T \, \widetilde{\Lambda}_T \big(\Gamma_{\mathrm{pool}} + \rho\sqrt{c_{\mathrm{align}}}\big)^2.$$

Hence the stated condition implies that the pooled term is at most $C_{\mathrm{SL}} \sqrt{Td \, L_T^3}$, i.e. no larger than the online fallback. □

## J Proof of Theorem 3

*Proof.* It is enough to prove the theorem with a fixed absolute constant $m_0 = 32$. Fix any (possibly randomized) policy $\pi$, and fix any subset $S \subset [d]$ with

$$m := |S| \ge 32.$$

Write

$$N := N(S) = \sum_{i \in S} n_i, \qquad \underline{v} := v(S) = \min_{i \in S} \frac{\rho}{\sqrt{b_i}}.$$

**Case 1: $N = 0$.** Since $n_i = 0$ for every $i \in S$, the offline sample carries no information about the coordinates in $S$. Consider the baseline instance

$$P_0 : \qquad \theta^* = \theta^\dagger = 0,$$

and, for each $j \in S$, the alternative

$$P_j : \qquad \theta_j^* = \theta_j^\dagger = \Delta, \qquad \theta_i^* = \theta_i^\dagger = 0 \quad \text{for } i \ne j.$$

Each $P_j$ belongs to $\mathcal{H}_{\mathrm{diag}}(G_{\mathrm{off}}, M_{\mathrm{bias}}, \rho)$ because the bias is zero.

Let

$$N_j := \sum_{t=1}^{T} \mathbf{1}\{a_t = j\}.$$

Under $P_0$, there exists $j^\star \in S$ with

$$\mathbb{E}_{P_0}[N_{j^\star}] \le \frac{T}{m},$$

since $\sum_{j\in S}\mathbb{E}_{P_0}[N_j] \le T$. Define

$$E := \{N_{j^\star} > T/2\}.$$

Then Markov's inequality gives

$$\Pr_{P_0}(E) \le \frac{\mathbb{E}_{P_0}[N_{j^\star}]}{T/2} \le \frac{2}{m} \le \frac{1}{16}.$$

The offline laws of $P_0$ and $P_{j^\star}$ are identical, and the online KL is

$$\mathrm{KL}(P_0, P_{j^\star}) = \frac{\Delta^2}{2}\mathbb{E}_{P_0}[N_{j^\star}] \le \frac{\Delta^2 T}{2m}.$$

Choose

$$\Delta := \frac{1}{8}\sqrt{\frac{m}{T}}.$$

Then $\mathrm{KL}(P_0, P_{j^\star}) \le 1/128 < 1/100$. By Bretagnolle–Huber,

$$\Pr_{P_{j^\star}}(E^c) \ge \frac{1}{2}e^{-1/100} - \frac{1}{16} > c_0$$

for a universal constant $c_0 > 0$.

Under $P_{j^\star}$, arm $j^\star$ is uniquely optimal with gap $\Delta$. On $E^c$, the optimal arm is pulled at most $T/2$ times, so at least $T/2$ rounds incur gap $\Delta$. Hence

$$\mathbb{E}_{P_{j^\star}}[R_T] \ge \frac{\Delta T}{2}\Pr_{P_{j^\star}}(E^c) \ge c_1\sqrt{mT}.$$

When $N = 0$,

$$\Psi(S) = \sqrt{\frac{mT^2}{T+N}} + \min\left\{\frac{NT}{T+N}\underline{v},\ \sqrt{\frac{mNT}{T+N}}\right\} = \sqrt{mT},$$

so

$$\sup_{I\in\mathcal{H}_{\mathrm{diag}}(G_{\mathrm{off}},M_{\mathrm{bias}},\rho)} \mathbb{E}_I[R_T] \ge c_1\Psi(S).$$

**Case 2:** $N > 0$**.** Define the clipped bias size

$$\bar{v} := \min\left\{\underline{v},\ \frac{1}{32}\sqrt{\frac{m(T+N)}{NT}}\right\}.$$

Consider the baseline instance

$$P_0: \qquad \theta^* = \theta^\dagger = 0,$$

and, for each $j \in S$, the alternative

$$P_j: \qquad \theta^*_j = \mu, \qquad \theta^\dagger_j = \mu - \bar{v}, \qquad \theta^*_i = \theta^\dagger_i = 0 \quad \text{for } i \ne j,$$

where

$$\mu := \frac{N}{N+T}\bar{v} + \frac{1}{64}\sqrt{\frac{m}{N+T}}.$$

Since $\bar{v} \le \underline{v} \le \rho/\sqrt{b_j}$ for every $j \in S$, each $P_j$ belongs to $\mathcal{H}_{\mathrm{diag}}(G_{\mathrm{off}}, M_{\mathrm{bias}}, \rho)$.

