CATCHING A MOVING SUBSPACE: LOW-RANK BANDITS BEYOND STATIONARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Many real bandit deployments (recommendation, ad targeting) share two structural facts handled only in isolation by prior work. Rewards live on a low-dimensional latent subspace, and that subspace drifts. Stationary low-rank bandits exploit the rank but assume a fixed subspace. Nonstationary linear bandits adapt to drift but their rates grow with the ambient dimension d. We study *piecewise-stationary low-rank* linear contextual bandits with scalar rewards whose parameter at round t factors as $\theta_t = B_k^\star w_t$, where $B_k^\star \in \mathbb{R}^{d\times r}$ is an orthonormal basis for the rank-r latent reward subspace in segment k and $w_t \in \mathbb{R}^r$ is the time-varying latent state. The subspace is constant within each of K unknown segments and may shift at the boundaries. The learner is a bandit with a sensing channel. Besides choosing actions, it may spend a round on a *probe*, a query drawn from a known exploration law that need not be a feasible action, and it pays for each probe its forgone reward plus a fee c. We give three results. **(i) The identification boundary** (when the moving subspace can be recovered). With one scalar reward per round, isotropic probes drawn independently of the noise recover the moving subspace, or the part of it that the latent state excites, through quadratic functionals of rewards. The noise variance need not be known. Probes that miss a direction cannot recover the subspace. If the noise may depend on the probe direction, identification can fail even with full coverage. **(ii) Algorithm and dynamic regret. SPSC** (Single-Play Subspace-Calibrated Optimism) interleaves isotropic probes with optimistic windowed ridge regression inside the learned r-dimensional subspace, given the segment boundaries. An adaptive variant detects them online. Measured against the best action for the mean reward vector, with probe costs charged, the dynamic regret over T rounds is $\widetilde{\mathcal{O}}(r\sqrt{KT} + (rd)^{1/3}K^{1/3}T^{2/3})$ when the mean reward vector is constant within segments. The $\sqrt{T}$ term depends on the rank r, and the ambient dimension enters only the additive cost of learning the subspace. Drift within segments adds the usual window tradeoff. **(iii) Empirics.** Against eleven baselines on nine benchmarks, including UCI datasets, images, MovieLens and production logs, an SPSC variant has the lowest regret in 51 of 58 real-data (dataset, d, r) cells. On a 40-cell synthetic grid, a rule of thumb suggested by the theory, that SPSC wins once $d - r > T^{1/6}$, picks the winner in 39 cells.

1 INTRODUCTION

Modern decision systems, including recommendation engines, ad allocation, and adaptive experiments, choose among high-dimensional actions and observe a single scalar outcome per round. Across these domains the useful signal concentrates on a far lower-dimensional latent structure. Each action is a vector of d features, but user preferences and ad-creative effects occupy a thin slice of the ambient space $\mathbb{R}^d$, a latent subspace of intrinsic rank $r \ll d$. Mining that structure is the difference between decision cost that scales with d and one that scales with r.

The structure does not stand still. Tastes drift and campaigns are replaced. The latent subspace itself shifts at unknown change points. Spotify's homepage contextual bandit targets daily shifts in user intent (Feijer et al., 2025). The Open Bandit logs (Saito et al., 2021) span successive ad campaigns at the fashion retailer ZOZOTOWN.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Formally, at round t the learner picks an action $x_t \in \mathcal{A}_t \subset \mathbb{R}^d$ and observes only the scalar reward

$$y_t \;=\; x_t^\top \theta_t + \varepsilon_t, \qquad \theta_t = B_k^\star w_t.$$

Here $B_k^\star \in \mathbb{R}^{d\times r}$ is an orthonormal basis of the reward subspace. It is constant on segment $\mathcal{I}_k$ and changes at unknown change points. The latent state $w_t \in \mathbb{R}^r$ may drift within a segment (§2).

Related work. *Low-rank methods assume the structure is fixed.* Low-rank and bilinear bandits (Jun et al., 2019; Kang et al., 2022; Jedra et al., 2024; Jang et al., 2024; Duong et al., 2024; Lu et al., 2021) exploit the rank and attain $\widetilde{\mathcal{O}}(r\sqrt{T})$ once the subspace is known, where $\widetilde{\mathcal{O}}$ hides polylogarithmic factors. *Nonstationary methods ignore the structure.* Nonstationary linear bandits (Russac et al., 2019; Cheung et al., 2019) absorb drift through sliding windows (SW-LinUCB) or discounting (D-LinUCB). Like us, they bound dynamic regret, the regret against the best action of each round, but in the ambient $\mathbb{R}^d$. *Costly-observation methods assume a fixed parameter.* Costly-observation bandits (Seldin et al., 2014; Tucker et al., 2023; Elumar et al., 2024) price information about a fixed parameter, not a moving subspace.

Our learner is *single-play*. It plays one action per round and sees one scalar reward. To our knowledge, no existing method recovers a moving subspace from such rewards when the parameter may change every round. The obstruction is structural. With full feature–response vectors, streaming principal component analysis would track the subspace. The square of one reward is a single rank-one quadratic measurement of $\theta_t\theta_t^\top$. So the subspace cannot be learned passively. The learner must spend rounds on *probes*, queries that need not be feasible actions, and each probe gives up a round of reward. Sliding-window ridge in $\mathbb{R}^d$ never recovers the subspace. A one-shot estimate breaks at the next change point. Restarting low-rank methods per segment requires oracle boundaries.

Contributions.

- **When can a moving subspace be recovered from scalar rewards? (§2, Theorem 2.1).** We find the *identification boundary*, the conditions that decide whether the subspace can be learned from the rewards at all. When the learner draws its probes independently of the noise, the probes must span every direction (*coverage*), and isotropic probes suffice. The noise variance need not be known, and the state–noise coupling cancels out. Without coverage, two different subspaces give the same probe rewards, however many probes are taken.
- **How can a learner act on this? (§3, Algorithm 1).** Building on the identification boundary, we propose *Single-Play Subspace-Calibrated Optimism* (**SPSC**). It couples learning the representation with acting on it. Occasional priced *probe* rounds recover the current subspace from squared scalar rewards. The other rounds, exploitation rounds, then act optimistically inside the recovered r-dimensional representation. An adaptive variant resets the estimate when a change detector fires. SPSC is a bandit with a sensing channel. The channel, not the action set, is what makes the subspace identifiable at this rate, and the fee c prices every query. For a practitioner the message is simple: a small, budgeted sensing cost lets a deployed recommender track regime changes and cut regret.
- **What does tracking the subspace cost? (§4, Theorem 4.1).** When the mean reward vector is constant within segments, SPSC's costed dynamic regret, the regret plus the probe fees, is $\widetilde{\mathcal{O}}(r\sqrt{KT}+(rd)^{1/3}K^{1/3}T^{2/3})$. The $\sqrt{T}$ term scales with r. The ambient dimension enters only the additive cost of acquiring the subspace. Drift within segments adds the usual window tradeoff (Corollary 4.2).
- **What if the change points are unknown? (§4, Proposition 4.3, App. D.2).** The adaptive variant, SPSC-Adaptive, estimates projectors from two probe windows and resets when they disagree. When changes are large enough to detect, it keeps the same bound.
- **Does it help in practice? (§5, App. F).** On nine benchmarks, SPSC reduces regret when d is large relative to r. Figure 1 sketches the setting and the method. Table 1 and Appendix A position this work in detail.

Table 1: **Problem classes and SPSC.** LR low rank, NS nonstationarity, Scalar scalar reward feedback, ID identification conditions. ✓ yes, ✗ no, ∼ partial. T rounds, K segments, N arms, rank r, dimension d, MDP a Markov decision process.

Problem class (representative works)	LR	NS	Scalar	ID	Typical regret
Low-rank and bilinear bandits (Jun et al., 2019; Jang et al., 2021; Lu et al., 2021; Jedra et al., 2024; Jang et al., 2024; Kang et al., 2024; Wang et al., 2025a; Cinus et al., 2026; Lee et al., 2026)	✓	✗	✓	∼	$\widetilde{\mathcal{O}}(r\sqrt{T})$ given the subspace
Nonstationary multi-armed bandits (Garivier & Moulines, 2011; Cao et al., 2019; Auer et al., 2019; Besson et al., 2022; Abbasi-Yadkori et al., 2023; Komiyama et al., 2024; Guo et al., 2026)	✗	✓	✓	✗	$\widetilde{\mathcal{O}}(\sqrt{KNT})$
Nonstationary linear bandits (Russac et al., 2019; Cheung et al., 2019; Zhao et al., 2021; Hou et al., 2024; Wang et al., 2025b; Lee & Oh, 2026; Hu et al., 2026)	✗	✓	✓	✗	ambient d, $\widetilde{\mathcal{O}}(T^{3/4})$ for SW/D-LinUCB
Costly-information and probing bandits (Seldin et al., 2014; Tucker et al., 2023; Sekhari et al., 2023; Elumar et al., 2024; Boyer et al., 2025; Guan, 2026)	✗	✗	✓	✗	varies
Subspace recovery (Király & Tomioka, 2012; Kacham & Woodruff, 2023; Stojanovic et al., 2023; Zhang & Zhou, 2024)	✓	✗	∼	✓	recovery, not regret
Nonstationary and low-rank, closest (Qin et al., 2022; Cheng et al., 2023; Enikeeva et al., 2025)	✓	✓	∼	∼	MDP, detection, or tasks
SPSC (this paper)	✓	✓	✓	✓	$\widetilde{\mathcal{O}}(r\sqrt{KT} + (rd)^{1/3}K^{1/3}T^{2/3})$

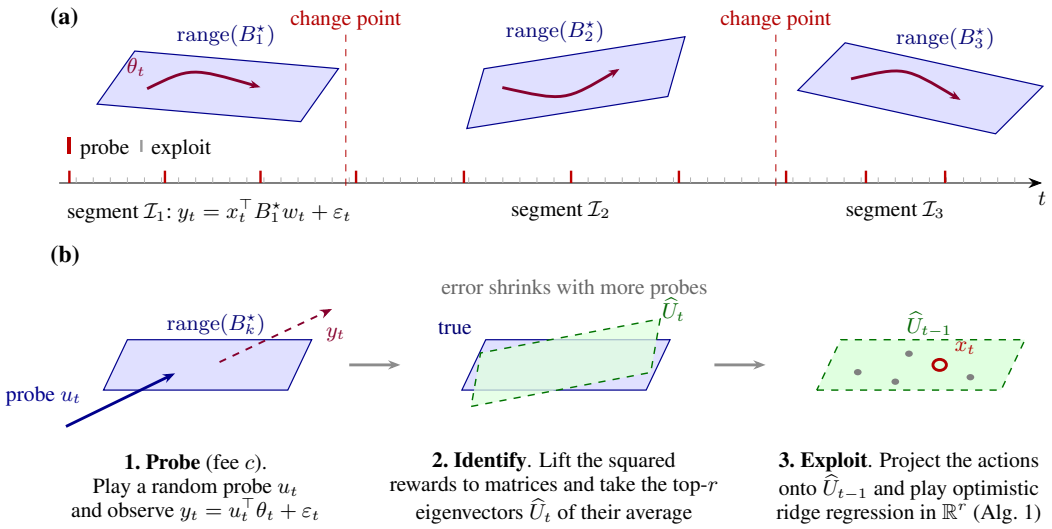


Figure 1: **Overview.** *(a)* The reward subspace $\mathrm{range}(B_k^\star)$ moves at unknown change points. The learner sees one scalar reward per round and inserts priced probe rounds (red ticks). *(b)* SPSC estimates a basis $\widehat{U}_t$ of the subspace from squared probe rewards and then plays optimistic windowed ridge regression inside it. SPSC-Adaptive resets the estimate when two probe windows disagree (§3).

2 Setting and Probe-Based Identification

A deployed system sees one scalar reward per round and must decide how much of its traffic to spend learning a structure that keeps moving. This section fixes the model, the probing protocol, and the regret, and then characterizes when the moving subspace can be recovered at all.

Piecewise low-rank model. At every round t the reward parameter admits a factorization $\theta_t = B_k^\star w_t$, where $B_k^\star \in \mathbb{R}^{d\times r}$ has orthonormal columns and is piecewise constant over $K \geq 1$ segments $\mathcal{I}_k = [\tau_{k-1}, \tau_k)$ of lengths $\ell_k := |\mathcal{I}_k|$, with unknown change points $1 = \tau_0 < \tau_1 < \cdots < \tau_K =$

$T+1$ over a horizon of T rounds, and the latent state $w_t \in \mathbb{R}^r$ follows a stable linear dynamical system around a mean path $\bar{a}_{k,t} \in \mathbb{R}^r$ inside each segment

$$w_t - \bar{a}_{k,t} = A_k(w_{t-1} - \bar{a}_{k,t-1}) + \eta_{t-1}, \qquad t \in \mathcal{I}_k,$$

with unknown A_k of spectral radius $\rho(A_k) \le 1-\alpha_0$ for some $\alpha_0 > 0$ (used only to ensure $\|w_t\| \le S_w$ via the stationary covariance bound) and zero-mean innovation η_{t-1}, independent of the past and of $\mathcal{A}_t$, with covariance $\Sigma_{\eta,k} \succeq 0$ ($\bar{a} \equiv 0$ gives a zero-mean system). We assume uniform action-boundedness $\sup_{x\in\mathcal{A}_t} \|x\| \le R_{\mathcal{A}}$, a known probe distribution Q with $\mathbb{E}_{u\sim Q}[uu^\top] \succeq \rho_Q I_d$ (full-dimensional *coverage*) and sub-Gaussian marginals (the theory uses scaled-sphere probes $u = \sqrt{d}\, g/\|g\|$, $g \sim \mathcal{N}(0, I_d)$, as our implementation does, App. C), and conditionally σ_ε-sub-Gaussian noise whose variance need not be known. Probes are drawn with the learner's own randomness. Conditional on the history $\mathcal{H}_{t-1}$ of the first $t-1$ rounds, u_t is independent of $(\theta_t, \varepsilon_t)$. We call this *probe independence*. Throughout, $\delta \in (0,1)$ denotes the global confidence parameter (concentration bounds hold with probability $\ge 1-\delta$) and $\lambda > 0$ denotes the ridge regularizer used by the windowed estimator of §3. The identification analysis uses the latent process only through the norm bound $\|w_t\| \le S_w$ and the second moment of θ_t on probe rounds, defined below. The regret analysis also uses the mean μ_t of the next paragraph and takes $A_k = 0$ (§4). The linear system is our running example.

Probes and costed regret. On probe rounds $t \in \mathcal{T}_{\text{probe}}$ the learner plays a *probe* $u_t \sim Q$, a sensing query drawn from a known exploration distribution to gather information rather than reward. This is a bandit with a sensing channel, not a standard contextual bandit. A probe need not be a feasible action, and the scaled-sphere probe has norm $\sqrt{d}$, ten times the action norm at $d = 100$ when $R_{\mathcal{A}} = 1$. Its norm and its freedom from $\mathcal{A}_t$ are what make the quadratic measurement below informative. We call this the *probing-permitted* protocol. It fits systems that can issue exploratory queries outside the action set, such as exploratory items shown to a user, survey-style questions, or simulator calls, and the fee c prices each query. Since every probe has the same norm, c can include a charge for it. A known design inside every $\mathcal{A}_t$ with the three properties of App. B.2 can take the place of Q. Performance is the costed dynamic regret

$$\mathrm{DynReg}_T^{(c)} := \sum_{t=1}^{T} \Big(\max_{x\in\mathcal{A}_t} x^\top \mu_t - x_t^\top \mu_t \Big) + c\,|\mathcal{T}_{\text{probe}}|, \tag{1}$$

with $x_t = u_t$ on probe rounds, so each probe is charged its forgone reward plus c. The comparator is the mean reward vector given the latent past, $\mu_t := \theta_t - B_k^\star \eta_{t-1} \in \mathrm{range}(B_k^\star)$ for $t \in \mathcal{I}_k$. The fresh innovation is revealed only after the action, so a comparator using the realized θ_t would charge every learner $\Omega(T)$ whenever it can flip the ranking of two actions. When θ_t is deterministic, as in the nonstationary bandits we compare with, $\mu_t = \theta_t$ and (1) is their dynamic regret.

Known and unknown quantities. The learner knows d, T, the rank r, the probe law Q, the probe cost c, the bounds $R_{\mathcal{A}}$ and S_w, a bound S_Δ on $\|\theta_t - \mu_t\|$ ($2S_w$ always works), and an upper bound on σ_ε. It does not need the noise variance (Theorem 2.1). It sees the action set $\mathcal{A}_t$ at each round and chooses λ, δ, the centering $\widehat{\sigma}^2$ (next paragraph), and the window W of §3. Algorithm 1 is also given the segment boundaries, and SPSC-Adaptive an upper bound on K, the common effective rank s, and a lower bound on $\lambda_{\min}$, both defined in the next paragraph (Proposition 4.3). The subspaces $B_k^\star$, the latent states w_t and their mean paths $\bar{a}_{k,t}$, the parameters θ_t and their means μ_t, the dynamics A_k and $\Sigma_{\eta,k}$, and the change points are unknown. Table 7 in Appendix B.1 lists the main symbols with their types.

Quadratic measurement identity. Let $\widehat{\sigma}^2 \ge 0$ be a fixed centering constant and define the probe statistic $\psi_t := y_t^2 - \widehat{\sigma}^2$ on every probe round $t \in \mathcal{T}_{\text{probe}}$. SPSC takes any fixed $\widehat{\sigma}^2$, and $\widehat{\sigma}^2 = 0$ needs no knowledge of the noise. With σ_t^2 the conditional noise variance and $\delta_\sigma := \widehat{\sigma}^2 - \sigma_t^2$,

$$\mathbb{E}[\psi_t \mid \mathcal{H}_{t-1}, u_t] = u_t^\top \widetilde{M}_t u_t + 2u_t^\top \mathbb{E}[\varepsilon_t \theta_t \mid \mathcal{H}_{t-1}, u_t] - \delta_\sigma, \tag{2}$$

where $\widetilde{M}_t := \mathbb{E}[\theta_t \theta_t^\top \mid \mathcal{H}_{t-1}]$ is the predictable second moment, whose range is contained in $\mathrm{range}(B_k^\star)$ for $t \in \mathcal{I}_k$. With exact centering and no state–noise coupling, $\mathbb{E}[\varepsilon_t \theta_t \mid \cdot] = 0$, this collapses to $\mathbb{E}[\psi_t \mid \cdot] = u_t^\top \widetilde{M}_t u_t$. *A scalar reward carries a quadratic measurement of* $\widetilde{M}_t$. The average of $\widetilde{M}_t$ over the probe rounds of segment k, $\bar{M}_k^{\text{probe}}$, has range $S_k \subseteq \mathrm{range}(B_k^\star)$, of dimension

$s_k \le r$, its *effective rank*. We call S_k the *effective subspace* of the segment and let $\lambda_{\min}$ lower-bound its smallest nonzero eigenvalue. Under full excitation, $\Sigma_{\eta,k} \succ 0$, $S_k = \mathrm{range}(B_k^\star)$ (Proposition B.9). A fixed $\theta^{(k)}$ per segment gives $S_k = \mathrm{span}(\theta^{(k)})$, only one direction of $\mathrm{range}(B_k^\star)$.

Lifted subspace estimator. The probe moment operator $\mathcal{K} : M \mapsto \mathbb{E}[(u^\top M u)\, uu^\top]$ admits a closed-form inverse on symmetric matrices for scaled-sphere probes $u = \sqrt{d}\, v$, $v \sim \mathrm{Unif}(\mathbb{S}^{d-1})$,

$$\mathcal{K}(M) = \tfrac{d}{d+2}\big(\mathrm{tr}(M)\, I_d + 2M\big), \qquad \mathcal{K}^{-1}(N) = \tfrac{d+2}{2d}\, N - \tfrac{\mathrm{tr}(N)}{2d}\, I_d, \qquad \|\mathcal{K}^{-1}\|_{\mathrm{op}\to\mathrm{op}} \le 1 \quad (3)$$

on the symmetric subspace (Lemma B.4). Define the *lifted probe sample*

$$G_t := \mathcal{K}^{-1}(\psi_t\, u_t u_t^\top), \qquad \mathbb{E}[G_t \mid \mathcal{H}_{t-1}] = \widetilde{M}_t + \widetilde{B}_t, \quad (4)$$

with bias $\widetilde{B}_t = -(\delta_\sigma/d)\, I_d$ exactly on scaled-sphere probes, hence $\|\widetilde{B}_t\|_{\mathrm{op}} = |\delta_\sigma|/d$ (Lemma B.6). Averaging lifted samples across the probe rounds of segment k yields $\widehat{M}_k := m_k^{-1} \sum_{t\in\mathcal{T}_k} G_t$ (with $\mathcal{T}_k := \mathcal{T}_{\mathrm{probe}} \cap \mathcal{I}_k$ the probe rounds in segment k, $m_k := |\mathcal{T}_k|$), whose top-r eigenvectors form $\widehat{U}_k \in \mathbb{R}^{d\times r}$. With $\widehat{P}_k := \widehat{U}_k \widehat{U}_k^\top$ and $P_k^\star$ the projector onto S_k, the part of S_k that the estimate misses satisfies $\|(I - \widehat{P}_k)P_k^\star\|_{\mathrm{op}} \le 2C_{\mathrm{sub}}\sqrt{d\log(4d/\delta)/m_k}$, with C_{sub} a constant free of d up to logarithms, once $m_k \ge d\log(4d/\delta)$ (Corollary B.10). Since $\widetilde{B}_t$ is a scaled identity, a wrong centering shifts eigenvalues but does not rotate eigenvectors.

The identification boundary. Can the moving subspace be learned from single-play scalar rewards at all? By the *identification boundary* we mean the conditions that answer this question. On one side of the boundary, the rewards carry the information, and averaging more lifted samples gives a better estimate, at rate $\sqrt{d/m_k}$. On the other side, two different subspaces produce exactly the same probe observations, and no amount of probing tells them apart. Theorem 2.1 locates the boundary. When the learner draws its probes independently of the noise, coverage is necessary, and isotropic probes suffice. If the probes miss a direction, no number of probes recovers the subspace. The noise variance need not be known, and the state–noise coupling cancels out by the symmetry of the probes. Both become obstructions only when the noise can depend on the probe direction. The boundary is a property of the sensing channel, not of SPSC. Each obstruction comes with two instances whose responses have the same law, so no algorithm can tell them apart, and sufficiency is constructive, yielding the estimator SPSC uses.

Theorem 2.1 (Identification under priced probing)**.** *Consider the model of §2 with scaled-sphere probes $u_t = \sqrt{d}\, v_t$, $v_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and the lifted probe sample $G_t = \mathcal{K}^{-1}(\psi_t u_t u_t^\top)$ from* (4)*.*

1. ***Sufficiency.*** *Under probe independence and for any fixed centering $\widehat{\sigma}^2$, the lifted sample satisfies $\mathbb{E}[G_t \mid \mathcal{H}_{t-1}] = \widetilde{M}_t - (\delta_\sigma/d)I_d$, whatever the size of the state–noise coupling. The top-s_k eigenspace of the segment average $\widehat{M}_k = m_k^{-1}\sum_{t\in\mathcal{T}_k} G_t$ recovers the effective subspace S_k. For every m_k, with probability at least $1-\delta$, its projector is within $\min\{1, C_{\mathrm{sub}}(\sqrt{d\log(4d/\delta)/m_k} + d\log(4d/\delta)/m_k)\}$ of $P_k^\star$. No knowledge of the noise variance is needed, since $\widehat{\sigma}^2 = 0$ is allowed, and any fixed bounded centering changes only C_{sub} (proof in App. B.2).*
2. ***Necessity.*** *Coverage is necessary for every probe law. If the probes lie in a proper subspace, two distinct second moments $\widetilde{M}_t \ne \widetilde{M}_t'$ give identical probe observations. If the noise variance or the state–noise coupling may vary with the probe direction, identification can fail even with full coverage. Knowing the variance removes the first obstruction, and a bound $\epsilon_\times$ on the coupling limits the second (proof in App. B.3).*

Building on this identification result, we now propose an algorithm that learns the subspace and earns reward at the same time.

3 Algorithm

Design challenges. Four obstacles shape the design. *The subspace is not observed.* Each round returns one scalar $y_t = x_t^\top \theta_t + \varepsilon_t$, so recovering $\mathrm{range}(B_k^\star)$ requires the quadratic lifting of §2.

Information is costly. Probe rounds give up reward and cost c each. Balancing m_k probes against the resulting $\sqrt{d/m_k}$ projector error gives $m_k \asymp \ell_k^{2/3}$. *The representation changes.* An estimate from an earlier segment is stale after a change, calling for per-segment identification and online resets. *Exploitation must track residual drift.* Within a fixed subspace θ_t may still vary, calling for windowed projected ridge regression with upper confidence bounds (ridge-UCB).

SPSC (Algorithm 1) takes the segment boundaries as input (SPSC-Adaptive below removes them). Probe rounds draw $u_t \sim Q$, accumulate the lifted samples G_t, and after every probe set $\widehat{U}$ to the top-r eigenvectors of their average ($\widehat{U}_t$ is the basis after round t and $\widehat{U}_k$ the last one of segment k). Exploitation rounds project each action to $z_t(x) = \widehat{U}_{t-1}^\top x \in \mathbb{R}^r$ and run ridge-UCB over the window $\mathcal{W}_t$ of the segment's exploitation rounds among the last W rounds, all projected with the current basis.

Algorithm 1 SPSC (Single-Play Subspace-Calibrated Optimism) with known segment boundaries. Probes need not lie in $\mathcal{A}_t$.

Input: segments $\{\mathcal{I}_k\}_{k=1}^K$, probe rounds $\mathcal{T}_{\text{probe}}$, probe law Q, rank r, window W, ridge λ, centering $\widehat{\sigma}^2$, and any orthonormal $\widehat{U} \in \mathbb{R}^{d\times r}$
for $k = 1, \dots, K$ **do**
  $M \leftarrow 0,\ m \leftarrow 0$ ▷ reset at each segment
  for $t \in \mathcal{I}_k$ **do**
    if $t \in \mathcal{T}_{\text{probe}}$ **then** ▷ probe round
      draw $u_t \sim Q$, play it, and observe y_t
      $G_t \leftarrow \mathcal{K}^{-1}\big((y_t^2 - \widehat{\sigma}^2)\, u_t u_t^\top\big)$ ▷ lifted probe sample
      $M \leftarrow M + G_t,\ m \leftarrow m + 1$
      $\widehat{U} \leftarrow \mathrm{TopEig}_r(M/m)$ ▷ refresh the basis
    else ▷ exploitation round
      $\mathcal{W}_t \leftarrow \{i \in \mathcal{I}_k \setminus \mathcal{T}_{\text{probe}} :\ t - W \le i < t\},\ z_t(x) \leftarrow \widehat{U}^\top x$
      $\widetilde{V}_t \leftarrow \lambda I_r + \sum_{i\in\mathcal{W}_t} z_t(x_i) z_t(x_i)^\top,\ \widetilde{b}_t \leftarrow \sum_{i\in\mathcal{W}_t} z_t(x_i)\, y_i$
      $\widehat{a}_t \leftarrow \widetilde{V}_t^{-1}\widetilde{b}_t$
      play $x_t = \arg\max_{x\in\mathcal{A}_t} \big\{ z_t(x)^\top \widehat{a}_t + \beta_t^{(r,W)} \|z_t(x)\|_{\widetilde{V}_t^{-1}} \big\}$ and observe y_t

The radius $\beta_t^{(r,W)} = \sigma_{\text{eff}} \sqrt{r \log(1 + WR_{\mathcal{A}}^2/(\lambda r)) + 2\log(2KT^2/\delta)} + \sqrt{\lambda} S_w$ is the usual ridge radius in r dimensions. It scales with $\sqrt{r}$, not $\sqrt{d}$. The innovation enters it through the effective noise scale $\sigma_{\text{eff}} := \sigma_\varepsilon + R_{\mathcal{A}} S_\Delta$ of $y_t - x_t^\top \mu_t$, which is σ_ε when θ_t is fixed within segments. The part of μ_t outside the estimated subspace is not covered by the radius. The analysis charges it to the regret directly, at the rate of the projector error, under condition (WC) of §4. An exploitation round costs $O(dr|\mathcal{A}_t| + r^2)$ vs. $O(d^2|\mathcal{A}_t|)$ for ambient LinUCB (Li et al., 2010), an $O(d/r)$ speed-up, and a probe round $O(d^2 r + Wdr)$. Memory is $O(d^2 + Wd)$ (App. D.1).

Adaptive variant (unknown boundaries). SPSC-Adaptive detects changes online. After each reset it plays a block of m_{relearn} probes and then probes at a fixed rate p. It keeps two non-overlapping windows of W_{det} recent and past probe rounds, forms the top-s projectors $\widehat{P}_t^{\text{rec}}, \widehat{P}_t^{\text{past}}$ of their lifted-moment estimates, and resets the subspace estimate when $\Gamma_t := \|\widehat{P}_t^{\text{rec}} - \widehat{P}_t^{\text{past}}\|_{\text{op}}$ exceeds $2\eta_P$. Here s is the effective rank and η_P the projector error of one window (Cor. B.10). With probability $1 - \delta_{\text{FA}}$, a chosen false-alarm level, there is no false alarm, and every change with $\|P_k^\star - P_{k-1}^\star\|_{\text{op}} > 4\eta_P$ is detected once the windows lie on opposite sides of it (App. B.6, pseudocode in App. D.2). App. D.2 also gives a reward-drop trigger, which resets when the mean reward of a recent window falls a threshold below the earlier mean (Proposition D.1).

4 Regret guarantees

Assumptions. Theorem 4.1 uses the model of §2 with $\|x\| \le R_{\mathcal{A}}$, $\|w_t\| \le S_w$, the noise and innovation of each round independent of earlier rounds and of $\mathcal{A}_t$, scaled-sphere probes, and probe independence, in the probing-permitted protocol. Within each segment the mean reward vector is constant. In the model this means $A_k = 0$ and a constant mean path, so $w_t = \bar{a}_k + \eta_{t-1}$ for $t \in \mathcal{I}_k$,

with innovations of covariance $\Sigma_{\eta,k} \succeq 0$ that satisfy the bounds above. A fixed $\theta^{(k)}$ per segment is the case $\Sigma_{\eta,k} = 0$. The predictable second moment is then the same in every round of a segment. Its range is the effective subspace S_k, of dimension at most r, and its smallest nonzero eigenvalue is at least $\lambda_{\min}$. The probe schedule is balanced. For an absolute constant a_0, every segment k and every $j \le m_k$, at most $a_0\lceil \ell_k/m_k \rceil$ exploitation rounds follow exactly j probes, which uniform spacing satisfies. The projected windows are well conditioned (WC). For constants $\rho_z > 0$ and n_0, (WC) asks that the smallest eigenvalue of $\sum_{i\in\mathcal{W}_t} z_t(x_i)z_t(x_i)^\top$ be at least $\rho_z|\mathcal{W}_t|$ at every exploitation round with $|\mathcal{W}_t| \ge n_0$, where $z_t(x) = \widehat{U}_{t-1}^\top x$ is the projection onto the basis in use. The trace bound gives $\rho_z \le R_{\mathcal{A}}^2/r$, and $\kappa_z := \rho_z r/R_{\mathcal{A}}^2 \le 1$ measures how far the windows are from isotropic. (WC) is a condition on the played actions, and App. G.1 finds that it holds at every round in three benchmarks. Algorithm 1 knows the segment boundaries.

Theorem 4.1 (Costed dynamic regret of Alg. 1)**.** *Under these assumptions, run Algorithm 1 with a window $W \ge \max_k \ell_k$ and $m_k = \min\{\ell_k, \lceil c_0 \ell_k^{2/3}\rceil\}$ probes in segment k, with c_0 given in App. B.5. There is a threshold T_0, polynomial in d, K and $1/\lambda_{\min}$, such that for $T \ge T_0$, with probability at least $1-\delta$,*

$$\mathrm{DynReg}_T^{(c)} \;\le\; \widetilde{\mathcal{O}}\big(r\sqrt{KT} + (rd)^{1/3}K^{1/3}T^{2/3}\big). \tag{5}$$

Here $\widetilde{\mathcal{O}}$ hides logarithmic factors and constants that do not depend on d for fixed κ_z and $\lambda_{\min}$. The constants scale as $\kappa_z^{-1/2}$ in the first term and as $(\lambda_{\min}^2\kappa_z)^{-1/3}$ in the second. They and T_0 are given in App. B.5.

The full bound, Theorem B.19 in App. B.5, holds for any window and for a drifting mean path. It adds a window term $\widetilde{\mathcal{O}}(rT/\sqrt{W})$, a drift term linear in W and in the variation of μ_t, and burn-in terms that $T \ge T_0$ makes lower order.

Reading the bound. The first term is the cost of learning inside the r-dimensional subspace. The second is the cost of learning the subspace, and it is the only place where d enters, as $d^{1/3}$. With $r = d$ and no probes SPSC is restarted LinUCB, with an oracle subspace it is projected OFUL, and the bound recovers both rates (Table 6). Against restarted LinUCB, which pays $\widetilde{\mathcal{O}}(d\sqrt{KT})$, the bound is smaller when $(d-r)\sqrt{KT}$ exceeds $(rd)^{1/3}K^{1/3}T^{2/3}$ up to constants, that is, for horizons $T \lesssim K(d-r)^6/(rd)^2$, or equivalently when the dimension gap satisfies $d - r \gtrsim (rd)^{1/3}(T/K)^{1/6}$. When θ is fixed within each segment, $S_k = \mathrm{span}(\theta^{(k)})$, $\lambda_{\min} = \min_k \|\theta^{(k)}\|^2$, and the theorem holds for any rank $r \ge 1$ given to Algorithm 1. Most of our real-data benchmarks are of this kind. Evenly spaced probes at a constant rate $p^\star = c_0(K/T)^{1/3}$ give the same bound (App. B.5).

Corollary 4.2 (Drifting means)**.** *Keep the other assumptions, including $A_k = 0$, but let $\Sigma_{\eta,k} \succ 0$ and let the mean path drift within segments, with variation $V_{\mathrm{in}} := \sum_k \sum_{t,t+1\in\mathcal{I}_k} \|\mu_{t+1} - \mu_t\|_2$. Algorithm 1 with window W adds $\widetilde{\mathcal{O}}(rT/\sqrt{W}) + O(R_{\mathcal{A}}^2 W V_{\mathrm{in}}/\sqrt{\rho_z})$ to (5). The window $W^\star = \min\{\max_k \ell_k, (rT\sqrt{\rho_z}/(R_{\mathcal{A}}^2 V_{\mathrm{in}}))^{2/3}\}$ makes the addition $\widetilde{\mathcal{O}}(r^{2/3}V_{\mathrm{in}}^{1/3}T^{2/3})$.*

Latent dynamics, $A_k \ne 0$, make V_{in} grow linearly in T, at rate $O(\|A_k\|)$.

Choice of W. Within segment k, with n_k exploitation rounds and variation V_k^{in} of μ_t, exploitation trades an estimation cost of order $r\,n_k/\sqrt{W}$ against a drift cost of order WV_k^{in}. The balance is $W_k^\star \asymp (r\,n_k/V_k^{\mathrm{in}})^{2/3}$, the classical variation-budget window with d replaced by r. When $V_k^{\mathrm{in}} = 0$ the whole segment is used. In practice one can minimize the bound over a dyadic grid of windows, or let a bandit-over-bandit layer (Cheung et al., 2019) choose among the grid, since W affects neither the probe schedule nor the identification analysis.

Proposition 4.3 (Adaptive SPSC)**.** *Suppose the assumptions of Theorem 4.1 hold with a common known effective rank $s \le r$, that an upper bound on K is known, and that every change satisfies $\|P_k^\star - P_{k-1}^\star\|_{\mathrm{op}} > 4\eta_P$. Run SPSC-Adaptive (Alg. 2) at probe rate $p^\star$, with detector windows of W_{det} probes, on segments longer than $m_{\mathrm{relearn}} + 3W_{\mathrm{det}}/p^\star$ rounds. There is a threshold $T_0' \ge T_0$ such that for $T \ge T_0'$, with probability at least $1 - \delta - \delta_{\mathrm{FA}}$, SPSC-Adaptive attains (5) with larger constants. App. B.6 gives the full bound, with its delay, recovery and burn-in terms, and the inequalities that define T_0'. Proposition D.1 (App. D.2) gives the same bound for the reward-drop trigger with its threshold and window set from Δ, when every change lowers the expected reward of the stale policy by Δ and no segment lowers it by more than $\Delta/2$.*

5 EXPERIMENTS

We ask three questions. Where does SPSC beat ambient methods once d is large relative to r (§5.1)? Does the advantage survive real data whose low-rank structure is only approximate (§5.2)? Does it beat stationary low-rank methods and the nonstationary baselines (§5.3)? We compare with eleven baselines (App. E), with Oracle-LinUCB (true subspace) as a reference. Given the same probes, LinUCB and SW-LinUCB still lose to an SPSC variant in 82 of 98 (d, r) cells (App. G.5). We report the costed regret (1) with the realized θ_t, which dominates it in expectation. SPSC-Adaptive resets on the reward-drop trigger of App. D.2. Appendix F gives the per-cell results and Appendix G the checks of the theory (App. G.2–G.4).

5.1 WHERE DOES SPSC BEAT AMBIENT LINUCB?

Theorem 4.1 trades a rank-r $\sqrt{T}$ term for an additive $(rd)^{1/3}K^{1/3}T^{2/3}$ cost of acquiring the subspace. So SPSC should lose to ambient LinUCB when r is close to d and win once $d - r$ is large. We locate this crossover on a synthetic grid of 40 (d, r) cells. In each segment the reward parameter is a fixed vector drawn from a new random r-dimensional subspace, and SPSC runs at rank r.

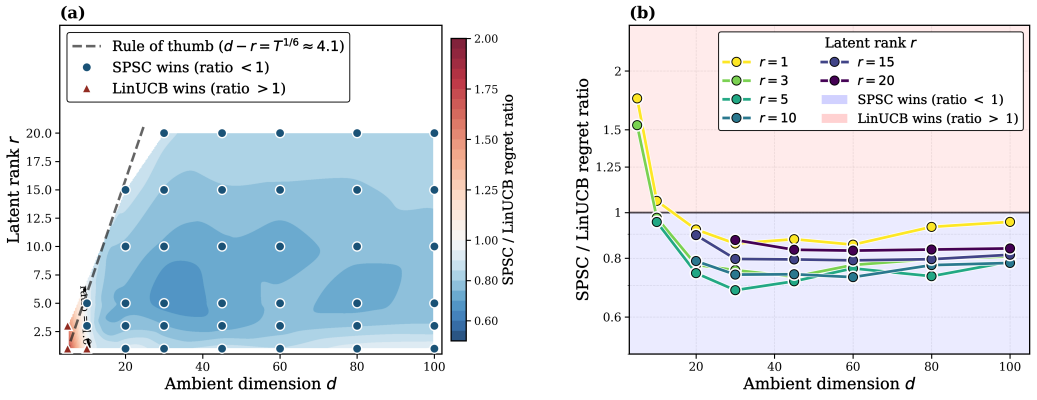


Figure 2: **Empirical crossover.** The regret ratio is SPSC's costed regret over LinUCB's. **(a)** Ratio on the (d, r) grid, with the rule of thumb $d - r = T^{1/6}$ (dashed) and the empirical ratio=1 contour (solid). Markers show the winner of each cell. **(b)** Ratio versus d for each r, log scale.