Again write

$$N_j := \sum_{t=1}^{T}\mathbf{1}\{a_t = j\}.$$

Under $P_0$, define

$$B_1 := \{j \in S:\ n_j > 4N/m\}, \qquad B_2 := \{j \in S:\ \mathbb{E}_{P_0}[N_j] > 4T/m\}.$$

Since

$$\sum_{j\in S} n_j = N \qquad \text{and} \qquad \sum_{j\in S}\mathbb{E}_{P_0}[N_j] \le T,$$

we have

$$|B_1| \le \frac{m}{4}, \qquad |B_2| \le \frac{m}{4}.$$

Therefore there exists an arm $j^\star \in S \setminus (B_1 \cup B_2)$ such that

$$n_{j^\star} \le \frac{4N}{m}, \qquad \mathbb{E}_{P_0}[N_{j^\star}] \le \frac{4T}{m}. \tag{10}$$

Only arm $j^\star$ differs between $P_0$ and $P_{j^\star}$. Since offline and online rewards are Gaussian with unit variance,

$$\mathrm{KL}(P_0, P_{j^\star}) = \frac{1}{2}\left(n_{j^\star}(\mu - \bar{v})^2 + \mathbb{E}_{P_0}[N_{j^\star}]\mu^2\right).$$

Using (10),

$$\mathrm{KL}(P_0, P_{j^\star}) \le \frac{1}{2}\left(\frac{4N}{m}(\mu - \bar{v})^2 + \frac{4T}{m}\mu^2\right).$$

Set

$$a := \frac{4N}{m}, \qquad b := \frac{4T}{m}.$$

Then

$$\mu = \frac{a}{a+b}\bar{v} + \frac{1}{32\sqrt{a+b}},$$

because $a + b = 4(N+T)/m$. A direct expansion gives

$$a(\mu - \bar{v})^2 + b\mu^2 = \frac{ab}{a+b}\bar{v}^2 + \frac{1}{32^2}.$$

Also,

$$\frac{ab}{a+b} = \frac{4NT}{m(N+T)}$$

and, by the definition of $\bar{v}$,

$$\bar{v}^2 \le \frac{1}{32^2}\frac{m(T+N)}{NT}.$$

Hence

$$\frac{ab}{a+b}\bar{v}^2 \le \frac{4}{32^2},$$

so

$$\mathrm{KL}(P_0, P_{j^\star}) \le \frac{1}{2}\left(\frac{4}{32^2} + \frac{1}{32^2}\right) = \frac{5}{2048} < \frac{1}{100}. \tag{$*$}$$

Define

$$E := \{N_{j^\star} > T/2\}.$$

By Markov's inequality and (10),

$$\Pr_{P_0}(E) \le \frac{\mathbb{E}_{P_0}[N_{j^\star}]}{T/2} \le \frac{8}{m} \le \frac{1}{4},$$

since $m \ge 32$. Bretagnolle–Huber and $(*)$ give

$$\Pr_{P_{j^\star}}(E^c) \ge \frac{1}{2}e^{-1/100} - \frac{1}{4} > c_2$$

for a universal constant $c_2 > 0$.

In instance $P_{j^\star}$, arm $j^\star$ is uniquely optimal with gap $\mu$. On $E^c$, the optimal arm is pulled at most $T/2$ times, so at least $T/2$ rounds incur gap $\mu$. Therefore

$$\mathbb{E}_{P_{j^\star}}[R_T] \ge \frac{\mu T}{2}\Pr_{P_{j^\star}}(E^c) \ge c_3\,\mu T.$$

Finally,

$$\mu T = \frac{NT}{N+T}\bar{v} + \frac{1}{64}\sqrt{\frac{mT^2}{N+T}},$$

**Algorithm 2** Data-Driven Directional Ellipsoidal-MINUCB

**Input:** $D_{\mathrm{off}}$, $T$, prespecified $1{=}\tau_1{<}\cdots{<}\tau_{K_{\mathrm{ep}}}{\le}T$ and $\tau_{K_{\mathrm{ep}}+1}{=}T{+}1$; confidence schedules $\{\delta^{\mathrm{on}}_{t,\ell}\}$, $\{\delta^{\mathrm{pool}}_{t,\ell}\}$ as in Algorithm 1, $\delta_{\mathrm{off}}$, and $\{\delta_k\}_{k=1}^{K_{\mathrm{ep}}}$ for $\beta^{\mathrm{on}}_k(\delta_k)$, satisfying $\delta_{\mathrm{off}}+\sum_{k=1}^{K_{\mathrm{ep}}}\delta_k{\le}\delta_{\mathrm{bias}}$