SPSC beats LinUCB on 37 of the 40 cells, losing only at the smallest dimensions ($d \leq 10$). The crossover $d - r \gtrsim (rd)^{1/3}(T/K)^{1/6}$ of §4 carries the instance factor $(rd)^{1/3}$ and hidden constants, so we test a parameter-free proxy that keeps its $T^{1/6}$ dependence. SPSC should win when $d - r > T^{1/6} \approx 4$. The rule is theory-motivated, not a corollary of Theorem 4.1. It classifies 39 of the 40 cells correctly, against 32 for the crossover with its instance factor, which over-penalizes small d. Its one exception, $(d, r)=(10, 1)$, is a 5.5% loss (Figure 2a). Beyond $d \approx 30$ the regret ratio levels off at 0.7–0.95, so the gain is a steady 16–29% for $d \geq 45$ and $r \geq 3$ (Figure 2b).

5.2 DOES THE ADVANTAGE HOLD ON REAL DATA?

Real data could erase the synthetic advantage. Reward parameters are ridge fits on real features, segments come from class or user shifts, and MovieLens rewards are real ratings. We recast six datasets (Covertype, Pendigits, Satimage, MNIST, Fashion-MNIST, and MovieLens-100K with real ratings) and the Open Bandit production logs as piecewise low-rank bandits.

Yes. At d=105 an SPSC variant beats every non-oracle baseline on all six datasets, with the smallest gains on Covertype. SPSC's regret is 61% below LinUCB's on Satimage and 10% below on MovieLens, whose rewards are real ratings (Table 2), and with real-rating probes it still leads there (App. F.1). At the smallest d=55 its edge on Covertype is 3%. SPSC-Adaptive beats LinUCB by 45% on Pendigits and 67% on Satimage at $(d, r)=(105, 10)$. On the Open Bandit logs an SPSC variant beats every non-oracle baseline on three of the four cells, SPSC by 13% relative to LinUCB at $(d, r)=(55, 5)$ and SPSC-Adaptive by 7% at $(105, 5)$.

Table 2: **Real data.** Costed regret of SPSC over LinUCB (S/Lin) and over the best non-oracle baseline (S/Best) at r=10. The last columns give the best cell. Values <1 favor SPSC.

	d=55		d=105		d=200		best	
Dataset	S/Lin	S/Best	S/Lin	S/Best	S/Lin	S/Best	cell	S/Lin
Covertype	0.97	1.00	0.88	0.93	—	—	$(105, 30)$	0.86
Pendigits	0.77	0.79	0.66	0.70	—	—	$(105, 20)$	0.66
Satimage	0.49	0.61	0.39	0.59	—	—	$(105, 10)$	0.39
MNIST	0.81	0.83	0.74	0.86	0.71	0.96	$(200, 20)$	0.70
Fashion-MNIST	0.87	0.89	0.70	0.83	0.66	0.95	$(200, 20)$	0.65
MovieLens	0.94	0.94	0.90	0.91	0.86	0.91	$(200, 5)$	0.83

5.3 How does SPSC compare with specialized baselines?

Specialized baselines are built for one of the two structures that SPSC combines. The stationary low-rank methods BOSS (Duong et al., 2024) and Jedra (Jedra et al., 2024) learn one fixed subspace, so comparing SPSC-Adaptive with them shows what tracking the subspace adds. The nonstationary methods track change in the ambient space. We compare SPSC with them on a small-d stress test that follows the dimensions of Russac et al. (2019).

Table 3: **Stationary low-rank methods.** Costed regret (mean±SE, 10 seeds), with no method told the change points. The last column compares SPSC-Adaptive with the better of BOSS and Jedra.

d	r	SPSC-Adaptive	BOSS	Jedra	vs. best
55	5	$\mathbf{342 \pm 14}$	399 ± 16	386 ± 17	-11.6%
55	10	$\mathbf{507 \pm 13}$	588 ± 26	602 ± 23	-13.8%
105	5	$\mathbf{274 \pm 10}$	300 ± 14	299 ± 12	-8.3%
105	10	$\mathbf{401 \pm 13}$	423 ± 17	429 ± 13	-5.1%
105	20	$\mathbf{608 \pm 13}$	654 ± 16	654 ± 13	-7.0%
200	5	$\mathbf{218 \pm 5}$	236 ± 8	226 ± 8	-3.4%
200	10	$\mathbf{292 \pm 12}$	294 ± 11	305 ± 14	-0.9%
200	20	$\mathbf{452 \pm 10}$	473 ± 15	489 ± 17	-4.4%

SPSC-Adaptive beats both stationary low-rank methods on every cell. Its regret is 0.9–13.8% below the better of the two, and the paired difference exceeds two standard errors on five of the eight cells (Table 3, App. F.2). On the stress test, SPSC cuts costed regret by 46.0% relative to D-LinUCB and 34.2% relative to SW-LinUCB, both of which reset at the true change points (App. F.3).

6 Conclusion

This paper characterizes when a drifting low-rank subspace can be learned from scalar rewards (Theorem 2.1). We build SPSC on this characterization and prove a costed dynamic regret of $\widetilde{\mathcal{O}}(r\sqrt{KT}) + \widetilde{\mathcal{O}}((rd)^{1/3}K^{1/3}T^{2/3}) + \widetilde{\mathcal{O}}(rT/\sqrt{W} + WV_{\mathrm{in}})$ (Corollary 4.2). Its $\sqrt{T}$ term depends on the rank r, and d enters only the additive acquisition cost. In extensive evaluations on recommendation, production-log, and benchmark streams, SPSC helps when the ambient dimension d is large relative to the intrinsic rank r.

Limitations and future work. SPSC relies on a sensing channel whose probes may lie outside the action sets. The regret analysis takes $A_k = 0$, and under latent dynamics the variation V_{in} grows linearly in T. The probe–exploit design opens several directions. Designs inside the action sets could replace the probes where no sensing channel exists, and rank-adaptive estimates could follow ranks that change between segments. Bilinear models, generalized linear rewards, and low-rank reinforcement learning also see one scalar per round, so priced probing may extend to them.

AI USE STATEMENT

We used generative AI tools to help draft and edit the manuscript, search for related literature and check references, and assist with debugging, refactoring, and editing software code. We did not use generative AI to develop the theory or conceptual framework, formulate mathematical claims or proofs, propose or refine hypotheses, design the experiments, implement the proposed methods, or interpret the results. All AI-assisted code was reviewed and tested by the authors, and all reported results were regenerated from the released code.

REPRODUCIBILITY STATEMENT

Code for all experiments is available anonymously at `https://anonymous.4open.science/r/spsc-supplement-code`. Complete proofs of every theoretical claim are in App. B, and the probe distributions are specified in App. C. Pseudocode for SPSC is in Section 3 and for SPSC-Adaptive in App. D. Datasets, preprocessing, hyperparameters, and seeds are in App. E, and results for every benchmark are in App. F.

REFERENCES

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

Yasin Abbasi-Yadkori, András György, and Nevena Lazić. A new look at dynamic regret for non-stationary stochastic bandits. *Journal of Machine Learning Research*, 24(288):1–37, 2023.

Peter Auer, Yifang Chen, Pratik Gajane, Chung-Wei Lee, Haipeng Luo, Ronald Ortner, and Chen-Yu Wei. Achieving optimal dynamic regret for non-stationary bandits without prior information. In *Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 159–163. PMLR, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 27, 2014.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(77):1–40, 2022.

Rajendra Bhatia. *Matrix analysis*. Springer Science & Business Media, 2013.

Nathan Boyer, Dorian Baudry, and Patrick Rebeschini. Best-of-both worlds for linear contextual bandits with paid observations. *arXiv preprint arXiv:2510.07424*, 2025.

T. Tony Cai and Anru Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 418–427, 2019.

Yuxin Chen, Yuejie Chi, and Andrea J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7): 4034–4059, 2015.

Yuan Cheng, Jing Yang, and Yingbin Liang. Provably efficient algorithm for nonstationary low-rank MDPs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Federico Cinus, Yuko Kuroki, Atsushi Miyauchi, and Francesco Bonchi. Online minimization of polarization and disagreement via low-rank matrix bandits. In *The Fourteenth International Conference on Learning Representations*, 2026.

Gautam Dasarathy, Parikshit Shah, Badri Narayan Bhaskar, and Robert D. Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388, 2015.

Yihan Du, Longbo Huang, and Wen Sun. Multi-task representation learning for pure exploration in linear bandits. In *International Conference on Machine Learning (ICML)*, 2023.

Dheeru Dua and Casey Graff. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2019.

Thang Duong, Zhi Wang, and Chicheng Zhang. Beyond task diversity: provable representation transfer for sequential multitask linear bandits. *Advances in Neural Information Processing Systems*, 37:37791–37822, 2024.

Eray Can Elumar, Cem Tekin, and Osman Yağan. Multi-armed bandits with costly probes. *IEEE Transactions on Information Theory*, 71(1):618–643, 2024.

Farida Enikeeva, Olga Klopp, and Mathilde Rousselot. Change-point detection in low-rank VAR processes. *Bernoulli*, 31(2):1058–1083, 2025.

Diego Feijer, Himan Abdollahpouri, Sanket Gupta, Alexander Clare, Yuxiao Wen, Todd Wasson, Maria Dimakopoulou, Zahra Nazari, Kyle Kretschman, and Mounia Lalmas. Calibrated recommendations with contextual bandits. *arXiv preprint arXiv:2509.05460*, 2025.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pp. 174–188. Springer, 2011.

Yongjie Guan. Bandits with probing: Optimal regret and the limits of winner feedback. *arXiv preprint arXiv:2609.15248*, 2026.

Mengzhuo Guo, Wuqi Zhang, Yiwen Shen, and Qingpeng Zhang. Bid shading in first-price auction: Nonstationary bayesian multiarmed bandit methods for real-time bidding. *Information Systems Research*, 2026. doi: 10.1287/isre.2025.1837. Articles in Advance.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015.

Yunlong Hou, Vincent Y Tan, and Zixin Zhong. Almost minimax optimal best arm identification in piecewise stationary linear bandits. *Advances in Neural Information Processing Systems*, 37: 128967–129041, 2024.

Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning (ICML)*, 2021.

Zihao Hu, Yuan Yao, Jiheng Zhang, and Zhengyuan Zhou. Dynamic regret for non-stationary linear bandits via misspecification reductions. *arXiv preprint arXiv:2607.02891*, 2026.

Kyoungseok Jang, Kwang-Sung Jun, Se-Young Yun, and Wanmo Kang. Improved regret bounds of bilinear bandits using action space analysis. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4744–4754, 2021.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Efficient low-rank matrix estimation, experimental design, and arm-set-dependent low-rank bandits. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 21329–21372, 2024.

Yassir Jedra, William Réveillard, Stefan Stojanovic, and Alexandre Proutiere. Low-rank bandits via tight two-to-infinity singular subspace recovery. In *International Conference on Machine Learning*, pp. 21430–21485. PMLR, 2024.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pp. 3163–3172. PMLR, 2019.

Praneeth Kacham and David P. Woodruff. Lower bounds on adaptive sensing for matrix recovery. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient frameworks for generalized low-rank matrix bandit problems. *Advances in Neural Information Processing Systems*, 35:19971–19983, 2022.

Yue Kang, Cho-Jui Hsieh, and Thomas C.M. Lee. Low-rank matrix bandits with heavy-tailed rewards. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1863–1889, 2024.

Franz J. Király and Ryota Tomioka. A combinatorial algebraic approach for the identifiability of low-rank matrix completion. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 967–974, 2012.

Junpei Komiyama, Edouard Fouché, and Junya Honda. Finite-time analysis of globally nonstationary multi-armed bandits. *Journal of Machine Learning Research*, 25(112):1–56, 2024.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Joongkyu Lee and Min-hwan Oh. Nonstationary generalized linear bandits with discounted online mirror descent. *arXiv preprint arXiv:2605.25590*, 2026.

Junghyun Lee, Kyoungseok Jang, Kwang-Sung Jun, Milan Vojnovic, and Se-Young Yun. GL-LowPopArt: A nearly instance-wise minimax-optimal estimator for generalized low-rank trace regression. In *International Conference on Artificial Intelligence and Statistics*, 2026.

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

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *International conference on artificial intelligence and statistics*, pp. 460–468. PMLR, 2021.

Jonas Mueller, Vasilis Syrgkanis, and Matt Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuzhen Qin, Tommaso Menara, Samet Oymak, ShiNung Ching, and Fabio Pasqualetti. Non-stationary representation learning in sequential linear bandits. *IEEE Open Journal of Control Systems*, 1:41–56, 2022. doi: 10.1109/OJCSYS.2022.3178540.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. *Advances in Neural Information Processing Systems*, 36:11261–11295, 2023.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pp. 280–287. PMLR, 2014.

Stefan Stojanovic, Yassir Jedra, and Alexandre Proutiere. Spectral entry-wise matrix estimation for low-rank reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 77056–77070, 2023.

Joel A Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Aaron D Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Uncertainty in Artificial Intelligence*, pp. 2147–2156. PMLR, 2023.

Yao Wang, Jiannan Li, Yue Kang, Shanxing Gao, and Zhenxin Xiao. Generalized low-rank matrix contextual bandits with graph information. *arXiv preprint arXiv:2507.17528*, 2025a.

Zhiyong Wang, Jize Xie, Yi Chen, John C. S. Lui, and Dongruo Zhou. Variance-dependent regret bounds for nonstationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2503–2511. PMLR, 2025b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jiaqi Yang, Wei Hu, Jason D. Lee, and Simon S. Du. Impact of representation learning in linear bandits. In *International Conference on Learning Representations (ICLR)*, 2021.

Anderson Y. Zhang and Harrison H. Zhou. Leave-one-out singular subspace perturbation analysis for spectral clustering. *The Annals of Statistics*, 52(5), 2024.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. *arXiv preprint arXiv:2103.05324*, 2021.

CONTENTS OF THE APPENDIX

A Related work in detail **14**

A.1 Low-rank and bilinear bandits 15

A.2 Nonstationary bandits 15

A.3 Nonstationary and low-rank 15

A.4 Costly information and probing 15

A.5 Subspace recovery from quadratic measurements 15

A.6 Representation learning in bandits 16

A.7 Terms and special cases of Theorem 4.1 16

B Proofs **16**

B.1 Preliminaries and notation 17

B.2 Concentration of the lifted probe moment 19

B.3 Necessity of the probe-side conditions 22

B.4 Lemmas for the interleaved analysis 24

B.5 Proof of Theorem 4.1 and Corollary 4.2 26

B.6 Proof of Proposition 4.3 (SPSC-Adaptive) 27

C Probe distributions: scaled sphere and Gaussian truncation **29**

D Algorithms: pseudocode and adaptive variant **29**

D.1 Computational cost 29

D.2 SPSC-Adaptive (unknown boundaries) 29

E Experimental setup **31**

F Additional results on the benchmarks **33**

F.1 Per-cell results for every dataset 33

F.2 BOSS and Jedra 38

F.3 Small-d stress test against nonstationary baselines 38

G Checks of the theory and controls **38**

G.1 The conditions of the analysis in the runs 39

G.2 Probe-rate ablation 39

G.3 Robustness and necessity experiments 40

G.4 Subspace recovery rate and change-point adaptation 40

G.5 The probes given to ambient baselines 41

Table 4: **Map of the experiments.** For each study, the question it asks, the result it tests, what it finds, and where it appears.

Experiment	Question	Theory	Finding	Where
Crossover grid	Where does SPSC beat LinUCB?	Thm. 4.1, the crossover	It wins 37 of 40 cells, losing only at $d \leq 10$	Tab. 12
Real data	Does the gain hold on real data?	—	An SPSC variant wins 51 of 58 cells	Tabs. 13–19
BOSS and Jedra	Does tracking the subspace beat learning it once?	Tab. 6, the stationary case	Both SPSC variants win every cell	Tab. 3, App. F.2
Small-d stress test	Does SPSC help at small d?	—	Lowest regret of all but the oracle	Fig. 3
Conditions of the analysis	Do (WC) and (R) hold in the runs?	Thm. 4.1, conditions (WC) and (R)	Both hold, and the refresh schedule does not matter	Tab. 20
Probe rate	Which probe rate is best?	Thm. 4.1, $m_k \asymp \ell_k^{2/3}$	A probe every 10 to 30 rounds	Fig. 4
Robustness and necessity	Which conditions of the theory matter?	Thm. 2.1(ii)	Only probe coverage	Fig. 5
Subspace recovery	Does the error behave as the theory says?	Thm. 2.1(i), Cor. B.10	It falls as probes arrive and jumps back at each change	Figs. 6, 7
Probes given to baselines	Do the probes alone explain the gain?	—	No, an SPSC variant wins 82 of 98 cells	Tab. 21

A Related work in detail

This appendix expands Table 1 with the model and rates of each line of work, and then reads Theorem 4.1 term by term.

A.1 Low-rank and bilinear bandits

In the bilinear model of Jun et al. (2019) the learner plays (x_t, z_t) and observes $y_t = x_t^\top \Theta^\star z_t + \eta_t$ with $\mathrm{rank}(\Theta^\star) = r$. Their method estimates the top-r singular subspaces of $\Theta^\star$ once and then runs a ridge regression that penalizes the other directions more heavily, with regret $\widetilde{\mathcal{O}}((d_1+d_2)^{3/2}\sqrt{rT})$. Later work sharpens the rate (Lu et al., 2021; Jang et al., 2021; Kang et al., 2022; Jang et al., 2024; Stojanovic et al., 2023), for example to $\widetilde{\mathcal{O}}(r^{5/4}(m+n)^{3/4}\sqrt{T})$ (Jedra et al., 2024). With a known subspace $B^\star$, OFUL on the projected actions $B^{\star\top}x$ has regret $\widetilde{\mathcal{O}}(r\sqrt{T})$. All of these assume a fixed subspace. After a change, an old estimate with $\mathrm{range}(\widehat{U}) \perp \mathrm{range}(B^\star_{k+1})$ gives $x^\top\theta_t = 0$ for every projected action, so the new reward directions are invisible. Within a segment our θ_t also moves. Its mean $\mathbb{E}[\theta_t] = B^\star_k \bar{a}_{k,t}$ spans at most one direction, while $\mathbb{E}[\theta_t\theta_t^\top] = B^\star_k(\bar{a}_{k,t}\bar{a}_{k,t}^\top + \Sigma_{\eta,k})B^{\star\top}_k$ has rank r. Least squares, which these methods build on, therefore cannot recover the subspace, and SPSC uses quadratic measurements instead.

A.2 Nonstationary bandits

With Γ_T changes among N arms, discounted, sliding-window, and change-detection methods reach $\widetilde{\mathcal{O}}(\sqrt{\Gamma_T NT})$ (Garivier & Moulines, 2011; Cao et al., 2019; Besson et al., 2022), also without knowing Γ_T (Auer et al., 2019). Under a variation budget V_T the rate is $(NV_T)^{1/3}T^{2/3}$ (Besbes et al., 2014). In linear bandits, SW-LinUCB (Cheung et al., 2019) runs ridge regression on the last W rounds, and D-LinUCB (Russac et al., 2019) discounts round s by γ^{t-s}. Their reported rate $\widetilde{\mathcal{O}}(T^{2/3}P_T^{1/3})$, with path length $P_T = \sum_t \|\theta_{t+1} - \theta_t\|_2$, was corrected to $\widetilde{\mathcal{O}}(T^{3/4}P_T^{1/4})$ by Zhao et al. (2021). Our drift step avoids the same flaw through the constant C_W of Lemma B.15. Extensions cover variance, generalized linear rewards, misspecification, and best-arm identification (Wang et al., 2025b; Lee & Oh, 2026; Hu et al., 2026; Hou et al., 2024). All these methods work in $\mathbb{R}^d$. Their confidence radius grows like $\sqrt{d}$ and their elliptical potential like $d\log T$, so their regret carries d even when every θ_t lies in an r-dimensional subspace.