Build $G_{\mathrm{off}}$, $\hat{\theta}_{\mathrm{off}}$, and $\beta_{\mathrm{off}}(\delta_{\mathrm{off}})$ from $D_{\mathrm{off}}$
$L \leftarrow \lceil \log_2 T \rceil$; initialize $\mathcal{H}_{1,\ell} \leftarrow \varnothing$ for $\ell = 0, 1, \ldots, L$
**for** each epoch $k = 1, 2, \ldots, K_{\mathrm{ep}}$ **do**
  $G^{\mathrm{on}}_k \leftarrow I + \sum_{s<\tau_k} x_{s,a_s} x_{s,a_s}^{\top}$;  $\hat{\theta}^{\mathrm{on}}_k \leftarrow (G^{\mathrm{on}}_k)^{-1} \sum_{s<\tau_k} x_{s,a_s} r_s$
  $\widehat{M}_{k,\mathrm{bias}} \leftarrow \big((G^{\mathrm{on}}_k)^{-1} + G_{\mathrm{off}}^{-1}\big)^{-1}$;  $\widehat{\Delta}_k \leftarrow \hat{\theta}^{\mathrm{on}}_k - \hat{\theta}_{\mathrm{off}}$
  $\widehat{\rho}_k \leftarrow \|\widehat{\Delta}_k\|_{\widehat{M}_{k,\mathrm{bias}}} + \beta^{\mathrm{on}}_k(\delta_k) + \beta_{\mathrm{off}}(\delta_{\mathrm{off}})$
  **for** each $t = \tau_k, \ldots, \tau_{k+1} - 1$ **do**
    Execute the per-round layer loop of Algorithm 1 at time $t$, using (12)–(11) in place of $\mathrm{rad}^{\mathrm{pool}}_{t,\ell}(\cdot)$ when forming the pooled UCB/LCB
  **end for**
**end for**

and by the definition of $\bar{v}$,

$$\frac{NT}{N+T}\bar{v} = \min\left\{\frac{NT}{N+T}\,\underline{v},\ \frac{1}{32}\sqrt{\frac{mNT}{N+T}}\right\}.$$

Hence

$$\sup_{I\in\mathcal{H}_{\mathrm{diag}}(G_{\mathrm{off}},M_{\mathrm{bias}},\rho)} \mathbb{E}_I[R_T] \ge c_4\left[\sqrt{\frac{mT^2}{T+N}} + \min\left\{\frac{NT}{T+N}\,\underline{v},\ \sqrt{\frac{mNT}{T+N}}\right\}\right] = c_4\,\Psi(S).$$

Both cases show that, for every subset $S \subset [d]$ with $|S| \ge 32$,

$$\sup_{I\in\mathcal{H}_{\mathrm{diag}}(G_{\mathrm{off}},M_{\mathrm{bias}},\rho)} \mathbb{E}_I[R_T] \ge c_5\,\Psi(S) \ge c_5 \min\{\sqrt{dT}, \Psi(S)\},$$

since $\min\{\sqrt{dT}, \Psi(S)\} \le \Psi(S)$. Taking the maximum over all such subsets proves the theorem with $m_0 = 32$. □

# K Data-driven directional bias certificates: definitions and proofs (Section 5)

## K.1 Definitions and Algorithm

In this section, we give the detailed definitions and proofs in Section 5. We first define the epoch-wise pooled width and the full algorithm, then record the geometric identity behind the learned certificate, and finally prove the validity and regret guarantees together with the doubling-epoch refinement. The pseudocode of the algorithms in given in Algorithm 2.

On rounds $t \in \{\tau_k, \ldots, \tau_{k+1} - 1\}$ one replaces $\rho\,\psi_{t,\ell}$ of the main text by $\widehat{\rho}_k\,\widehat{\psi}_{k,t,\ell}$ with

$$\widehat{\psi}_{k,t,\ell}(x) := \left\| \widehat{M}_{k,\mathrm{bias}}^{-1/2} G_{\mathrm{off}} (A_{t,\ell} + G_{\mathrm{off}})^{-1} x \right\|_2 \tag{11}$$

and

$$\widehat{\mathrm{rad}}^{\mathrm{pool,ep}}_{t,\ell}(x) := \big(\gamma_{t,\ell} + \beta_{\mathrm{off}}\big)\|x\|_{(A_{t,\ell}+G_{\mathrm{off}})^{-1}} + \widehat{\rho}_k\,\widehat{\psi}_{k,t,\ell}(x). \tag{12}$$

## K.2 A Geometric Identity for the Parallel Sum

We begin with the geometric identity underlying the learned certificate.