A.3 Nonstationary and low-rank

Qin et al. (2022) study tasks with fixed parameters $\theta_s = Bw_s$, each played for N rounds, where B changes with the environment. They estimate each θ_s by least squares in $\mathbb{R}^d$, so their regret over τ tasks keeps a $dr\sqrt{\tau N}$ term. Our θ_t changes every round, so the subspace must come from second moments. Cheng et al. (2023), Mueller et al. (2019), and Enikeeva et al. (2025) treat low-rank MDPs, dynamic pricing, and change-point detection, not bandit regret under a moving subspace.

A.4 Costly information and probing

Paid-observation and costly-probe bandits charge c per observation (Seldin et al., 2014; Tucker et al., 2023; Boyer et al., 2025; Elumar et al., 2024), with regret $\Theta(c^{1/3}T^{2/3})$ in the basic case (Tucker et al., 2023). The rate comes from balancing cm against $\ell/\sqrt{m}$, which gives $c^{1/3}\ell^{2/3}$ at $m \asymp (\ell/c)^{2/3}$. SPSC faces the same balance in each segment, $Am_k + B\ell_k/\sqrt{m_k}$, but its B carries the $\sqrt{d}$ of estimating a subspace in $\mathbb{R}^d$, which gives the $d^{1/3}$ of Theorem 4.1. These works learn a fixed parameter, and their observations reveal rewards, not second moments. Guan (2026) instead lets the learner probe up to k of n arms per round and observe the largest reward, a budget rather than a fee.

A.5 Subspace recovery from quadratic measurements

Our probe statistic follows the quadratic-sampling model $\mathbb{E}[\psi_t \mid \mathcal{H}_{t-1}, u_t] = u_t^\top \widetilde{M}_t u_t$. In batch form, a convex program recovers a rank-r matrix from $m \gtrsim dr$ such measurements (Chen et al., 2015), and rank-one projections and covariance sketches give similar guarantees (Cai & Zhang, 2015; Dasarathy et al., 2015). SPSC instead streams the moment estimator $\widehat{M}_k = m_k^{-1}\sum_{t\in\mathcal{T}_k} \mathcal{K}^{-1}(\psi_t u_t u_t^\top)$, at $O(d^2)$ per probe, while its target changes at unknown times. Identifiability and recovery results for matrix completion, adaptive sensing, and a fixed subspace (Király

& Tomioka, 2012; Kacham & Woodruff, 2023; Zhang & Zhou, 2024; Jedra et al., 2024) do not ask when a moving subspace can be recovered from single-play rewards.

A.6 Representation learning in bandits

Multi-task bandits learn a representation shared by all tasks, $\theta_s = Bw_s$, and replace d by r in the dominant regret term (Yang et al., 2021; Hu et al., 2021; Du et al., 2023). Duong et al. (2024) transfer it across a sequence of tasks, and their BOSS method is one of our baselines. In all of these works the representation is fixed within a stream.

A.7 Terms and special cases of Theorem 4.1

Table 5 gives the source of each term of the full bound, Theorem B.19, and what it depends on. The ambient dimension enters only the acquisition term, as $d^{1/3}$, and the threshold T_0.

Table 5: **The full bound term by term**, with the parameters each term depends on. The first two terms form Theorem 4.1.

Term	Source	Depends on
$\widetilde{\mathcal{O}}(r\sqrt{KT})$	projected ridge-UCB, restarted per segment	r, K, λ, S_w, $\kappa_z^{-1/2}$
$\widetilde{\mathcal{O}}((rd)^{1/3}\Xi K^{1/3}T^{2/3})$	online subspace acquisition	$(rd/\kappa_z)^{1/3}$; via Ξ: $\lambda_{\min}^{-2/3}$, S_w, $R_{\mathcal{A}}$, c
$\widetilde{\mathcal{O}}(rT/\sqrt{W})$	a window shorter than the segments	r, W, $\kappa_z^{-1/2}$ (Cor. 4.2)
$O(C_W R_{\mathcal{A}} W V_{\text{in}})$	drift inside a window, basis fixed	W, V_{in}, C_W (Cor. 4.2)
burn-in, $\propto q^\star K^{2/3}T^{1/3}$	first probes of each segment	made lower order by $T \geq T_0$

Table 6 compares the rate with existing methods and lists special cases. With $r = d$ and no probes SPSC is restarted LinUCB, and with a known subspace it is projected OFUL, so the bound recovers both rates.

Table 6: **Rates of comparable methods and special cases of Theorem 4.1.**

Method or case	Setting	Rate	Requires or matches
LinUCB, restarted	nonstat., ambient	$\widetilde{\mathcal{O}}(d\sqrt{KT})$	oracle boundaries
D-LinUCB, SW-LinUCB	nonstat., ambient	$\widetilde{\mathcal{O}}(T^{3/4})$ after correction (Zhao et al., 2021)	variation budget
LowOFUL	stat. low-rank	$\widetilde{\mathcal{O}}(r\sqrt{T})$	oracle subspace
ESTR (Jun et al., 2019)	stat. low-rank	$\widetilde{\mathcal{O}}((d_1+d_2)^{3/2}\sqrt{rT})$	fixed subspace
BOSS	stat. low-rank	$\widetilde{\mathcal{O}}(r\sqrt{dT})$	fixed subspace
Jedra	stat. low-rank	$\widetilde{\mathcal{O}}(r^{5/4}(m+n)^{3/4}\sqrt{T})$	fixed subspace
SPSC (Thm. 4.1)	piecewise low-rank	$\widetilde{\mathcal{O}}(r\sqrt{KT}) + \widetilde{\mathcal{O}}((rd)^{1/3}\Xi K^{1/3}T^{2/3})$	none of these
SPSC, $r = d$, no probes		$\widetilde{\mathcal{O}}(d\sqrt{KT})$	restarted LinUCB
SPSC, oracle subspace		$\widetilde{\mathcal{O}}(r\sqrt{KT})$	projected OFUL
SPSC, drifting means (Cor. 4.2)		adds $\widetilde{\mathcal{O}}(r^{2/3}V_{\text{in}}^{1/3}T^{2/3})$ at $W^\star$	well-conditioned windows
SPSC, fixed $\theta^{(k)}$ per segment		$\widetilde{\mathcal{O}}(r\sqrt{KT}) + \widetilde{\mathcal{O}}((rd)^{1/3}K^{1/3}T^{2/3})$	$\lambda_{\min} = \min_k \|\theta^{(k)}\|^2$

B Proofs

This section proves the statements of §4 in dependency order:

1. §B.1 fixes notation and establishes the probe-time excitation lemma used under full excitation.
2. §B.2 derives the *identification chain*. It gives the closed-form inverse of the probe operator $\mathcal{K}$, the quadratic-measurement identity, the lifted probe sample's near-unbiasedness, a variance-aware matrix Freedman bound, and the projector error via Davis–Kahan.

3. §B.3 establishes the necessity half of Theorem 2.1 via three propositions (Prop. B.11–B.13). Coverage is necessary for every probe law. Without known variance or bounded coupling, identification can fail once the noise may depend on the probe direction.
4. §B.4 states and proves the interleaved-analysis lemmas: projected nonstationary UCB with time-varying basis (Lem. B.15), uniform prefix subspace concentration (Lem. B.16), and cumulative interleaved subspace error (Lem. B.17).
5. §B.5 proves a full bound that keeps every term (Thm. B.19), using the probe-cost lemma (Lem. B.18) and Lem. B.15. It then derives Theorem 4.1 and Corollary 4.2 and defines the threshold T_0.
6. §B.6 lifts the guarantee to SPSC-Adaptive (Prop. 4.3) by controlling the projector-based detector.

The windowed projected self-normalized bound (Lem. B.3), the width bound under (WC) (Lem. B.2), and the projected UCB lemma with an estimated basis (Lem. B.15) are the linear-bandit ingredients; constants that do not affect leading order are absorbed into $\widetilde{\mathcal{O}}(\cdot)$.

B.1 Preliminaries and notation

Table 7 collects the notation of the main text, with the type of each quantity and whether the learner knows it.

Throughout, $\mathcal{I}_k = [\tau_{k-1}, \tau_k)$ denotes segment k of length $\ell_k := |\mathcal{I}_k|$, so $\sum_k \ell_k = T$. Inside $\mathcal{I}_k$ the learner performs $m_k = |\mathcal{T}_k|$ probe rounds ($\mathcal{T}_k := \mathcal{T}_{\text{probe}} \cap \mathcal{I}_k$) and $n_k = \ell_k - m_k$ exploitation rounds ($E_k := \mathcal{I}_k \setminus \mathcal{T}_k$). We write $\widehat{P}_t = \widehat{U}_t \widehat{U}_t^\top$ for the estimated rank-r projector, $P_k^\star$ for the projector onto the effective subspace S_k defined below, and $\varepsilon_k := \|(I - \widehat{P}_k)P_k^\star\|_{\text{op}}$ for the part of S_k that the estimate misses. We work on the high-probability event $\mathcal{G}_w := \{\max_{t \le T} \|w_t\|_2 \le S_w\}$. For sub-Gaussian innovations one may take $S_w = \widetilde{\mathcal{O}}(\sqrt{r})$ via union bound (the failure probability is absorbed into the overall δ budget); for bounded innovations $\mathcal{G}_w$ holds deterministically. In what follows we implicitly condition on $\mathcal{G}_w$.

The *predictable second moment* is $\widetilde{M}_t := \mathbb{E}[\theta_t \theta_t^\top \mid \mathcal{H}_{t-1}] = B_k^\star \mathbb{E}[w_t w_t^\top \mid \mathcal{H}_{t-1}](B_k^\star)^\top$, and the probe-time average is $\bar{M}_k^{\text{probe}} := m_k^{-1} \sum_{t \in \mathcal{T}_k} \widetilde{M}_t$. The *effective subspace* of segment k is $S_k := \text{range}(\bar{M}_k^{\text{probe}}) \subseteq \text{range}(B_k^\star)$. Its dimension is $s_k \le r$, and $P_k^\star$ is the orthogonal projector onto it. The proofs use the following condition.

Condition (E). For every segment k and every $t \in \mathcal{I}_k$, $\text{range}(\widetilde{M}_t) \subseteq S_k$ and $\mu_t \in S_k$. Moreover, every prefix average over the probe rounds of $\mathcal{I}_k$ has s_k-th eigenvalue at least $\lambda_{\min} > 0$, that is, $\lambda_{s_k}(q^{-1} \sum_{j=1}^q \widetilde{M}_{t_j}) \ge \lambda_{\min}$ for the first $q \ge 1$ probe rounds $t_1 < \cdots < t_q$ of the segment.

Two cases satisfy Condition (E). Under full excitation, $\Sigma_{\eta,k} \succ 0$, the effective subspace is all of $\text{range}(B_k^\star)$ and Lemma B.1 gives $\lambda_{\min} = \min_k \lambda_r(\Sigma_{\eta,k})$. Under the constant-mean model of Theorem 4.1, $A_k = 0$ and $\bar{a}_{k,t} \equiv \bar{a}_k$, so $\widetilde{M}_t = B_k^\star(\bar{a}_k \bar{a}_k^\top + \Sigma_{\eta,k}) B_k^{\star\top}$ is the same in every round of the segment. Its range contains $\mu_t = B_k^\star \bar{a}_k$, and $\lambda_{\min}$ is the smallest nonzero eigenvalue of $\bar{a}_k \bar{a}_k^\top + \Sigma_{\eta,k}$ over k. A fixed $\theta^{(k)}$ per segment is the case $\Sigma_{\eta,k} = 0$. It gives $S_k = \text{span}(\theta^{(k)})$ and $\lambda_{\min} = \min_k \|\theta^{(k)}\|^2$. Constants $C, C', \ldots$ depend only on $\sigma_\varepsilon, \widehat{\sigma}^2, R_{\mathcal{A}}, \rho_Q, S_w, \lambda_{\min}$ and may change line to line.

Lemma B.1 (Predictable probe-time excitation)**.** *If $\Sigma_{\eta,k} \succ 0$, then for each segment k and every nonempty set $\mathcal{T}$ of probe rounds of $\mathcal{I}_k$,*

$$\lambda_r\left(\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbb{E}[w_t w_t^\top \mid \mathcal{H}_{t-1}]\right) \;\ge\; \lambda_r(\Sigma_{\eta,k}) \;>\; 0 \qquad \textit{almost surely.}$$

Proof. For each $t \in \mathcal{I}_k$, write $w_t = \bar{w}_t + \eta_{t-1}$ with $\bar{w}_t := \bar{a}_{k,t} + A_k(w_{t-1} - \bar{a}_{k,t-1})$, which is independent of the zero-mean η_{t-1}. Then $\mathbb{E}[w_t w_t^\top \mid \mathcal{H}_{t-1}] = \mathbb{E}[\bar{w}_t \bar{w}_t^\top \mid \mathcal{H}_{t-1}] + \Sigma_{\eta,k} \succeq \Sigma_{\eta,k}$ (LDS recursion + non-degenerate innovations). Hence $\lambda_r(\mathbb{E}[w_t w_t^\top \mid \mathcal{H}_{t-1}]) \ge \lambda_r(\Sigma_{\eta,k}) > 0$ in every round. Averaging over any set of probe rounds preserves this PSD lower bound, which gives the claim. Stability of A_k is not needed for the lemma. The innovation covariance alone suffices. □

Table 7: **Main notation, types, and what the learner knows.**

Symbol	Type	Meaning	Status
$d,\ r,\ T$	integers	ambient dimension, rank, horizon	known[a]
$K,\ \tau_k,\ \mathcal{I}_k$	integers, sets	number of segments, change points, segments	Alg. 1 only[d]
$\mathcal{A}_t$	set in $\mathbb{R}^d$	actions offered at round t, $\|x\| \le R_{\mathcal{A}}$	revealed
$\mathcal{H}_{t-1}$	history	actions, probes and rewards of the first $t-1$ rounds	revealed
$\theta_t,\ \mu_t$	vectors in $\mathbb{R}^d$	reward parameter ($y_t = x_t^\top \theta_t + \varepsilon_t$), its mean	unknown
$B_k^\star$	$d \times r$ matrix	orthonormal basis of segment k's subspace	unknown
$w_t,\ \bar{a}_{k,t}$	vectors in $\mathbb{R}^r$	latent state ($\theta_t = B_k^\star w_t$), its mean path	unknown
$R_{\mathcal{A}},\ S_w,\ S_\Delta$	scalars	bounds $\|x\| \le R_{\mathcal{A}}$, $\|w_t\| \le S_w$, $\|\theta_t - \mu_t\| \le S_\Delta$	known
$A_k,\ \Sigma_{\eta,k}$	$r \times r$ matrices	latent dynamics, innovation covariance	unknown
$\varepsilon_t,\ \sigma_\varepsilon$	scalars	observation noise, its sub-Gaussian scale	upper bound known[b]
$Q,\ u_t,\ c$	law, vector, scalar	exploration distribution, probe, probe cost	known
$\widetilde{M}_t,\ \widehat{M}_k$	$d \times d$ matrices	predictable second moment, lifted estimate	$\widehat{M}_k$ computed
$S_k,\ s_k$	subspace, integer	effective subspace $\mathrm{range}(\bar{M}_k^{\mathrm{probe}}) \subseteq \mathrm{range}(B_k^\star)$, its dimension	unknown
$\widehat{U}_t,\ \widehat{U}_k,\ \widehat{P}_k,\ P_k^\star$	matrices	basis after round t, last basis of segment k, its projector; projector onto S_k	first three computed
$W,\ \lambda,\ \delta,\ \widehat{\sigma}^2$	scalars	exploitation window, ridge parameter, confidence, centering	chosen
$\lambda_{\min}$	scalar	smallest nonzero eigenvalue of the probe-time second moment	in bounds only[c]
$\rho_z,\ n_0,\ \kappa_z$	scalars	window conditioning (WC), its burn-in, and $\kappa_z = \rho_z r / R_{\mathcal{A}}^2$	in bounds only
V_{in}	scalar	within-segment variation of μ_t	in bounds only
$\mathcal{T}_{\mathrm{probe}},\ m_k,\ p$	set, integer, scalar	probe rounds, their number in segment k, probe rate	chosen
$\psi_t,\ G_t$	scalar, $d \times d$ matrix	probe statistic $y_t^2 - \widehat{\sigma}^2$, lifted probe sample $\mathcal{K}^{-1}(\psi_t u_t u_t^\top)$	computed
$\mathcal{W}_t,\ z_t(x),\ \beta_t^{(r,W)}$	set, vector in $\mathbb{R}^r$, scalar	exploitation window, projected action $\widehat{U}_{t-1}^\top x$, ridge radius	computed
$s,\ \eta_P,\ W_{\mathrm{det}},\ \delta_{\mathrm{FA}}$	integer, scalars	common effective rank, projector-error bound, detector window, false-alarm level	SPSC-Adaptive inputs

[a] Any run rank $r \ge s_k$ suffices (Corollary B.10). [b] The noise variance itself need not be known (Theorem 2.1). [c] SPSC-Adaptive's detector (§3) takes a lower bound on it. [d] SPSC-Adaptive takes an upper bound on K.

Notation for the interleaved analysis. For each segment $\mathcal{I}_k = [\tau_{k-1}, \tau_k)$, recall $\ell_k = |\mathcal{I}_k|$, $\mathcal{T}_k = \mathcal{T}_{\mathrm{probe}} \cap \mathcal{I}_k$, $m_k = |\mathcal{T}_k|$, $E_k = \mathcal{I}_k \setminus \mathcal{T}_k$, $n_k = |E_k| = \ell_k - m_k$. For $t \in \mathcal{I}_k$, let $q_k(t) := |\mathcal{T}_k \cap [\tau_{k-1}, t)|$ be the number of probes collected in segment k strictly before round t. Denote $\widehat{P}_t = \widehat{U}_t \widehat{U}_t^\top$ and $\varepsilon_{k,t} := \|(I - \widehat{P}_t) P_k^\star\|_{\mathrm{op}}$, the part of S_k that the basis in use after round t misses. The within-segment variation is $V_k^{\mathrm{in}} := \sum_{s, s+1 \in \mathcal{I}_k} \|\mu_{s+1} - \mu_s\|_2$, $V_{\mathrm{in}} := \sum_k V_k^{\mathrm{in}}$, where $\mu_t = \theta_t - B_k^\star \eta_{t-1}$ is the mean given the latent past (§2). Since η_{t-1} is independent of the past and of $\mathcal{A}_t$, $\xi_t := y_t - x_t^\top \mu_t = \varepsilon_t + x_t^\top B_k^\star \eta_{t-1}$ has conditional mean zero given the past, the latent path up to $t-1$, and x_t. For bounded innovations, $\|\mu_t\| \le S_w$ (as μ_t averages θ_t over η_{t-1}) and $|x_t^\top B_k^\star \eta_{t-1}| = |x_t^\top (\theta_t - \mu_t)| \le R_{\mathcal{A}} S_\Delta \le 2 R_{\mathcal{A}} S_w$, so ξ_t is σ_{eff}-sub-Gaussian with $\sigma_{\mathrm{eff}} := \sigma_\varepsilon + R_{\mathcal{A}} S_\Delta$; sub-Gaussian innovations change these constants by logarithmic factors.

Balanced interleaved probe schedule. The probe schedule is balanced: there exists an absolute constant a_0 such that, for every segment k and every $j \in \{0, 1, \dots, m_k\}$,

$$\big|\{t \in E_k : q_k(t) = j\}\big| \;\le\; a_0\lceil \ell_k/m_k \rceil. \tag{B}$$

Uniformly spaced probes satisfy this condition.