**Proposition 4** (Geometric characterization of the parallel sum)**.** *For any epoch $k$ and any vector $x \in \mathbb{R}^d$,*

$$\|x\|^2_{\widehat{M}_{k,\mathrm{bias}}} = \inf_{x=u+v}\Big\{\|u\|^2_{G^{\mathrm{on}}_k} + \|v\|^2_{G_{\mathrm{off}}}\Big\}.$$

*Proof.* Write $A := G_k^{\mathrm{on}}$ and $B := G_{\mathrm{off}}$ for brevity, and let $M := (A^{-1} + B^{-1})^{-1}$. For any decomposition $x = u + v$,

$$\|u\|_A^2 + \|v\|_B^2 = u^\top A u + (x-u)^\top B (x-u).$$

This is a strictly convex quadratic in $u$, whose unique minimizer satisfies

$$(A+B)u = Bx, \qquad \text{hence} \qquad u = (A+B)^{-1}Bx, \quad v = (A+B)^{-1}Ax.$$

Substituting these expressions gives

$$\inf_{x=u+v} \left\{ \|u\|_A^2 + \|v\|_B^2 \right\} = x^\top B(A+B)^{-1}Ax.$$

It remains to identify the matrix on the right with $M$. Since

$$A^{-1} + B^{-1} = A^{-1}(A+B)B^{-1},$$

we have

$$M = (A^{-1} + B^{-1})^{-1} = B(A+B)^{-1}A.$$

Therefore the minimum equals

$$x^\top M x = \|x\|_M^2 = \|x\|_{\widehat{M}_{k,\mathrm{bias}}}^2,$$

which proves the claim. □

### K.3 Validity of the Learned Directional Certificate

We next verify that the learned pair $(\widehat{M}_{k,\mathrm{bias}}, \widehat{\rho}_k)$ defines a valid epoch-wise directional certificate.

**Theorem 6** (Validity of the data-driven directional bias certificate)**.** *Suppose the standard offline ridge and epoch-wise online events hold, i.e.* $\|\theta^\dagger - \hat{\theta}_{\mathrm{off}}\|_{G_{\mathrm{off}}} \le \beta_{\mathrm{off}}(\delta_{\mathrm{off}})$ *and, for every* $k$, $\|\theta_* - \hat{\theta}_k^{\mathrm{on}}\|_{G_k^{\mathrm{on}}} \le \beta_k^{\mathrm{on}}(\delta_k)$. *If* $\delta_{\mathrm{off}} + \sum_{k=1}^{K_{\mathrm{ep}}} \delta_k \le \delta_{\mathrm{bias}}$ *then, with probability at least* $1 - \delta_{\mathrm{bias}}$*,*

$$\|\theta_* - \theta^\dagger\|_{\widehat{M}_{k,\mathrm{bias}}} \le \widehat{\rho}_k \qquad \forall k \in \{1, \dots, K_{\mathrm{ep}}\}.$$

*Proof.* With probability at least $1 - \delta_{\mathrm{off}}$, $\|\theta^\dagger - \hat{\theta}_{\mathrm{off}}\|_{G_{\mathrm{off}}} \le \beta_{\mathrm{off}}(\delta_{\mathrm{off}})$. For each epoch $k$, the estimator $\hat{\theta}_k^{\mathrm{on}}$ uses only data strictly before $\tau_k$; the self-normalized ridge inequality at time $\tau_k$ yields $\|\theta_* - \hat{\theta}_k^{\mathrm{on}}\|_{G_k^{\mathrm{on}}} \le \beta_k^{\mathrm{on}}(\delta_k)$ on an event of probability at least $1 - \delta_k$ (in the predictable epoch case, the same time-uniform bound or a union over $t$ and evaluation at $\tau_k$ applies). A union bound gives simultaneous validity with probability at least $1 - \delta_{\mathrm{bias}}$ when $\delta_{\mathrm{off}} + \sum_k \delta_k \le \delta_{\mathrm{bias}}$.

On this event, for fixed $k$, the parallel-sum definition implies $\widehat{M}_{k,\mathrm{bias}} \preceq G_k^{\mathrm{on}}$ and $\widehat{M}_{k,\mathrm{bias}} \preceq G_{\mathrm{off}}$ because $(G_k^{\mathrm{on}})^{-1} + G_{\mathrm{off}}^{-1} \succeq (G_k^{\mathrm{on}})^{-1}$ and the symmetric inequality for $G_{\mathrm{off}}$. Hence, for all $v$, $\|v\|_{\widehat{M}_{k,\mathrm{bias}}} \le \|v\|_{G_k^{\mathrm{on}}}$ and $\|v\|_{\widehat{M}_{k,\mathrm{bias}}} \le \|v\|_{G_{\mathrm{off}}}$. Decompose $\theta_* - \theta^\dagger = \widehat{\Delta}_k + (\theta_* - \hat{\theta}_k^{\mathrm{on}}) - (\theta^\dagger - \hat{\theta}_{\mathrm{off}})$ with $\widehat{\Delta}_k := \hat{\theta}_k^{\mathrm{on}} - \hat{\theta}_{\mathrm{off}}$. The triangle inequality in the $\widehat{M}_{k,\mathrm{bias}}$ norm, together with the confidence events and Loewner domination, gives

$$\|\theta_* - \theta^\dagger\|_{\widehat{M}_{k,\mathrm{bias}}} \le \|\widehat{\Delta}_k\|_{\widehat{M}_{k,\mathrm{bias}}} + \beta_k^{\mathrm{on}}(\delta_k) + \beta_{\mathrm{off}}(\delta_{\mathrm{off}}) = \widehat{\rho}_k.$$