Refresh schedule and window conditioning. Alg. 1 recomputes $\widehat{U}$ after every probe, so the basis in use at round t is built from all $q_k(t)$ probes of the segment before t. The analysis needs only condition (R): the basis in use at round t is built from the first $j_k(t)$ probes of the segment, for some $j_k(t) \ge q_k(t)/2$. A refresh after the 1st, 2nd, 4th, … probe also satisfies it. Write $\omega_t(x) := \|z_t(x)\|_{\widetilde{V}_t^{-1}}$ with $z_t(x) = \widehat{U}_{t-1}^\top x$, and $L_W := \log(1 + WR_{\mathcal{A}}^2/(\lambda r))$. Condition (WC) of §4 asks that the smallest eigenvalue of $\sum_{s\in\mathcal{W}_t} z_t(x_s)z_t(x_s)^\top$ be at least $\rho_z|\mathcal{W}_t|$ whenever $|\mathcal{W}_t| \ge n_0$. The trace of this matrix is at most $|\mathcal{W}_t|R_{\mathcal{A}}^2$, so $\rho_z \le R_{\mathcal{A}}^2/r$, and $\kappa_z := \rho_z r/R_{\mathcal{A}}^2 \in (0, 1]$.

Lemma B.2 (Widths under (WC))**.** *Assume (WC), $W \ge 2n_0+2$, and probes evenly spaced at a rate of at most one per two rounds. Then $|\mathcal{W}_t| < n_0$ only on the first $2n_0$ rounds of the segment. For every exploitation round t with $|\mathcal{W}_t| \ge n_0$ and every x with $\|x\| \le R_{\mathcal{A}}$, $\omega_t(x)^2 \le R_{\mathcal{A}}^2/(\lambda + \rho_z|\mathcal{W}_t|)$, and*

$$\sum_{t\in E_k:\,|\mathcal{W}_t|\ge n_0} \min\{1, \omega_t(x_t)^2\} \le \frac{R_{\mathcal{A}}^2}{\rho_z}\Big(1 + \ln n_k + \frac{4n_k}{W}\Big).$$

Proof. $\omega_t(x)^2 \le \|z_t(x)\|^2/\lambda_{\min}(\widetilde{V}_t)$ and $\|z_t(x)\| \le \|x\|$, which gives the width bound. $|\mathcal{W}_t|$ counts the exploitation rounds of the segment in $[t - W, t)$. During the first W rounds of the segment it is the number e_t of earlier exploitation rounds, which grows by one at each exploitation round, so $e_t \ge n_0$ after the first $2n_0$ rounds, and these rounds contribute at most $\sum_{e=1}^{n_k} R_{\mathcal{A}}^2/(\rho_z e)$. Later, the last W rounds contain at most $W/2 + 1$ probes, so $|\mathcal{W}_t| \ge W/2 - 1 \ge \max\{n_0, W/4\}$, and each such round contributes at most $4R_{\mathcal{A}}^2/(\rho_z W)$. □

Lemma B.3 (Windowed projected self-normalized concentration)**.** *Fix a segment k, write $z_t(x) := \widehat{U}_{t-1}^\top x$, and let $A_k = 0$ with the noise and innovation of each round independent of earlier rounds. The noise $\xi_s = y_s - x_s^\top \mu_s$ is σ_{eff}-sub-Gaussian. With probability at least $1 - \delta/K$, simultaneously for all $t \in E_k$,*

$$\Big\|\sum_{s\in\mathcal{W}_t} z_t(x_s)\,\xi_s\Big\|_{\widetilde{V}_t^{-1}} + \sqrt{\lambda}\,S_w \le \beta_t^{(r,W)},$$

$$\beta_t^{(r,W)} := \sigma_{\text{eff}}\sqrt{r\log\big(1 + WR_{\mathcal{A}}^2/(\lambda r)\big) + 2\log(2KT^2/\delta)} + \sqrt{\lambda}\,S_w.$$

This is the windowed projected specialization of Abbasi-Yadkori et al. (2011, Thm. 1, 2), applied once for each basis and each window start. Each basis is computed from probe rounds only. Under the two assumptions, which Theorem 4.1 and Corollary 4.2 make, the probe rounds after an exploitation round s are independent of ξ_s. Conditioning on all probe data therefore fixes the bases and keeps (ξ_s) a martingale difference sequence, so the rounds of earlier blocks may be projected with the current basis. A union bound over the at most T bases and T window starts of a segment gives the term $2\log(2KT^2/\delta)$. Under latent dynamics, $A_k \ne 0$, later probes depend on earlier innovations. The lemma then holds for the variant that empties the window at each refresh, since every round in its window comes after the basis it uses, and it needs only that ξ_s has conditional mean zero.

B.2 Concentration of the lifted probe moment

This is the non-standard piece of the analysis; we derive it in full. We work with scaled-sphere probes $u_t = \sqrt{d}\,v_t$, $v_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, which satisfy $\|u_t\|_2 = \sqrt{d}$ *deterministically* and $\mathbb{E}[u_t u_t^\top] = I_d$. The probe moment operator $\mathcal{K} : M \mapsto \mathbb{E}[(u^\top M u)uu^\top]$ is a linear bijection on the space of symmetric $d \times d$ matrices.

Lemma B.4 (Closed-form inverse of $\mathcal{K}$ on symmetric matrices)**.** *For scaled-sphere probes $u = \sqrt{d}\,v$, $v \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, $\mathcal{K}$ acts on symmetric $d \times d$ matrices as $\mathcal{K}(M) = \frac{d}{d+2}\big(\mathrm{tr}(M)\,I_d + 2M\big)$ with inverse $\mathcal{K}^{-1}(N) = \frac{d+2}{2d}\,N - \frac{\mathrm{tr}(N)}{2d}\,I_d$. Moreover $\|\mathcal{K}^{-1}\|_{\mathrm{op}\to\mathrm{op}} \le 1$ on symmetric inputs.*

Proof. Use the fourth-moment identity for the uniform sphere $\mathbb{E}[v_i v_j v_k v_l] = (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})/(d(d+2))$ and the rescaling $u = \sqrt{d}\, v$: $\mathbb{E}[(u^\top M u)u_k u_l] = d^2\mathbb{E}[(v^\top M v)v_k v_l] = d^2 \cdot (2M_{kl} + \mathrm{tr}(M)\delta_{kl})/(d(d+2)) = \frac{d}{d+2}(2M_{kl} + \mathrm{tr}(M)\delta_{kl})$ (using M symmetric). So $\mathcal{K}(M) = \frac{d}{d+2}(\mathrm{tr}(M)I_d + 2M)$. Taking traces of $N = \mathcal{K}(M)$ gives $\mathrm{tr}(N) = d\ \mathrm{tr}(M)$, so $\mathrm{tr}(M) = \mathrm{tr}(N)/d$ and $M = \frac{d+2}{2d}N - \frac{\mathrm{tr}(N)}{2d}I_d$.

For the operator-norm bound: $\mathcal{K}^{-1}(N)$ is diagonalized in the eigenbasis of N; if N has eigenvalues $\mu_1, \ldots, \mu_d$, then $\mathcal{K}^{-1}(N)$ has eigenvalues $\nu_i = \frac{d+2}{2d}\mu_i - \frac{\mathrm{tr}(N)}{2d}$. Maximizing $|\nu_i|$ over $\|N\|_{\mathrm{op}} = 1$ (i.e. $\max_i |\mu_i| = 1$): at $\mu_1 = 1, \mu_2 = \cdots = \mu_d = -1$, $\mathrm{tr}(N) = 2 - d$ and $\nu_1 = \frac{d+2-(2-d)}{2d} = 1$; at $\mu_1 = -1, \mu_2 = \cdots = \mu_d = 1$, $\nu_1 = -1$. All other eigenvalue patterns give $|\nu_i| \le 1$ by direct case analysis. So $\|\mathcal{K}^{-1}\|_{\mathrm{op}\to\mathrm{op}} = 1$ on symmetric matrices for $d \ge 2$. □

Lemma B.5 (Quadratic measurement identity)**.** *On every probe round $t \in \mathcal{T}_{\mathrm{probe}}$, let $\sigma_t^2 := \mathbb{E}[\varepsilon_t^2 \mid \mathcal{H}_{t-1}, u_t]$ be the conditional noise variance, $\delta_\sigma := \widehat{\sigma}^2 - \sigma_t^2$, and $m_t := \mathbb{E}[\varepsilon_t \theta_t \mid \mathcal{H}_{t-1}, u_t]$. Then*

$$\mathbb{E}[\psi_t \mid \sigma(\mathcal{H}_{t-1}, u_t)] = u_t^\top \widetilde{M}_t u_t + 2u_t^\top m_t - \delta_\sigma.$$

Under probe independence neither σ_t^2 nor m_t depends on u_t. Both may vary with t. If $\delta_\sigma = 0$ and $m_t = 0$, the identity collapses to $\mathbb{E}[\psi_t \mid \cdot] = u_t^\top \widetilde{M}_t u_t$.

Proof. Expand $y_t^2 = (u_t^\top \theta_t)^2 + 2\varepsilon_t u_t^\top \theta_t + \varepsilon_t^2$ and condition on $\sigma(\mathcal{H}_{t-1}, u_t)$. The middle term contributes $2u_t^\top m_t$. The last term contributes σ_t^2, and subtracting $\widehat{\sigma}^2$ leaves $-\delta_\sigma$. The step $\mathbb{E}[(u_t^\top \theta_t)^2 \mid \mathcal{H}_{t-1}, u_t] = u_t^\top \widetilde{M}_t u_t$ uses probe independence, the independence of u_t from $(\theta_t, \varepsilon_t)$ given $\mathcal{H}_{t-1}$. The same property makes σ_t^2 and m_t free of u_t. □

Lemma B.6 (Lifted probe sample is near-unbiased)**.** *Let $G_t := \mathcal{K}^{-1}(\psi_t u_t u_t^\top)$ for scaled-sphere $u_t = \sqrt{d}\, v_t$, $v_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and assume probe independence. Then $\mathbb{E}[G_t \mid \mathcal{H}_{t-1}] = \widetilde{M}_t + \widetilde{B}_t$ with bias*

$$\widetilde{B}_t = -\tfrac{\delta_\sigma}{d}\, I_d, \qquad \|\widetilde{B}_t\|_{\mathrm{op}} = \tfrac{|\delta_\sigma|}{d},$$

whatever the size of the cross-moment m_t. The bias is a scaled identity, hence shifts every eigenvalue uniformly and does not rotate eigenvectors of $\widetilde{M}_t$.

Proof. By linearity of $\mathcal{K}^{-1}$ and Lemma B.5, $\mathbb{E}[G_t \mid \mathcal{H}_{t-1}] = \mathcal{K}^{-1}(\mathcal{K}(\widetilde{M}_t)) + 2\,\mathcal{K}^{-1}(\mathbb{E}[(u_t^\top m_t)u_t u_t^\top \mid \mathcal{H}_{t-1}]) - \delta_\sigma\, \mathcal{K}^{-1}(\mathbb{E}[u_t u_t^\top]) = \widetilde{M}_t + \widetilde{B}_t$, with $\widetilde{M}_t = \mathcal{K}^{-1}\mathcal{K}(\widetilde{M}_t)$.

Variance-bias term. $\mathbb{E}[u_t u_t^\top] = I_d$ exactly for scaled-sphere probes, so $-\delta_\sigma \mathcal{K}^{-1}(I_d) = -\delta_\sigma\big[\frac{d+2}{2d} I_d - \frac{d}{2d} I_d\big] = -\delta_\sigma \cdot \frac{2}{2d} I_d = -\frac{\delta_\sigma}{d} I_d$. This is a scaled identity, hence does not rotate eigenvectors of $\widetilde{M}_t$.

State-noise coupling term. The remaining contribution is $2\,\mathcal{K}^{-1}\big(\mathbb{E}[(u_t^\top m_t)\, u_t u_t^\top \mid \mathcal{H}_{t-1}]\big)$. Under probe independence, $m_t = \mathbb{E}[\varepsilon_t \theta_t \mid \mathcal{H}_{t-1}, u_t] = \mathbb{E}[\varepsilon_t \theta_t \mid \mathcal{H}_{t-1}]$ does not depend on u_t. The (a, b)-entry of the inner expectation is $\sum_c (m_t)_c\, \mathbb{E}[(u_t)_a (u_t)_b (u_t)_c]$, a degree-three polynomial in u_t. For sphere-symmetric probes $u_t \mapsto -u_t$ leaves the distribution invariant, so every odd-degree moment vanishes and $\mathbb{E}[(u_t^\top m_t)\, u_t u_t^\top \mid \mathcal{H}_{t-1}] = 0$ exactly. Hence the coupling term contributes zero, and $\widetilde{B}_t = -\frac{\delta_\sigma}{d}\, I_d$. □

Sensitivity to probe independence. Probe independence is a modeling assumption, so we check how much the estimate changes when it fails. A coupling that depends on the probe direction biases the lifted estimate by at most twice its size, with no factor of d, and a coupling aligned with the signal leaves the subspace unchanged. Only the part of $m_t(u) := \mathbb{E}[\varepsilon_t \theta_t \mid \mathcal{H}_{t-1}, u]$ that is odd in u can bias G_t, since the even part cancels by sphere symmetry. For the leading linear case $m_t^{\mathrm{odd}}(u) = Mu$ we have $\mathbb{E}[(u^\top M u)\, uu^\top] = \mathcal{K}(\mathrm{sym}\, M)$, so the bias is exactly $2\,\mathrm{sym}(M)$. Davis–Kahan then gives $\|\widehat{P}_k - P_k^\star\|_{\mathrm{op}} \le 8\|\mathrm{sym}\, M\|_{\mathrm{op}}/\lambda_{\min}$ plus sampling error whenever $\|\mathrm{sym}\, M\|_{\mathrm{op}} \le \lambda_{\min}/4$. If the coupling is aligned with the signal, $y_t = (1 + \epsilon_\times)u_t^\top \theta_t + \varepsilon_t$ as in the robustness sweep of App. G.3, then $\mathbb{E}[G_t \mid \mathcal{H}_{t-1}] = (1 + \epsilon_\times)^2 \widetilde{M}_t + \widetilde{B}_t$. This is a positive rescaling, which leaves the eigenvectors unchanged. Simulations with scaled-sphere probes confirm both statements. The subspace error

does not grow with $\epsilon_\times$ (Table 8), and the measured bias matches $2\,\mathrm{sym}(M)$ up to Monte-Carlo error, with relative error 0.0023, 0.0108, and 0.0360 at $d = 5, 20, 60$ (1.5×10^6 probe rounds per cell).

Table 8: **Subspace error under the aligned coupling** $y_t = (1+\epsilon_\times)u_t^\top \theta_t + \varepsilon_t$ (scaled-sphere probes, 8×10^5 probe rounds per cell).

Measured quantity	$\epsilon_\times = 0$	$\epsilon_\times = 0.5$	$\epsilon_\times = 1.0$
$\|\widehat{P} - P^\star\|_{\mathrm{op}}$, $d = 10, r = 2$	0.020	0.015	0.011
$\|\widehat{P} - P^\star\|_{\mathrm{op}}$, $d = 30, r = 3$	0.042	0.033	0.040
$\|\widehat{M} - (1+\epsilon_\times)^2\widetilde{M}\|/\|(1+\epsilon_\times)^2\widetilde{M}\|$, $d = 30$	0.028	0.033	0.028

Lemma B.7 (Dimension-free truncation of the probe statistic)**.** *Let* $L := \log(8Td^4/\delta)$ *and* $\tilde{R}_s := 4(S_w^2 + \sigma_\varepsilon^2)L + \widehat{\sigma}^2$. *With probability at least* $1 - \delta/2$, *every probe round has* $|u_t^\top \theta_t| \le S_w\sqrt{2L}$ *and* $|\varepsilon_t| \le \sigma_\varepsilon\sqrt{2L}$, *hence* $|\psi_t| \le \tilde{R}_s$, *with no factor of* d *outside the logarithm.*

Proof. Write $u_t = \sqrt{d}\, v_t$ with v_t uniform on the sphere. For a fixed unit vector e, $\Pr(|v^\top e| \ge a) \le 2e^{-da^2/2}$. By probe independence, conditional on $\mathcal{H}_{t-1}$, v_t is independent of θ_t and $\|\theta_t\| \le S_w$, so $\Pr(|u_t^\top \theta_t| > S_w\sqrt{2L} \mid \mathcal{H}_{t-1}) \le 2e^{-L} = \delta/(4Td^4)$; the sub-Gaussian noise obeys the same bound. A union bound over the at most T probe rounds gives probability at least $1 - \delta/(2d^4) \ge 1 - \delta/2$, and on this event $|\psi_t| \le (|u_t^\top \theta_t| + |\varepsilon_t|)^2 + \widehat{\sigma}^2 \le \tilde{R}_s$. □

Theorem B.8 (Matrix concentration for $\widehat{M}_k$)**.** *Let* $\widehat{M}_k := m_k^{-1}\sum_{t\in\mathcal{T}_k} G_t$. *For every* $m_k \ge 1$, *with probability at least* $1 - \delta$,

$$\|\widehat{M}_k - \bar{M}_k^{\mathrm{probe}} - \widetilde{B}\|_{\mathrm{op}} \le \tau(m_k, \delta) := \sqrt{2}\,\tilde{R}_s\sqrt{\frac{d\log(4d/\delta)}{m_k}} + \frac{4d\tilde{R}_s\log(4d/\delta)}{3m_k} + \frac{C(S_w^2 + \sigma_\varepsilon^2 + \widehat{\sigma}^2)}{\sqrt{T}}, \tag{6}$$

where $\widetilde{B} := m_k^{-1}\sum_{t\in\mathcal{T}_k}\widetilde{B}_t$ *averages the biases of Lemma B.6. It is a multiple of* I_d. *The second term of* τ *is at most the first once* $m_k \ge 8d\log(4d/\delta)$, *and then* $\tau(m_k, \delta) \le 2\sqrt{2}\,\tilde{R}_s\sqrt{d\log(4d/\delta)/m_k} + C(S_w^2 + \sigma_\varepsilon^2 + \widehat{\sigma}^2)/\sqrt{T}$.