Since $k$ is arbitrary, the claim holds for all $k$. □

### K.4 Proof of Theorem 4

*Proof of Theorem 4.* The online branch is unchanged, so the ordinary fallback regret term $C_{\mathrm{SL}}\sqrt{T d L_T^3}$ and its proof are unchanged. For the pooled branch, on the event of Theorem 6,

$$\|\theta_* - \theta^\dagger\|_{\widehat{M}_{k(t),\mathrm{bias}}} \le \widehat{\rho}_{k(t)} \qquad \forall t.$$

The epoch start times are part of the prespecified (or past-measurable) rule, and $(\widehat{M}_{k,\mathrm{bias}}, \widehat{\rho}_k)$ uses only data before $\tau_k$; therefore the pair used in round $t$ is $\mathcal{F}_{t-1}$-measurable and the policy remains non-anticipatory, so the martingale and self-normalized concentration events from the proof of Theorem 1 are unchanged. On the directional bias certificate event, for every round-$t$ queried direction $x$,

$$|x^\top(\theta_* - \theta^\dagger)| \le \widehat{\rho}_{k(t)} \|x\|_{\widehat{M}_{k(t),\mathrm{bias}}^{-1}}.$$

In the pooled confidence step (Lemma 7 in the main appendix proof), the fixed bias width $\rho\,\|M_{\mathrm{bias}}^{-1/2}G_{\mathrm{off}}(A_{t,\ell}+G_{\mathrm{off}})^{-1}x\|_2$ is therefore replaced by $\widehat{\rho}_{k(t)}\big\|\widehat{M}_{k(t),\mathrm{bias}}^{-1/2}G_{\mathrm{off}}(A_{t,\ell}+G_{\mathrm{off}})^{-1}x\big\|_2$, i.e., $\widehat{\rho}_{k(t)}\widehat{\psi}_{k(t),t,\ell}(x)$. The UCB/LCB validity, elimination, and stopping-layer arguments from the proof of Theorem 1 then go through unchanged except that the routed bias sum is $\widetilde{\Psi}_T^{\mathrm{ep}}$, while the statistical (non-bias) part of the pooled width still produces the $8\sum_\ell(\beta_{\mathrm{off}}+\gamma_\ell^{\mathrm{fin}})\sqrt{2|Q_\ell|\,\Lambda_\ell}$ term by the same pooled elliptical potential calculation. Intersecting the directional bias certificate event with the original $1-T^{-2}$-probability event for the online, offline, and pooled self-normalized events yields the stated probability and regret bound. □

### K.5 Doubling-Epoch Refinement

We now specialize to a standard doubling schedule in order to obtain a closed-form refinement.

A default is $K_{\mathrm{ep}} := \lfloor \log_2 T \rfloor + 1$, $\tau_k := 2^{k-1}$, $\tau_{K_{\mathrm{ep}}+1} := T+1$. To combine Theorems 6–4 with closed-form radii, split $\delta_{\mathrm{off}} := \delta_{\mathrm{bias}}/2$ and $\delta_k := \delta_{\mathrm{bias}}/(2K_{\mathrm{ep}})$ and set

$$\beta_k^{\mathrm{on}} := \sigma\sqrt{d\log(1+\tau_k) + 2\log(2K_{\mathrm{ep}}/\delta_{\mathrm{bias}})}+S, \qquad \beta_{\mathrm{off}}^{\mathrm{bias}} := \sigma\sqrt{2\log\big(\det(G_{\mathrm{off}})^{1/2}/(\delta_{\mathrm{bias}}/2)\big)}+S.$$

In each epoch, with $\widehat{\Delta}_k$ and $\widehat{M}_{k,\mathrm{bias}}$ built from $D_{\mathrm{off}}$ and online data before $\tau_k$ (as in Algorithm 2), set $\widehat{\rho}_k := \|\widehat{\Delta}_k\|_{\widehat{M}_{k,\mathrm{bias}}} + \beta_k^{\mathrm{on}} + \beta_{\mathrm{off}}^{\mathrm{bias}}$.