Proof. Let $\mathcal{E}_t$ be the event of Lemma B.7 at round t; on the intersection of these events $G_t = G_t\mathbf{1}_{\mathcal{E}_t}$ at every probe round. Put $X_t := G_t\mathbf{1}_{\mathcal{E}_t} - \mathbb{E}[G_t\mathbf{1}_{\mathcal{E}_t} \mid \mathcal{H}_{t-1}]$. *Predictable variance.* For $\|u\|^2 = d$, $\mathcal{K}^{-1}(s\,uu^\top) = s[\frac{d+2}{2d}uu^\top - \frac{1}{2}I_d]$, and since $(uu^\top)^2 = d\,uu^\top$, $G_t^2 = \psi_t^2[\frac{d+2}{4}u_tu_t^\top + \frac{1}{4}I_d]$. With $(\psi_t\mathbf{1}_{\mathcal{E}_t})^2 \le \tilde{R}_s^2$ and $\mathbb{E}[u_tu_t^\top \mid \mathcal{H}_{t-1}] = I_d$, $\mathbb{E}[X_t^2 \mid \mathcal{H}_{t-1}] \preceq \mathbb{E}[(G_t\mathbf{1}_{\mathcal{E}_t})^2 \mid \mathcal{H}_{t-1}] \preceq \frac{d+3}{4}\tilde{R}_s^2 I_d$. *Envelope.* The eigenvalues of G_t are $\psi_t\frac{d+1}{2}$ and $-\psi_t/2$, so $\|X_t\|_{\mathrm{op}} \le 2d\tilde{R}_s$. *Truncation bias.* By Cauchy–Schwarz, $\|\mathbb{E}[G_t\mathbf{1}_{\mathcal{E}_t^c} \mid \mathcal{H}_{t-1}]\|_{\mathrm{op}} \le (\mathbb{E}\|G_t\|_{\mathrm{op}}^2)^{1/2}\Pr(\mathcal{E}_t^c)^{1/2} \le C(S_w^2 + \sigma_\varepsilon^2 + \widehat{\sigma}^2)/\sqrt{T}$, using $\|G_t\|_{\mathrm{op}} \le d|\psi_t|$, the second moment of ψ_t, and $\Pr(\mathcal{E}_t^c) \le \delta/(2Td^4)$; so $\mathbb{E}[G_t\mathbf{1}_{\mathcal{E}_t} \mid \mathcal{H}_{t-1}]$ is within this distance of $\widetilde{M}_t + \widetilde{B}_t$. *Freedman.* The matrix Freedman inequality (Tropp, 2011) gives, for every $m_k \ge 1$, with probability at least $1 - \delta/2$, $\|m_k^{-1}\sum_{t\in\mathcal{T}_k} X_t\|_{\mathrm{op}} \le \tilde{R}_s\sqrt{(d+3)\log(4d/\delta)/(2m_k)} + \frac{4d\tilde{R}_s\log(4d/\delta)}{3m_k}$. Since $(d+3)/2 \le 2d$, the first term is at most $\sqrt{2}\,\tilde{R}_s\sqrt{d\log(4d/\delta)/m_k}$, which gives (6). For $m_k \ge 8d\log(4d/\delta)$, $\frac{4}{3}\sqrt{d\log(4d/\delta)/m_k} \le \sqrt{2}$, so the second term is at most the first. □

Proposition B.9 (Segment factorization)**.** $\bar{M}_k^{\mathrm{probe}} = B_k^\star\bar{S}_k^{\mathrm{probe}}(B_k^\star)^\top$ *where* $\bar{S}_k^{\mathrm{probe}} := m_k^{-1}\sum_{t\in\mathcal{T}_k}\mathbb{E}[w_tw_t^\top \mid \mathcal{H}_{t-1}]$. *In particular,* $S_k = \mathrm{range}(\bar{M}_k^{\mathrm{probe}}) = B_k^\star\,\mathrm{range}(\bar{S}_k^{\mathrm{probe}}) \subseteq \mathrm{range}(B_k^\star)$, *with equality when* $\bar{S}_k^{\mathrm{probe}}$ *has rank* r, *for example under full excitation.*

Proof. $\widetilde{M}_t = B_k^\star\mathbb{E}[w_tw_t^\top \mid \mathcal{H}_{t-1}](B_k^\star)^\top$ for $t \in \mathcal{I}_k$. Average over $\mathcal{T}_k$ and pull $B_k^\star$ out. Since $B_k^\star$ has orthonormal columns, the range is $B_k^\star\mathrm{range}(\bar{S}_k^{\mathrm{probe}})$. Under full excitation $\lambda_r(\bar{S}_k^{\mathrm{probe}}) \ge \lambda_r(\Sigma_{\eta,k}) > 0$ (Lem. B.1), so the range is all of $\mathrm{range}(B_k^\star)$. □

Corollary B.10 (Projector concentration). *Let $s := s_k$ and let $\widehat{P}_k^{(s)}$ project onto the top-s eigenspace of $\widehat{M}_k$. For every $m_k \geq 1$, with probability at least $1-\delta$,*

$$\|\widehat{P}_k^{(s)} - P_k^\star\|_{\mathrm{op}} \leq \min\Big\{1,\ C_{\mathrm{sub}}\Big(\sqrt{\frac{d\log(4d/\delta)}{m_k}} + \frac{d\log(4d/\delta)}{m_k}\Big)\Big\}, \tag{7}$$

where $C_{\mathrm{sub}} := \big(8\sqrt{2}\,\tilde{R}_s + 4C(S_w^2 + \sigma_\varepsilon^2 + \widehat{\sigma}^2)\big)/\lambda_{\min}$. The second term in the parentheses is at most the first once $m_k \geq d\log(4d/\delta)$, so the bound is then at most $2C_{\mathrm{sub}}\sqrt{d\log(4d/\delta)/m_k}$. For every $r \geq s$, the top-r projector $\widehat{P}_k$ satisfies $\varepsilon_k = \|(I-\widehat{P}_k)P_k^\star\|_{\mathrm{op}} \leq \|\widehat{P}_k^{(s)} - P_k^\star\|_{\mathrm{op}}$, so (7) bounds ε_k as well. Here C is the constant of Theorem B.8, and the centering $\widehat{\sigma}^2$ enters only through C_{sub}.

Proof. The bias $\widetilde{B}$ is a multiple of I_d (Lemma B.6). So $\bar{M}_k^{\mathrm{probe}} + \widetilde{B}$ has the same eigenvectors as $\bar{M}_k^{\mathrm{probe}}$. Its top-$s$ eigenspace is S_k, and the gap between its s-th and $(s{+}1)$-th eigenvalues is the same, at least $\lambda_{\min}$ by Condition (E). Write $\Delta := \widehat{M}_k - \bar{M}_k^{\mathrm{probe}} - \widetilde{B}$. Two projectors of equal rank are at operator distance at most 1, and Davis–Kahan (Bhatia, 2013, Thm. VII.3.1) with $\bar{M}_k^{\mathrm{probe}} + \widetilde{B}$ as reference gives $\|\widehat{P}_k^{(s)} - P_k^\star\|_{\mathrm{op}} \leq 4\|\Delta\|_{\mathrm{op}}/\lambda_{\min}$ whenever $\|\Delta\|_{\mathrm{op}} \leq \lambda_{\min}/2$. Theorem B.8 gives $\|\Delta\|_{\mathrm{op}} \leq \tau(m_k,\delta)$. If $4\tau(m_k,\delta)/\lambda_{\min} \geq 1$ the bound holds trivially, and otherwise $\tau(m_k,\delta) < \lambda_{\min}/4$ and the Davis–Kahan bound applies. Since $m_k \leq T$ gives $1/\sqrt{T} \leq \sqrt{d\log(4d/\delta)/m_k}$,

$$\frac{4\tau(m_k,\delta)}{\lambda_{\min}} \leq \frac{4\sqrt{2}\,\tilde{R}_s + 4C(S_w^2+\sigma_\varepsilon^2+\widehat{\sigma}^2)}{\lambda_{\min}}\sqrt{\frac{d\log(4d/\delta)}{m_k}} + \frac{16\tilde{R}_s}{3\lambda_{\min}}\,\frac{d\log(4d/\delta)}{m_k},$$

and both coefficients are at most C_{sub}. For $r \geq s$ the range of $\widehat{P}_k$ contains that of $\widehat{P}_k^{(s)}$. Hence $\|(I-\widehat{P}_k)P_k^\star\|_{\mathrm{op}} \leq \|(I-\widehat{P}_k^{(s)})P_k^\star\|_{\mathrm{op}}$, which equals $\|\widehat{P}_k^{(s)} - P_k^\star\|_{\mathrm{op}}$ for two projectors of equal rank. □

This chain (Lemmas B.4–B.7, Theorem B.8, Corollary B.10, Proposition B.9) gives a self-contained subspace-recovery guarantee for SPSC. It needs no knowledge of the noise variance. SPSC takes any fixed $\widehat{\sigma}^2$. Every such centering gives the same population eigenspace and changes only $\tilde{R}_s$.

General probe laws. The argument uses three properties of the scaled sphere. First, $\mathcal{K}$ is invertible on symmetric matrices. Second, the law is centrally symmetric, $u \mapsto -u$, which together with probe independence cancels the cubic cross-moment term. Third, $\mathcal{K}^{-1}(\mathbb{E}[uu^\top])$ is a multiple of I_d, so an unknown variance does not rotate the eigenvectors. Any probe law with these three properties gives the same conclusion, with constants that depend on $\|\mathcal{K}^{-1}\|_{\mathrm{op}\to\mathrm{op}}$ and the norm of the probes. A known centrally symmetric design inside every action set whose second and fourth moments match those of the scaled sphere is one example. Probes that merely span $\mathbb{R}^d$ are not enough. Probes along the coordinate axes see only the diagonal of $\widetilde{M}_t$.

B.3 Necessity of the probe-side conditions

This section proves Theorem 2.1(ii). Props. B.11–B.13 each build two instances of the model of §2 whose effective subspaces differ but whose probe observations have the same law. Both instances have rank r, stable latent dynamics, a bounded latent state, and sub-Gaussian noise with conditional mean zero, so the obstructions hold inside the model class. Prop. B.11 matches the second moments with bounded states and the full law with Gaussian states, which leave the class only through the norm bound $\|w_t\| \leq S_w$. Coverage (Prop. B.13) is needed for every probe law. Props. B.11 and B.12 let the noise variance or the state–noise cross-moment $\mathbb{E}[\varepsilon_t\theta_t \mid \mathcal{H}_{t-1}, u]$ depend on the probe direction u, and they show that identification can fail even with full coverage. Knowing the variance removes the first obstruction, and a bound $\epsilon_\times$ on the cross-moment limits the second (the sensitivity paragraph of App. B.2). Under probe independence neither obstruction arises (Theorem 2.1(i)).

The rotated twin. Fix a segment with orthonormal basis $B = [b_1,\dots,b_r]$ and a unit vector $e \perp \mathrm{range}(B)$, which exists since $r < d$. For $\varphi \in (0,\pi/2)$ let G_φ rotate the plane $\mathrm{span}\{b_1, e\}$ by the angle φ and fix its orthogonal complement. Then $G_\varphi B$ is orthonormal, $G_\varphi b_1 = \cos\varphi\, b_1 + \sin\varphi\, e$,

$G_\varphi b_j = b_j$ for $j \geq 2$, and the two ranges satisfy $\|P_{\mathrm{range}(G_\varphi B)} - P_{\mathrm{range}(B)}\|_{\mathrm{op}} = \sin\varphi > 0$. Write $M := B\Sigma B^\top$ for a latent covariance $\Sigma \succ 0$ and $M_\varphi := G_\varphi M G_\varphi^\top$. Since $\|G_\varphi - I\|_{\mathrm{op}} = 2\sin(\varphi/2)$, $\|M_\varphi - M\|_{\mathrm{op}} \leq 4\sin(\varphi/2)\|M\|_{\mathrm{op}}$ and $\|(I-G_\varphi)M(I-G_\varphi)^\top\|_{\mathrm{op}} \leq 4\sin^2(\varphi/2)\|M\|_{\mathrm{op}}$.

Proposition B.11 (Non-identification without known noise variance)**.** *Suppose $\epsilon_\times = 0$ (the noise is independent of the state) and probe coverage is full, but the conditional noise variance $v_t(u) := \mathbb{E}[\varepsilon_t^2 \mid \mathcal{H}_{t-1}, u]$ may depend on u. Then the effective subspace is not identifiable from the observable $g_t(u) := \mathbb{E}[y_t^2 \mid \mathcal{H}_{t-1}, u]$, and in particular not by any method that uses the probe responses through their second moments, such as the lifted estimator. If the variance may also depend on the played action, two such instances with Gaussian latent states give every response the same conditional law, so no method identifies it.*

Proof. Let the first instance have basis B, i.i.d. zero-mean latent states w_t with covariance $\Sigma \succ 0$ and $\|w_t\| \leq S_w$ (latent dynamics $A_k = 0$, which are stable), and Gaussian noise of variance σ_ε^2 independent of everything else. Then $\widetilde{M}_t = M$ and $g_t(u) = u^\top M u + \sigma_\varepsilon^2$. Let the second instance have basis $G_\varphi B$, the same latent states, so $\widetilde{M}'_t = M_\varphi$ with the same nonzero eigenvalues, and Gaussian noise independent of everything else with variance $v'_t(u) := \sigma_\varepsilon^2 - u^\top(M_\varphi - M)u$. For $\|u\|^2 = d$, $|u^\top(M_\varphi - M)u| \leq 4d\sin(\varphi/2)\|M\|_{\mathrm{op}} \leq \sigma_\varepsilon^2/2$ once $\sin(\varphi/2) \leq \sigma_\varepsilon^2/(8d\|M\|_{\mathrm{op}})$, so $v'_t(u) \in [\sigma_\varepsilon^2/2, 3\sigma_\varepsilon^2/2]$ is a valid conditional variance that depends on u. Then $g'_t(u) = u^\top M_\varphi u + v'_t(u) = u^\top M u + \sigma_\varepsilon^2 = g_t(u)$ for every u, while the two effective subspaces are at distance $\sin\varphi$.

Every method. Fix scaled-sphere probes and exogenous action sets, the same in both instances. Let both instances have i.i.d. latent states $w_t \sim \mathcal{N}(0, \Sigma)$ in place of bounded ones, and let the second instance's noise be $\varepsilon'_t = \sqrt{v'(x_t)}\,\zeta_t$ with $\zeta_t \sim \mathcal{N}(0,1)$ independent of everything else and $v'(x) := \sigma_\varepsilon^2 - x^\top(M_\varphi - M)x$ for every played vector x, probe or action. Since $\|u\|^2 = d$ and $\|x\| \leq R_{\mathcal{A}}$, $v'(x) \in [\sigma_\varepsilon^2/2, 3\sigma_\varepsilon^2/2]$ once $\sin(\varphi/2) \leq \sigma_\varepsilon^2/(8\max\{d, R_{\mathcal{A}}^2\}\|M\|_{\mathrm{op}})$, and both noises are $\sqrt{3/2}\,\sigma_\varepsilon$-sub-Gaussian. Given the history, the action set and the played vector x, the response is $\mathcal{N}(0, x^\top M x + \sigma_\varepsilon^2)$ in the first instance and $\mathcal{N}(0, x^\top M_\varphi x + v'(x)) = \mathcal{N}(0, x^\top M x + \sigma_\varepsilon^2)$ in the second, since the state and the noise of round t are fresh. The learner's choices, probe times and directions included, depend only on the history, the action set and its own randomness, so by induction over rounds the law of the whole interaction is the same in both instances. Any estimate of the subspace then has the same law in both, while the two subspaces are $\sin\varphi$ apart, so no method identifies the subspace. Gaussian states are sub-Gaussian and leave the model of §2 only through the norm bounds S_w and S_Δ. □

Proposition B.12 (Non-identification without bounded state-noise coupling)**.** *Suppose σ_ε^2 is known, $\mathbb{E}[\varepsilon_t^2 \mid \mathcal{H}_{t-1}, u] = \sigma_\varepsilon^2$ for every u, and probe coverage is full, but the cross-moment $m_t(u) := \mathbb{E}[\theta_t \varepsilon_t \mid \mathcal{H}_{t-1}, u]$ is unrestricted (no bound $\epsilon_\times$). Then the effective subspace is not identifiable. The two instances below produce the same responses on every path, so no estimator identifies it.*

Proof. Take the latent states of Prop. B.11, let $\theta_t = B w_t$ and $\theta'_t = G_\varphi B w_t$, and put $\Delta_t := \theta_t - \theta'_t = (I - G_\varphi)B w_t$ and $D := \mathbb{E}[\Delta_t \Delta_t^\top] = (I-G_\varphi)M(I-G_\varphi)^\top$. Let ξ_t be Gaussian, independent of everything else, with variance $\sigma_\varepsilon^2 - \frac{1}{4}u^\top D u$, which is at least $\sigma_\varepsilon^2/2$ once $\sin^2(\varphi/2) \leq \sigma_\varepsilon^2/(2d\|M\|_{\mathrm{op}})$. Set $\varepsilon_t := \xi_t - \frac{1}{2}u_t^\top \Delta_t$ and $\varepsilon'_t := \xi_t + \frac{1}{2}u_t^\top \Delta_t$. Then $u_t^\top \theta_t + \varepsilon_t = \frac{1}{2}u_t^\top(\theta_t + \theta'_t) + \xi_t = u_t^\top \theta'_t + \varepsilon'_t$ on every path. Both noises are sub-Gaussian, since Δ_t is bounded, and have conditional mean zero, since $\mathbb{E}[w_t \mid \mathcal{H}_{t-1}] = 0$. Both have conditional variance $\sigma_\varepsilon^2 - \frac{1}{4}u^\top D u + \frac{1}{4}u^\top D u = \sigma_\varepsilon^2$. Their cross-moments $\mp\frac{1}{2}\mathbb{E}[\theta_t \Delta_t^\top]u$ are nonzero and depend on u. The two effective subspaces are at distance $\sin\varphi$. □

Proposition B.13 (Non-identification without full-dimensional coverage)**.** *Suppose the probes u_t lie in a proper subspace $U \subsetneq \mathbb{R}^d$ almost surely. Then the effective subspace is not identifiable from the responses, linear or quadratic, whatever the latent dynamics and the noise.*

Proof. Pick a unit vector $e \in U^\perp$ and an instance whose basis B has $e \perp \mathrm{range}(B)$. Let the twin have basis $G_\varphi B$ and latent state $w'_t := D_\varphi w_t$ with $D_\varphi := \mathrm{diag}(1/\cos\varphi, 1, \ldots, 1)$, and the same noise. Then $\theta'_t = G_\varphi B D_\varphi w_t = \theta_t + \tan\varphi\, w_{t,1} e$, so $u^\top \theta'_t = u^\top \theta_t$ for every $u \in U$ and the responses coincide on every path. The twin's latent dynamics $D_\varphi A_k D_\varphi^{-1}$ have the spectrum of A_k,

its innovation covariance is $D_\varphi \Sigma_{\eta,k} D_\varphi$, and $\|w'_t\| \le S_w/\cos\varphi$, so the twin lies in the model with S_w scaled by $1/\cos\varphi$. The two effective subspaces are at distance $\sin\varphi$. □

Remark B.14 (Scope of the three constructions)**.** *Each construction lies in the model of §2, fails exactly one condition, and satisfies the other two. Prop. B.11 fails known variance, Prop. B.12 bounded coupling, and Prop. B.13 coverage. So when the noise may depend on the probe direction, no two of the conditions suffice. Under probe independence Props. B.11 and B.12 do not apply. Coverage is then necessary, and isotropic probes suffice (Theorem 2.1).*

B.4 Lemmas for the interleaved analysis

Lemma B.15 (Projected UCB with an estimated basis)**.** *Fix a segment $\mathcal{I}_k$. For each exploit round $t \in E_k$, define $U_t := \widehat{U}_{t-1}$, $P_t := U_t U_t^\top$, $z_t(x) := U_t^\top x$, $a_t^\star := U_t^\top \mu_t$, and (as in Alg. 1) $\widetilde{V}_t = \lambda I_r + \sum_{s\in\mathcal{W}_t} z_t(x_s) z_t(x_s)^\top$ and $\widehat{a}_t = \widetilde{V}_t^{-1} \sum_{s\in\mathcal{W}_t} z_t(x_s) y_s$. Assume on segment $\mathcal{I}_k$ that (i) the self-normalized event of Lem. B.3 holds uniformly over E_k, (ii) $\|\mu_t\| \le S_w$ and $\mu_t \in S_k$ for $t \in \mathcal{I}_k$, and $\|x\| \le R_{\mathcal{A}}$ for $x \in \mathcal{A}_t$, and (iii) (WC) holds, $W \ge 2n_0+2$, and probes are evenly spaced at a rate of at most one per two rounds. Let $V_{k,t}(W) := \sum_{s,s+1\in\mathcal{I}_k,\, \min\mathcal{W}_t \le s < t} \|\mu_{s+1} - \mu_s\|_2$ be the variation over the span of the window, $C_W := R_{\mathcal{A}}/\sqrt{\rho_z}$, and*

$$\delta_{k,t} := \Big(\frac{R_{\mathcal{A}}^2}{\sqrt{\rho_z}} + R_{\mathcal{A}}\Big) S_w\, \varepsilon_{k,t-1}. \tag{8}$$

Then on the same event, for every $t \in E_k$ with $|\mathcal{W}_t| \ge n_0$ and every $x \in \mathcal{A}_t$,

$$|x^\top \mu_t - z_t(x)^\top \widehat{a}_t| \le \beta_t^{(r,W)} \omega_t(x) + \delta_{k,t} + C_W R_{\mathcal{A}} V_{k,t}(W), \tag{9}$$

and

$$\begin{aligned}\sum_{t\in E_k} \big(\max_{x\in\mathcal{A}_t} x^\top \mu_t - x_t^\top \mu_t\big) &\le 2\big(\beta^{(r,W)} + R_{\mathcal{A}} S_w\big)\sqrt{n_k\,(R_{\mathcal{A}}^2/\rho_z)\big(1 + \ln n_k + 4n_k/W\big)} \\ &\quad + 2\Big(\frac{R_{\mathcal{A}}^2}{\sqrt{\rho_z}} + R_{\mathcal{A}}\Big) S_w \sum_{t\in E_k} \varepsilon_{k,t-1} \\ &\quad + 4R_{\mathcal{A}} S_w n_0 + 2C_W R_{\mathcal{A}}(W+1) V_k^{\mathrm{in}}.\end{aligned} \tag{10}$$

Proof. Fix $t \in E_k$ with $|\mathcal{W}_t| \ge n_0$ and decompose, for $x \in \mathcal{A}_t$, $x^\top \mu_t = z_t(x)^\top a_t^\star + \zeta_t(x)$ with $\zeta_t(x) = x^\top (I - P_t)\mu_t$. Since $\mu_t = P_k^\star \mu_t$ by (ii),

$$|\zeta_t(x)| \le R_{\mathcal{A}} S_w\, \|(I - P_t) P_k^\star\|_{\mathrm{op}} = R_{\mathcal{A}} S_w\, \varepsilon_{k,t-1}. \tag{11}$$

Rewrite the windowed regression in basis U_t. For $s \in \mathcal{W}_t$, $y_s = z_t(x_s)^\top a_t^\star + b_{s,t} + \xi_s$ with bias $b_{s,t} = x_s^\top (I - P_t)\mu_s + x_s^\top P_t(\mu_s - \mu_t)$ and noise $\xi_s = y_s - x_s^\top \mu_s$. The first term has $|x_s^\top (I - P_t)\mu_s| \le R_{\mathcal{A}} S_w \varepsilon_{k,t-1}$, again because $\mu_s = P_k^\star \mu_s$. The second is at most $R_{\mathcal{A}} \sum_{u=s}^{t-1} \|\mu_{u+1} - \mu_u\|_2$. By definition, $\widehat{a}_t - a_t^\star = \widetilde{V}_t^{-1} \sum_{s\in\mathcal{W}_t} z_t(x_s)\xi_s + \widetilde{V}_t^{-1} \sum_{s\in\mathcal{W}_t} z_t(x_s) b_{s,t} - \lambda \widetilde{V}_t^{-1} a_t^\star$. On the self-normalized event of Lem. B.3 the noise and ridge parts have weighted norm at most $\beta_t^{(r,W)}$, so their contribution to $z_t(x)^\top(\widehat{a}_t - a_t^\star)$ is at most $\beta_t^{(r,W)} \omega_t(x)$.

Subspace bias. Let $b_t^{\mathrm{sub}} \in \mathbb{R}^{|\mathcal{W}_t|}$ have entries $x_s^\top (I - P_t)\mu_s$, so $\|b_t^{\mathrm{sub}}\|_\infty \le R_{\mathcal{A}} S_w \varepsilon_{k,t-1}$. Let Z_t be the matrix with rows $z_t(x_s)^\top$, $s \in \mathcal{W}_t$. Since $Z_t^\top Z_t \preceq \widetilde{V}_t$, $\|\widetilde{V}_t^{-1/2} Z_t^\top\|_{\mathrm{op}} \le 1$, hence $\|\sum_{s\in\mathcal{W}_t} z_t(x_s) b_{s,t}^{\mathrm{sub}}\|_{\widetilde{V}_t^{-1}} \le \|b_t^{\mathrm{sub}}\|_2 \le \sqrt{|\mathcal{W}_t|}\, R_{\mathcal{A}} S_w \varepsilon_{k,t-1}$. Therefore $|z_t(x)^\top \widetilde{V}_t^{-1} \sum_s z_t(x_s) b_{s,t}^{\mathrm{sub}}| \le \omega_t(x) \sqrt{|\mathcal{W}_t|}\, R_{\mathcal{A}} S_w \varepsilon_{k,t-1}$, and Lemma B.2 bounds $\omega_t(x)\sqrt{|\mathcal{W}_t|}$ by $R_{\mathcal{A}}/\sqrt{\rho_z}$. This is where (WC) enters. It controls the width of every action in $\mathcal{A}_t$, not only of the played one, so the bias is bounded at the optimal action too.

Drift bias. Write $a_s := U_t^\top \mu_s$ and $\Delta_u := a_u - a_{u+1}$, so $\|\Delta_u\| \le \|\mu_{u+1} - \mu_u\|$. Summing by parts, $\sum_{s\in\mathcal{W}_t} z_t(x_s) z_t(x_s)^\top (a_s - a_t) = \sum_u J_u \Delta_u$, where u runs over the span of $\mathcal{W}_t$ and $J_u := \sum_{s\in\mathcal{W}_t,\, s\le u} z_t(x_s) z_t(x_s)^\top \preceq \widetilde{V}_t$. Hence $\|\widetilde{V}_t^{-1/2} J_u^{1/2}\|_{\mathrm{op}} \le 1$, and with $\|J_u\|_{\mathrm{op}} \le |\mathcal{W}_t| R_{\mathcal{A}}^2$,

$\|\widetilde{V}_t^{-1/2} J_u\|_{\mathrm{op}} \le R_{\mathcal{A}}\sqrt{|\mathcal{W}_t|}$. For any x, $|z_t(x)^\top \widetilde{V}_t^{-1} \sum_u J_u \Delta_u| \le \omega_t(x)\, R_{\mathcal{A}}\sqrt{|\mathcal{W}_t|}\, V_{k,t}(W) \le C_W R_{\mathcal{A}} V_{k,t}(W)$ by Lemma B.2. Without a conditioning condition the familiar bound $C R_{\mathcal{A}} W V_k^{\mathrm{in}}$ does not follow from this step (Zhao et al., 2021).

Approximate optimism. The three parts and (11) give (9). Let $x_t^\star$ maximize $x^\top \mu_t$ over $\mathcal{A}_t$ and let $I_t(x) := z_t(x)^\top \widehat{a}_t + \beta_t^{(r,W)} \omega_t(x)$ be the index. By (9), $I_t(x_t^\star) \ge x_t^{\star\top} \mu_t - \delta_{k,t} - C_W R_{\mathcal{A}} V_{k,t}(W)$ and $I_t(x_t) \le x_t^\top \mu_t + 2\beta_t^{(r,W)} \omega_t(x_t) + \delta_{k,t} + C_W R_{\mathcal{A}} V_{k,t}(W)$. Since x_t maximizes the index,

$$\Delta_t := x_t^{\star\top} \mu_t - x_t^\top \mu_t \le 2\beta_t^{(r,W)} \omega_t(x_t) + 2\delta_{k,t} + 2 C_W R_{\mathcal{A}} V_{k,t}(W).$$

The drift enters only through the regret, so the index does not need $V_{k,t}(W)$.

Summation. Also $\Delta_t \le 2R_{\mathcal{A}} S_w$, which we use on the at most $2n_0$ rounds with $|\mathcal{W}_t| < n_0$ (Lemma B.2). On the other rounds, $\min\{2R_{\mathcal{A}} S_w, 2\beta_t^{(r,W)} \omega_t(x_t)\} \le 2(\beta^{(r,W)} + R_{\mathcal{A}} S_w) \min\{1, \omega_t(x_t)\}$. Cauchy–Schwarz and Lemma B.2 give $\sum_t \min\{1, \omega_t(x_t)\} \le \sqrt{n_k\,(R_{\mathcal{A}}^2/\rho_z)(1 + \ln n_k + 4n_k/W)}$, the first term of (10). The terms $2\delta_{k,t}$ give the second term, after adding the nonnegative terms of the exempt rounds. Each increment $\mu_{u+1} - \mu_u$ lies in the span of at most $W+1$ windows, so the drift terms sum to at most $2C_W R_{\mathcal{A}}(W+1) V_k^{\mathrm{in}}$. □

Lemma B.16 (Uniform prefix subspace concentration)**.** *Assume probe independence, scaled-sphere probes, and Condition (E). Let $L_T := \log(12KdT/\delta)$ and $\bar{\varepsilon}(q) := \min\{1, C_{\mathrm{sub}}(\sqrt{dL_T/q} + dL_T/q)\}$. With probability at least $1 - \delta/3$, simultaneously for all segments k and all $q \le m_k$, the top-r projector $\widehat{P}_k^{(q)}$ built from the first q probes of the segment satisfies*

$$\|(I - \widehat{P}_k^{(q)}) P_k^\star\|_{\mathrm{op}} \le \bar{\varepsilon}(q).$$

Proof. Fix a segment k and a prefix size q. By Condition (E) the prefix average of the $\widetilde{M}_{t_j}$ has range S_k and s_k-th eigenvalue at least $\lambda_{\min}$, so Corollary B.10, applied to the prefix at confidence $\delta/(3KT)$, gives the bound for the top-s_k estimate and hence for the top-r estimate, with $\log(4d \cdot 3KT/\delta) = L_T$. A union bound over the K segments and the at most T prefixes yields the simultaneous statement. □

Lemma B.17 (Cumulative interleaved subspace error)**.** *Let $\bar{\varepsilon}_{k,t-1} := \min\{1, C_{\mathrm{sub}}(\sqrt{2dL_T/q_k(t)} + 2dL_T/q_k(t))\}$, let $q^\star := C_{\mathrm{sub}}\, dL_T(1 + \ln T)$, and let $E_k^{\mathrm{burn}} := \{t \in E_k : q_k(t) < 2dL_T\}$. On the event of Lemma B.16, $\varepsilon_{k,t-1} \le \bar{\varepsilon}_{k,t-1}$ for the basis in use at every exploit round t, and for every segment k,*

$$\sum_{t \in E_k} \varepsilon_{k,t-1} \le 4\sqrt{2}\, a_0 C_{\mathrm{sub}} \sqrt{dL_T}\, \frac{\ell_k}{\sqrt{m_k}} + 2a_0 q^\star \Big\lceil \frac{\ell_k}{m_k} \Big\rceil, \tag{12}$$

$$\sum_{t \in E_k \setminus E_k^{\mathrm{burn}}} \bar{\varepsilon}_{k,t-1}^2 \le 8 a_0 C_{\mathrm{sub}}^2\, dL_T\, (1 + \ln m_k) \Big\lceil \frac{\ell_k}{m_k} \Big\rceil, \qquad |E_k^{\mathrm{burn}}| \le 2a_0\, dL_T \Big\lceil \frac{\ell_k}{m_k} \Big\rceil.$$

Proof. By condition (R) the basis in use at round t is built from the first $j_k(t) \ge q_k(t)/2$ probes of the segment, so Lemma B.16 gives $\varepsilon_{k,t-1} \le \bar{\varepsilon}(j_k(t)) \le \bar{\varepsilon}_{k,t-1}$. Group exploit rounds by $j = q_k(t)$. By the balanced-schedule condition (B) each group has at most $a_0 \lceil \ell_k/m_k \rceil$ exploit rounds. Since $\bar{\varepsilon}_{k,t-1} \le \sqrt{2}\, C_{\mathrm{sub}} \sqrt{dL_T/j} + 2C_{\mathrm{sub}} dL_T/j$, the bounds $\sum_{j=1}^{m_k} j^{-1/2} \le 2\sqrt{m_k}$, $\sum_{j=1}^{m_k} j^{-1} \le 1 + \ln m_k \le 1 + \ln T$ and $\lceil \ell_k/m_k \rceil \le 2\ell_k/m_k$ give (12). At most $2dL_T$ groups have $j < 2dL_T$, which bounds $|E_k^{\mathrm{burn}}|$. For $j \ge 2dL_T$, $2dL_T/j \le \sqrt{2dL_T/j}$, so $\bar{\varepsilon}_{k,t-1} \le 2\sqrt{2}\, C_{\mathrm{sub}} \sqrt{dL_T/j}$ and $\bar{\varepsilon}_{k,t-1}^2 \le 8 C_{\mathrm{sub}}^2 dL_T/j$, which gives the second bound. □

Lemma B.18 (Probe cost)**.** *Let $\mathcal{F}_{t-1}$ contain the history and the latent path up to round $t-1$, so that μ_t is $\mathcal{F}_{t-1}$-measurable, and let the decision to probe at round t be $\mathcal{F}_{t-1}$-measurable. Assume probe independence. There is a universal constant C such that, with probability at least $1 - \delta$, simultaneously for every interval I of rounds that contains m_I probes,*

$$\sum_{t \in \mathcal{T}_{\mathrm{probe}} \cap I} \Big(\max_{x \in \mathcal{A}_t} x^\top \mu_t - u_t^\top \mu_t + c \Big) \le m_I\, (R_{\mathcal{A}} S_w + c) + S_w \sqrt{2C\, m_I \log(2T^2/\delta)}.$$

Proof. Since $|x^\top \mu_t| \le R_{\mathcal{A}} S_w$ for $x \in \mathcal{A}_t$, it suffices to bound $-\sum u_t^\top \mu_t$. Let $t_1 < t_2 < \cdots$ be the probe rounds and $Y_j := u_{t_j}^\top \mu_{t_j}$. By probe independence u_{t_j} is a scaled-sphere draw independent of $\mathcal{F}_{t_j-1}$, while μ_{t_j} is fixed given it. For a unit vector e and v uniform on the sphere, $v^\top e$ is sub-Gaussian with variance proxy C/d for a universal C. Hence $\mathbb{E}[e^{\lambda Y_j} \mid \mathcal{F}_{t_j-1}] \le e^{C\lambda^2 \|\mu_{t_j}\|^2/2} \le e^{C\lambda^2 S_w^2/2}$. The Chernoff bound for martingales with conditionally sub-Gaussian increments gives, for each fixed pair $a \le b$ of probe indices, $|\sum_{j=a}^{b} Y_j| \le S_w \sqrt{2C(b-a+1)\log(2T^2/\delta)}$ with probability at least $1 - \delta/T^2$. Every interval of rounds contains a contiguous block of probe indices, and there are at most T^2 such blocks. A union bound finishes the proof. □

A per-probe tail bound, $|u_t^\top \mu_t| \le S_w \sqrt{2\log(8Td^4/\delta)}$ by the tail argument of Lemma B.7, gives the same linear term up to a factor $\sqrt{\log(Td/\delta)}$. Lemma B.18 removes that factor. Neither bound introduces a factor of d.

B.5 Proof of Theorem 4.1 and Corollary 4.2

Both results follow from one bound that keeps every term. Let $A := R_{\mathcal{A}} S_w + c$, $D := (R_{\mathcal{A}}^2/\sqrt{\rho_z} + R_{\mathcal{A}}) S_w$, and

$$B'' := 8\sqrt{2}\, a_0 D\, C_{\mathrm{sub}} \sqrt{d \log(12KdT/\delta)}. \tag{13}$$

Since $R_{\mathcal{A}}^2/\sqrt{\rho_z} = R_{\mathcal{A}}\sqrt{r/\kappa_z}$, up to logarithms B'' is proportional to $R_{\mathcal{A}} S_w C_{\mathrm{sub}} \sqrt{rd/\kappa_z}$, and C_{sub} of Corollary B.10 is free of d. The probe budget is $m_k := \min\{\ell_k, \lceil c_0 \ell_k^{2/3} \rceil\}$ with $c_0 := (B''/2A)^{2/3}$, the probes evenly spaced from the first round of the segment. A segment with $\lceil c_0 \ell_k^{2/3} \rceil > \ell_k/2$ has $\ell_k < 8c_0^3 + 8$, and we charge it $2A$ per round.

Theorem B.19 (Full bound)**.** *Assume scaled-sphere probes, probe independence, the boundedness assumptions of §2, Condition (E), condition (WC), the balanced schedule* (B)*, and either $A_k = 0$ with the noise and innovation of each round independent of earlier rounds, or the variant of Lemma B.3 that empties the window at each refresh. Run Algorithm 1 with the given boundaries, the radius $\beta_t^{(r,W)}$, any refresh schedule satisfying (R), any window $W \ge 2n_0 + 2$, and the budget m_k above. Then with probability at least $1 - \delta$,*

$$\begin{aligned} \mathrm{DynReg}_T^{(c)} \le\ & \widetilde{\mathcal{O}}\Big(r\sqrt{KT/\kappa_z} + rT/\sqrt{\kappa_z W}\Big) + \Xi_0\, K^{1/3} T^{2/3} + 2C_W R_{\mathcal{A}} (W+1) V_{\mathrm{in}} \\ & + 4a_0 D q^\star \big(K^{2/3} T^{1/3}/c_0 + K\big) + AK(16c_0^3 + 17) \\ & + 4R_{\mathcal{A}} S_w n_0 K + S_w \sqrt{2CK(c_0 K^{1/3} T^{2/3} + K)\log(2T^2/\delta)}, \end{aligned} \tag{14}$$

where $\Xi_0 := 3 \cdot 2^{-2/3} A^{1/3} B''^{2/3} = (rd/\kappa_z)^{1/3}\, \Xi$ with $\Xi := 3 \cdot 2^{-2/3} A^{1/3} (B'' \sqrt{\kappa_z/(rd)})^{2/3}$. Up to logarithms, Ξ is free of d, r, κ_z and W, and scales as $\lambda_{\min}^{-2/3}$.

Proof. Fix a segment $\mathcal{I}_k$ with $m_k = \lceil c_0 \ell_k^{2/3} \rceil \le \ell_k/2$.

Probe regret. By Lemma B.18 the probe rounds of $\mathcal{I}_k$ cost at most $Am_k + S_w\sqrt{2Cm_k \log(2T^2/\delta)}$.

Exploitation regret. Lemma B.17 bounds the sum of $\varepsilon_{k,t-1}$ in (10). With $\lceil \ell_k/m_k \rceil \le \ell_k/m_k + 1$, the second term of (10) is at most $B''\ell_k/\sqrt{m_k} + 4a_0 D q^\star(\ell_k/m_k + 1)$. The first term is $\widetilde{\mathcal{O}}(r\sqrt{n_k(1+\ell_k/W)/\kappa_z})$, since $\beta^{(r,W)} = \widetilde{\mathcal{O}}(\sqrt{r})$, $R_{\mathcal{A}}^2/\rho_z = r/\kappa_z$ and $n_k \le \ell_k$.

Per-segment bound. Combining,

$$\begin{aligned} \mathrm{DynReg}^{(c)}(\mathcal{I}_k) \le\ & \widetilde{\mathcal{O}}\big(r\sqrt{n_k(1+\ell_k/W)/\kappa_z}\big) + Am_k + B''\frac{\ell_k}{\sqrt{m_k}} + 4a_0 D q^\star\big(\ell_k/m_k + 1\big) \\ & + 4R_{\mathcal{A}} S_w n_0 + S_w\sqrt{2Cm_k \log(2T^2/\delta)} + 2C_W R_{\mathcal{A}} (W+1) V_k^{\mathrm{in}}. \end{aligned} \tag{15}$$

Probe budget. With $m_k = \lceil c_0 \ell_k^{2/3} \rceil$, $Am_k + B''\ell_k/\sqrt{m_k} \le (Ac_0 + B''/\sqrt{c_0})\ell_k^{2/3} + A = \Xi_0 \ell_k^{2/3} + A$, and c_0 minimizes $Ac_0 + B''/\sqrt{c_0}$. Also $\ell_k/m_k \le \ell_k^{1/3}/c_0$. Hence

$$\begin{aligned}\mathrm{DynReg}^{(c)}(\mathcal{I}_k) \le \widetilde{\mathcal{O}}\big(r\sqrt{n_k(1+\ell_k/W)/\kappa_z}\big) + \Xi_0 \ell_k^{2/3} + A + 4a_0 D q^\star \big(\ell_k^{1/3}/c_0 + 1\big) \\ + 4R_{\mathcal{A}} S_w n_0 + S_w\sqrt{2Cm_k \log(2T^2/\delta)} + 2C_W R_{\mathcal{A}}(W+1)V_k^{\mathrm{in}}.\end{aligned} \tag{16}$$

A segment with $\lceil c_0 \ell_k^{2/3} \rceil > \ell_k/2$ costs at most $2A\ell_k \le 2A(8c_0^3 + 8)$.