**Theorem 7** (High probility bound for the data-driven directional bias certificate)**.** *Under the doubling schedule, $\delta$-split, and $(\widehat{\rho}_k, \widehat{M}_{k,\mathrm{bias}})$ from the previous paragraph, define $n_k := \tau_{k+1}-\tau_k$,*

$$\widehat{c}_k := \lambda_{\max}\big(G_{\mathrm{off}}^{1/2}\widehat{M}_{k,\mathrm{bias}}^{-1}G_{\mathrm{off}}^{1/2}\big), \qquad \mathcal{S}_{\mathrm{ep}} := \sum_{k=1}^{K_{\mathrm{ep}}} n_k\,\widehat{\rho}_k^2\,\widehat{c}_k.$$

*With probability at least $1-T^{-2}-\delta_{\mathrm{bias}}$, the regret satisfies*

$$R_T \le \min\left\{C_{\mathrm{SL}}\sqrt{Td\,L_T^3},\ 8\Gamma_{\mathrm{pool}}\sqrt{2T\,\widetilde{\Lambda}_T} + 8\sqrt{2\,\widetilde{\Lambda}_T\,\mathcal{S}_{\mathrm{ep}}}\right\},$$

*and also the same bound with $8\Gamma_{\mathrm{pool}}\sqrt{2T\,\widetilde{\Lambda}_T}$ replaced by the full layerwise statistical term $8\sum_\ell(\beta_{\mathrm{off}}+\gamma_\ell^{\mathrm{fin}})\sqrt{2|Q_\ell|\,\Lambda_\ell}$, where $\Gamma_{\mathrm{pool}}$ is as in Theorem 1.*

*Proof of Theorem 7.* This step specializes the epoch-wise bound to the doubling schedule introduced above. Start from Theorem 4. For $t$ in epoch $k$, write $A := A_{t,\ell_t}$ and $G := G_{\mathrm{off}}$. By definition of $\widehat{\psi}$,

$$\widehat{\psi}_{k,t,\ell_t}(x_{t,a_t})^2 = x_{t,a_t}^\top(A+G)^{-1}G\widehat{M}_{k,\mathrm{bias}}^{-1}G(A+G)^{-1}x_{t,a_t}.$$

The choice $\widehat{c}_k := \lambda_{\max}(G^{1/2}\widehat{M}_{k,\mathrm{bias}}^{-1}G^{1/2})$ is equivalent to $G\widehat{M}_{k,\mathrm{bias}}^{-1}G \preceq \widehat{c}_k\,G$ in Loewner order, hence $\widehat{\psi}_{k,t,\ell_t}(x_{t,a_t})^2 \le \widehat{c}_k\,x_{t,a_t}^\top(A+G)^{-1}G(A+G)^{-1}x_{t,a_t}$. Since $G \preceq A+G$, $(A+G)^{-1}G(A+G)^{-1} \preceq (A+G)^{-1}$ and

$$\widehat{\psi}_{k,t,\ell_t}(x_{t,a_t}) \le \sqrt{\widehat{c}_k}\,\|x_{t,a_t}\|_{(A_{t,\ell_t}+G_{\mathrm{off}})^{-1}}.$$

Therefore

$$\widetilde{\Psi}_T^{\mathrm{ep}} \le \sum_{t=1}^T \widehat{\rho}_{k(t)}\sqrt{\widehat{c}_{k(t)}}\,\|x_{t,a_t}\|_{(A_{t,\ell_t}+G_{\mathrm{off}})^{-1}}.$$

By Cauchy–Schwarz,

$$\widetilde{\Psi}_T^{\mathrm{ep}} \le \Big(\sum_{t=1}^T \widehat{\rho}_{k(t)}^2\widehat{c}_{k(t)}\Big)^{1/2}\Big(\sum_{t=1}^T \|x_{t,a_t}\|^2_{(A_{t,\ell_t}+G_{\mathrm{off}})^{-1}}\Big)^{1/2}.$$

The first factor equals $\sqrt{\mathcal{S}_{\mathrm{ep}}}$ because $(\widehat{\rho}_k, \widehat{c}_k)$ are constant on each epoch. The second is at most $\sqrt{2\,\widetilde{\Lambda}_T}$ by Lemma 14 and the definition $\widetilde{\Lambda}_T = \sum_\ell \Lambda_\ell$ in Theorem 1. Thus $\widetilde{\Psi}_T^{\mathrm{ep}} \le$

$\sqrt{2\,\widetilde{\Lambda}_T\,\mathcal{S}_{\mathrm{ep}}}$. Substitute into the minimum expression of Theorem 4 and use the Cauchy–Schwarz step $\sum_\ell(\beta_{\mathrm{off}}+\gamma_\ell^{\mathrm{fin}})\sqrt{2|Q_\ell|\,\Lambda_\ell} \le \Gamma_{\mathrm{pool}}\sqrt{2T\,\widetilde{\Lambda}_T}$ as in the proof of Theorem 1 to obtain the two displayed bounds. □

*Proof of Corollary 2.* Theorem 7 gives

$$R_T \le \min\left\{C_{\mathrm{SL}}\sqrt{Td\,L_T^3},\ 8\Gamma_{\mathrm{pool}}\sqrt{2T\,\widetilde{\Lambda}_T}+8\sqrt{2\,\widetilde{\Lambda}_T\,\mathcal{S}_{\mathrm{ep}}}\right\}$$

with probability at least $1-T^{-2}-\delta_{\mathrm{bias}}$. Factor the pooled term as

$$8\Gamma_{\mathrm{pool}}\sqrt{2T\,\widetilde{\Lambda}_T}+8\sqrt{2\,\widetilde{\Lambda}_T\,\mathcal{S}_{\mathrm{ep}}} = 8\Big(\Gamma_{\mathrm{pool}}+\sqrt{\mathcal{S}_{\mathrm{ep}}/T}\Big)\sqrt{2T\,\widetilde{\Lambda}_T},$$

which is exactly the displayed bound in the corollary. □

## L Additional Experiments

All experiments use $K$=5 arms, $d$=5 features, horizon $T$=20000, sub-Gaussian reward noise $\sigma$=0.1, and $n_{\mathrm{off}}$=2000 offline points unless stated otherwise. Contexts are independent Gaussian draws in $\mathbb{R}^d$ for each arm and round, normalized to the Euclidean unit ball. We set ridge $\lambda_{\mathrm{reg}}$=1, confidence budget $\delta$=0.01, norm bound $S$=1, and offline covariate diversity multiplier div=1.0. Reported curves are means with variability bands over $n_{\mathrm{runs}}$=50 trials with base seed 42. The figures are multi-panel versions of the same protocol, with one panel per setting. All experiments run on a single CPU core.

**Randomized bias composition.** The online parameter is drawn with independent coordinates in $[0.25, 0.55]$. The offline mean is $\theta^\dagger = s\,\theta_* + u$, where $s \in (0,1]$ and $u$ is uniform on $[0,c]^d$ with $c \in \{0, 0.2, 0.4, 0.6, 0.8\}$. The pairs $(s,c)$ are chosen so that as $c$ increases, the additive offset $u$ has larger range while the scale factor $s$ toward $\theta_*$ decreases; at $c=0$ one has $s=1$ and $u=0$ (a pure re-scaling parameterization of the online vector). Figure 2 stacks these five misspecification strengths; we plot Ellipsoidal-MINUCB and the best purely offline policy on the same axis for each panel.

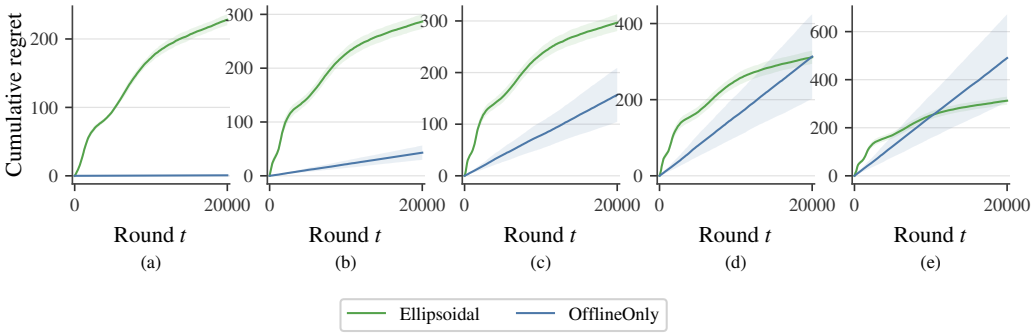



Figure 2: Randomized bias sweep: for each $c$, mean cumulative regret under the $(s,u)$ law above (left to right, increasing $c$; five panels in the published figure).

**Fixed online parameter, varying offline mean on one coordinate.** We fix $\theta_* = (2,1,1,1,1)$ and change only $\theta_1^\dagger$ among $\{0,1,2,3,4\}$ (with the remaining coordinates 1 in each case). Figure 3 compares the full method to two stripped variants: one that only uses the offline design matrix in the online branch, and one that only uses the online ridge design (no offline pooling in the same sense as the main text).

**Covariate shift on offline data only.** The online and offline parameters are generated from the same random law, but offline covariates are drawn from an axis-aligned box $\prod_j[\ell_j, h_j]$ inside the support used at play time. In particular, the first feature uses $[0,1]$, and coordinates $2,\dots,d$ use $[0,w]$

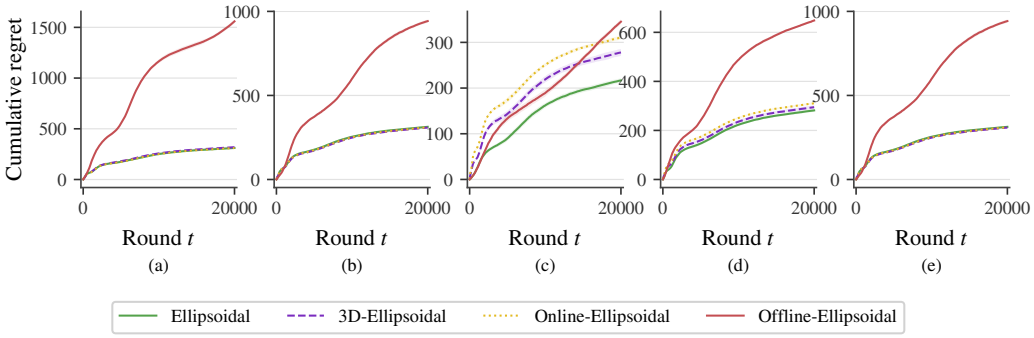



Figure 3: Ablation: five values of $\theta_1^\dagger$ with $\theta_*$ as above, same hyperparameters otherwise.

with $w$ in a finite decreasing set that includes 1, a moderate value, and small values down to $0.01$. For small $w$ the offline design places almost no energy on several directions, so the pooled statistics in those directions are driven almost entirely by the online data even when $\theta^\dagger = \theta_*$ on average. Figure 4 reports the same algorithm comparison as in Figure 2, here indexed by $w$.

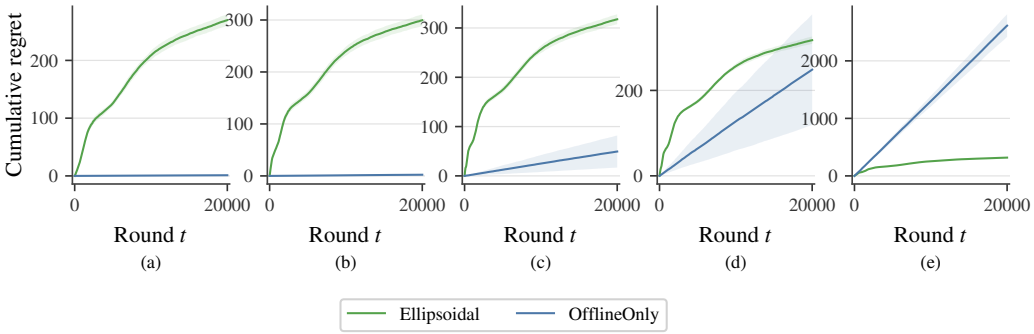



Figure 4: Offline-coverage study: each panel is a value of $w$ in the family above (narrow offline support on several coordinates, wide online support; five panels in the published figure).

## M Broader Impact

This work studies how to reuse offline data in linear contextual bandits when the offline and online environments may differ. A potential positive impact is that safer reuse of historical data can reduce online exploration, which may lower deployment cost and reduce exposure to poor actions in applications such as recommendation, advertising, and pricing. A key risk is that if offline bias is underestimated, the method may over-trust historical data and propagate existing biases or spurious correlations. Our framework partly mitigates this through conservative confidence bounds and an explicit fallback to the online branch, but practical use in high-stakes settings still requires careful auditing and monitoring.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the abstract and Section 1, and are matched by the formal results in Section 4 and the supporting experiments in Section 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses the main modeling and theoretical limitations in Section 8 and Appendix M. Future work includes a refined lower bound on the non-daigonal action transfer hardness.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in Section 2 and Section 4. Full proofs for the stated theoretical results are provided in the appendix, with a short interpretation in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 6 and Appendix L describe all our synthetic setups and the quantities reported in the figures.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The current draft uses repository-local scripts and result files, but it does not yet provide a finalized public release link or archival package for code and data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 and Appendix L states the experimental setup, the compared regrets and the interpretation of all figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main-text figures report mean cumulative regret curves over repeated runs with shaded variability bands, and the text explains what is being compared in each plot.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix L reports that all experiments run on a single CPU core.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is a theoretical and simulation-based study and does not involve human subjects, sensitive deployment, or model release practices that conflict with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix M includes a paragraph discussing the impact of safer offline-data reuse in deployed sequential decision systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets and does not study a misuse-sensitive generative system.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper uses no external datasets, pre-trained models, or proprietary code assets; experiments rely only on standard open-source Python libraries used under their respective permissive open-source licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new public dataset, model, or benchmark asset for release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the proposed bandit method, theory, or experimental pipeline described in the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.