Sum over segments. By Cauchy–Schwarz, $\sum_k r\sqrt{n_k(1+\ell_k/W)} \le r\sqrt{KT} + rT/\sqrt{W}$ and $\sum_k \sqrt{m_k} \le \sqrt{K\sum_k m_k} \le \sqrt{K(c_0 K^{1/3}T^{2/3} + K)}$. By Hölder, $\sum_k \ell_k^{2/3} \le K^{1/3}T^{2/3}$ and $\sum_k \ell_k^{1/3} \le K^{2/3}T^{1/3}$, and $\sum_k V_k^{\mathrm{in}} = V_{\mathrm{in}}$. Summing (16), with the short segments charged $2A(8c_0^3 + 8)$ each, gives (14). The events of Lemmas B.3, B.7, B.16 and B.18 hold together with probability at least $1 - \delta$ after rescaling δ. □

The threshold T_0. Let T_0 be the smallest T for which the last two lines of (14) together are at most $\Xi_0 K^{1/3}T^{2/3}$, that is,

$$\begin{aligned}4a_0 D q^\star\big(K^{2/3}T^{1/3}/c_0 + K\big) + AK(16c_0^3 + 17) + 4R_{\mathcal{A}} S_w n_0 K \\ + S_w\sqrt{2CK(c_0 K^{1/3}T^{2/3} + K)\log(2T^2/\delta)} \le \Xi_0 K^{1/3}T^{2/3}.\end{aligned} \tag{17}$$

The left side grows like $K^{2/3}T^{1/3}$ up to a logarithm and the right side like $K^{1/3}T^{2/3}$, so the inequality holds for all large T. Its leading part, $4a_0 D q^\star K^{2/3}T^{1/3}/c_0 \le \frac{1}{2}\Xi_0 K^{1/3}T^{2/3}$, reads $T \ge C\, a_0^3 D^3 A\, q^{\star 3} K/B''^4$ for an absolute constant C. Since $q^\star \propto dC_{\mathrm{sub}}$ and $B'' \propto DC_{\mathrm{sub}}\sqrt{d}$ up to logarithms, this part of T_0 grows linearly in K and in d up to logarithms and does not grow with C_{sub}. The part that comes from the short-segment charge $AK(16c_0^3 + 17)$, with $c_0^3 = (B''/2A)^2$, is of order KB''^2/A^2, so it grows linearly in K and in d and as $C_{\mathrm{sub}}^2 \propto \lambda_{\min}^{-2}$.

Proof of Theorem 4.1. Under the constant-mean model Condition (E) holds (App. B.1), and μ_t is constant within each segment, so $V_{\mathrm{in}} = 0$. With $W \ge \max_k \ell_k$, $\ell_k/W \le 1$ and $\sum_k r\sqrt{n_k(1+\ell_k/W)} \le \sqrt{2}\, r\sqrt{KT}$. For $T \ge T_0$ the last two lines of (14) are at most $\Xi_0 K^{1/3}T^{2/3}$. Hence $\mathrm{DynReg}_T^{(c)} \le \widetilde{\mathcal{O}}(r\sqrt{KT/\kappa_z}) + 2(rd/\kappa_z)^{1/3}\Xi K^{1/3}T^{2/3}$. A window $W \ge \kappa \max_k \ell_k$ with $\kappa \in (0, 1]$ changes only the constant, since then $\ell_k/W \le 1/\kappa$. □

Proof of Corollary 4.2. Under full excitation Condition (E) holds by Lemma B.1, with $S_k = \mathrm{range}(B_k^\star)$. Under (WC), $C_W = R_{\mathcal{A}}/\sqrt{\rho_z}$. Take $W^\star := \min\{\max_k \ell_k, (rT/(C_W R_{\mathcal{A}} V_{\mathrm{in}}))^{2/3}\}$. If the minimum is the second term, $rT/\sqrt{W^\star} + C_W R_{\mathcal{A}} W^\star V_{\mathrm{in}} = 2\, r^{2/3}(C_W R_{\mathcal{A}} V_{\mathrm{in}})^{1/3} T^{2/3}$. Otherwise $W^\star = \max_k \ell_k$, the window term is at most $\sqrt{2}\, r\sqrt{KT}$, and $C_W R_{\mathcal{A}} W^\star V_{\mathrm{in}} \le r^{2/3}(C_W R_{\mathcal{A}} V_{\mathrm{in}})^{1/3} T^{2/3}$. Substituting in (14) with $T \ge T_0$ proves the corollary. □

Constant probe rate. Let the probes of each segment be evenly spaced at rate p, starting with its first round, so $m_k = \lceil p\ell_k \rceil$ and (B) holds with $a_0 = 2$. Then $Am_k + B''\ell_k/\sqrt{m_k} \le A(p\ell_k + 1) + B''\sqrt{\ell_k/p}$ and $\ell_k/m_k \le 1/p$. Summing with $\sum_k \sqrt{\ell_k} \le \sqrt{KT}$ gives $ApT + B''\sqrt{KT/p} + AK$ in place of $\Xi_0 K^{1/3}T^{2/3} + AK$, and $4a_0 D q^\star K(1/p + 1)$ in place of the burn-in term. At $p^\star := \min\{1, (B''/2A)^{2/3}(K/T)^{1/3}\} < 1$, the first part equals $\Xi_0 K^{1/3}T^{2/3}$ and $K/p^\star = K^{2/3}T^{1/3}/c_0$, so (14) is unchanged. The prefactor $(B''/2A)^{2/3}$ carries a factor $(rd/\kappa_z)^{1/3}$.

B.6 Proof of Proposition 4.3 (SPSC-Adaptive)

Algorithm analyzed. Algorithm 2 with the projector detector. After each reset it plays m_{relearn} probes. After that, probes are evenly spaced at rate p, restarting at every reset. Within each estimated segment it refreshes after every probe and uses the radius $\beta_t^{(r,W)}$, so Theorem B.19 applies under (WC). The detector compares the top-s projectors $\widehat{P}_t^{\mathrm{rec}}$ and $\widehat{P}_t^{\mathrm{past}}$ of the lifted-moment averages of the last W_{det} probes and of the W_{det} probes before them, counting only probes after the recovery block. It resets when $\Gamma_t := \|\widehat{P}_t^{\mathrm{rec}} - \widehat{P}_t^{\mathrm{past}}\|_{\mathrm{op}} > 2\eta_P$. Here $\eta_P := \min\{1, C_{\mathrm{sub}}(\sqrt{dL_{\mathrm{FA}}/W_{\mathrm{det}}} +$

$dL_{\mathrm{FA}}/W_{\mathrm{det}})\}$ with $L_{\mathrm{FA}} := \log(8dT/\delta_{\mathrm{FA}})$ is the projector radius of Corollary B.10 for one window at confidence $\delta_{\mathrm{FA}}/(2T)$.

(i) No false alarm. Fix a probe round whose two windows lie in one segment k. Under the constant mean of Theorem 4.1, $\widetilde{M}_t$ is the same in every round of the segment, so each window's average of $\widetilde{M}_t$ has range S_k and s-th eigenvalue at least $\lambda_{\min}$. Corollary B.10, applied to each window, gives $\|\widehat{P}_t^{\bullet} - P_k^{\star}\|_{\mathrm{op}} \le \eta_P$ with probability at least $1 - \delta_{\mathrm{FA}}/(2T)$ each, so $\Gamma_t \le 2\eta_P$. A union bound over $t \le T$ gives no false alarm with probability at least $1 - \delta_{\mathrm{FA}}$.

(ii) Detection delay. Assume every segment has $\ell_k \ge m_{\mathrm{relearn}} + 3W_{\mathrm{det}}/p$ and that the previous change was detected within $W_{\mathrm{det}}/p + 1$ rounds. Then before τ_{k-1}, the part of segment $k-1$ after its recovery block holds at least $2W_{\mathrm{det}}$ probes. So the detector is armed and both windows lie in segment $k-1$. After W_{det} more probes, that is within $D_k \le W_{\mathrm{det}}/p + 1$ rounds, the recent window lies in segment k and the past window in segment $k-1$. The bounds of (i) then give $\Gamma_t \ge \Delta_k - 2\eta_P > 2\eta_P$, where $\Delta_k := \|P_k^{\star} - P_{k-1}^{\star}\|_{\mathrm{op}} > 4\eta_P$, so the detector fires. By (i) it does not fire before τ_{k-1}. By induction, every change is detected with delay $D_k \le W_{\mathrm{det}}/p + 1$.

(iii) Regret on estimated segments. On the events of (i) and (ii), the interval $[\widehat{\tau}_k, \tau_k)$ lies in segment k. On it the algorithm is Algorithm 1 started at the known time $\widehat{\tau}_k$, with m_{relearn} initial probes followed by probes at rate p. Exploitation starts after the initial probes, and at most $\lceil 1/p \rceil$ exploitation rounds share each probe count. The proof of Theorem 4.1 uses (B) only through such a count, so it applies with $\lceil 1/p \rceil$ in place of $a_0 \lceil \ell_k/m_k \rceil$, and the initial probes only shorten the burn-in. Lemmas B.3 and B.16 hold simultaneously for all start times after a union bound over at most T of them, which replaces $\log(12KdT/\delta)$ by $\log(12dT^2/\delta)$. With $\ell_k'' := \tau_k - \widehat{\tau}_k$ and at least $p\ell_k''$ probes, the per-segment bound (15) gives an acquisition part at most $A(p\ell_k'' + m_{\mathrm{relearn}} + 1) + B''\sqrt{\ell_k''/p}$ and a burn-in part at most $4a_0 Dq^{\star}(1/p+1) + 4R_{\mathcal{A}} S_w n_0$.

(iv) Stale rounds. The rounds in $[\tau_{k-1}, \widehat{\tau}_k)$ still use the statistics of segment $k-1$. Each exploit round among them costs at most $2R_{\mathcal{A}} S_w$, and the probe rounds are covered by Lemma B.18. Together they add at most $(2R_{\mathcal{A}} S_w + A)D_k$ per change.

(v) Sum. Summing over segments with $\sum_k \sqrt{\ell_k''} \le \sqrt{KT}$, under the constant-mean model ($V_{\mathrm{in}} = 0$) and with $W \ge \max_k \ell_k$,

$$\begin{aligned}\mathrm{DynReg}_T^{(c)} \le\ & \widetilde{\mathcal{O}}(r\sqrt{KT/\kappa_z}) + ApT + B''\sqrt{KT/p} + 4a_0 Dq^{\star}K(1/p+1) + 4R_{\mathcal{A}} S_w n_0 K \\ & + AK(m_{\mathrm{relearn}} + 1) + (2R_{\mathcal{A}} S_w + A)K(W_{\mathrm{det}}/p + 1) \\ & + S_w\sqrt{2CK\big(pT + K(m_{\mathrm{relearn}} + 1)\big)\log(2T^2/\delta)}.\end{aligned} \tag{18}$$

At $p = p^{\star}$ (App. B.5) the terms $ApT + B''\sqrt{KT/p}$ equal $\Xi_0 K^{1/3}T^{2/3}$. Setting $p^{\star}$ requires K or an upper bound on it, which then takes the place of K.

The threshold T_0'. Let $T_0' \ge T_0$ be the smallest T for which, at $p = p^{\star}$, each of

$$4a_0 Dq^{\star}K(1/p^{\star} + 1) + 4R_{\mathcal{A}} S_w n_0 K, \qquad AK(m_{\mathrm{relearn}} + 1), \qquad (2R_{\mathcal{A}} S_w + A)K(W_{\mathrm{det}}/p^{\star} + 1)$$

is at most $\frac{1}{3}\Xi_0 K^{1/3}T^{2/3}$, and the last term of (18) is at most $\Xi_0 K^{1/3}T^{2/3}$. For $T \ge T_0'$, (18) gives $\mathrm{DynReg}_T^{(c)} \le \widetilde{\mathcal{O}}(r\sqrt{KT/\kappa_z}) + 3\,\Xi_0 K^{1/3}T^{2/3}$, the bound of Theorem 4.1 with larger constants, with probability at least $1 - \delta - \delta_{\mathrm{FA}}$. Since $1/p^{\star} = (2A/B'')^{2/3}(T/K)^{1/3}$, the delay condition reads approximately $KW_{\mathrm{det}}^3 A^4 \lesssim B''^4 T$, that is, $KW_{\mathrm{det}}^3 \lesssim (rd/\kappa_z)^2 T$ up to the constants in $B''\sqrt{\kappa_z/(rd)}$. Detectability needs W_{det} of order $C_{\mathrm{sub}}^2 d\log(dT)/\Delta_{\min}^2$ with $\Delta_{\min} := \min_k \Delta_k$, so this condition holds once T is linear in Kd, up to r, $\lambda_{\min}$, $\Delta_{\min}$ and logarithms. □

C PROBE DISTRIBUTIONS: SCALED SPHERE AND GAUSSIAN TRUNCATION

The analysis in §B.2 uses scaled-sphere probes $u_t = \sqrt{d}\, v_t$, $v_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, which satisfy $\|u_t\|_2 = \sqrt{d}$ exactly. The probe norm needs no truncation; the concentration argument truncates only the probe signal $u_t^\top \theta_t$ and the noise (Lemma B.7).

In our implementation, probes are generated via $u \leftarrow \sqrt{d}\,\tilde{u}/\|\tilde{u}\|$ for $\tilde{u} \sim \mathcal{N}(0, I_d)$, which is exactly the scaled-sphere distribution.

For isotropic-Gaussian probes (no rescaling), $u_t \sim \mathcal{N}(0, I_d)$, the Laurent–Massart χ^2 tail (Laurent & Massart, 2000, Lem. 1) gives $\Pr(\|u_t\| > L) \le \delta/T$ at $L = \sqrt{2d\log(2T/\delta)}$; union-bounding over $|\mathcal{T}_{\mathrm{probe}}| \le T$ gives $\Pr(\bigcap_t \{\|u_t\| \le L\}) \ge 1 - \delta$. On this event the results of §B.2 apply with $\|u_t\| \le L$ (replacing $\sqrt{d}$), changing $\tilde{R}_s$ and C_{sub} only by logarithmic factors that is absorbed in $\widetilde{\mathcal{O}}(\cdot)$.

D ALGORITHMS: PSEUDOCODE AND ADAPTIVE VARIANT

Algorithm 1 is given in Section 3. This appendix gives its computational cost and the pseudocode of SPSC-Adaptive (online change-point detection, Alg. 2), and explains the adaptive design choices (detector form, threshold, recovery phase, burn-in) that tie the tuning back to the guarantee of Proposition 4.3.

D.1 COMPUTATIONAL COST

Table 9 lists the cost of each operation of Algorithm 1 for a finite action set.

Table 9: **Cost of the operations of Algorithm 1** for a finite action set $\mathcal{A}_t$, window W and an iterative top-r eigensolver.

Operation	Time	Space
Project candidate actions, $z_t(x) = \widehat{U}_{t-1}^\top x$	$O(dr\|\mathcal{A}_t\|)$	$O(dr)$
UCB scores	$O(r^2\|\mathcal{A}_t\|)$	$O(r^2)$
Rank-r ridge update, per exploitation round	$O(r^2)$	$O(r^2)$
Window buffer of raw actions	$O(d)$ incremental	$O(Wd)$
Re-projecting the window after a refresh, once per probe	$O(Wdr)$	$O(Wr)$
Lifted probe-moment update, per probe	$O(d^2)$	$O(d^2)$
Top-r eigenspace refresh, once per probe (subspace iteration)	$O(d^2 r)$	$O(d^2)$

An exploitation round therefore costs $O((dr + r^2)|\mathcal{A}_t| + r^2)$, and a probe round $O(d^2 r + Wdr)$, against $O(d^2|\mathcal{A}_t|)$ for ambient LinUCB, and the memory is $O(d^2 + Wd)$. The window keeps raw actions so that a refresh can re-project them. The variant that empties the window at each refresh needs only $O(Wr)$. With a structured action-maximization oracle, the $|\mathcal{A}_t|$ scoring cost is replaced by that oracle's cost. SPSC-Adaptive additionally keeps two $d \times d$ detector accumulators and, to slide its windows, the last $2W_{\mathrm{det}}$ probes, $O(d^2 + W_{\mathrm{det}} d)$ memory. Restarting at the boundaries and estimating cumulatively removes the window buffer; its drift term has the segment length ℓ_k in place of W, so it matches the windowed variant when $V_k^{\mathrm{in}} = 0$ and is worse under drift. With unknown boundaries a window also limits stale data after a delayed reset, whereas cumulative statistics stay contaminated until the next restart.

D.2 SPSC-ADAPTIVE (UNKNOWN BOUNDARIES)

SPSC-Adaptive recovers segment boundaries from probe data, removing the oracle assumption of Alg. 1. It probes at a fixed rate p on a schedule that restarts after each reset.

Algorithm 2 SPSC-Adaptive: unknown boundaries via change detection

1: **Input:** probe law Q, probe rate $p \in (0, 1]$, rank r, effective rank $s \le r$, detector window W_{det}, threshold b, recovery budget m_{relearn}, W, λ, δ, centering $\widehat{\sigma}^2$, a lower bound on $\lambda_{\min}$, and an upper bound on K.
2: start an estimated segment: set `phase` $\leftarrow$ RECOVERY and clear $\widehat{M}$, m, the detector windows, and the exploitation window.
3: **for** $t = 1, \dots, T$ **do**
4: **if** `phase` = RECOVERY **then** ▷ block of probes to rebuild the subspace
5: probe: play $u_t \sim Q$, observe y_t, add $\mathcal{K}^{-1}((y_t^2 - \widehat{\sigma}^2)u_t u_t^\top)$ to $\widehat{M}$, and set $m \leftarrow m + 1$.
6: after m_{relearn} such probes, set $\widehat{U} \leftarrow \mathrm{TopEig}_r(\widehat{M}/m)$ and `phase` $\leftarrow$ NORMAL.
7: **else if** t is on the period-$1/p$ probe schedule, restarted at the end of recovery **then**
8: probe as above and set $\widehat{U} \leftarrow \mathrm{TopEig}_r(\widehat{M}/m)$.
9: add the probe to the detector windows. Once both hold W_{det} probes, compute Γ_t from their top-s projectors (Eq. 19).
10: **if** $\Gamma_t > b$ **then**
11: start a new estimated segment as in line 2.
12: **else**
13: exploit as in Alg. 1, with the radius $\beta_t^{(r,W)}$.

Detector. At each probe round the algorithm maintains two non-overlapping windows of W_{det} probe rounds, the recent and the past, counting only probes taken after the last recovery block. It forms the top-s eigenprojectors of their lifted-moment averages, where s is the effective rank, and computes

$$\Gamma_t := \left\| \widehat{P}_t^{\mathrm{rec}} - \widehat{P}_t^{\mathrm{past}} \right\|_{\mathrm{op}}. \tag{19}$$

Once both windows are full, it declares a change whenever $\Gamma_t > b$ and enters a RECOVERY phase. Proposition 4.3 analyzes $b = 2\eta_P$.

Recovery phase. Recovery deploys m_{relearn} consecutive probe rounds with no exploitation. The block of probes serves two purposes. Its probes start the estimate of the new effective subspace, which shortens the burn-in of the new segment. And by clearing stale buffers it keeps the windowed ridge estimator from carrying information from the previous regime into the new one. After m_{relearn} rounds the algorithm returns to NORMAL probe–exploitation interleaving at rate p.

Reward-drop trigger. A second trigger needs no probe windows. Let $\bar{y}_{\mathrm{rec}}$ be the mean of the last n exploitation rewards and $\bar{y}_{\mathrm{hist}}$ the mean of the earlier exploitation rewards since the last reset. Once $\tau_{\mathrm{burn}} + n$ exploitation rounds have passed since the reset, the trigger resets when $\bar{y}_{\mathrm{hist}} - \bar{y}_{\mathrm{rec}}$ exceeds a threshold. The experiments use the threshold $z_\star \widehat{\sigma}/\sqrt{n}$, with $\widehat{\sigma}$ the standard deviation of the earlier rewards, and the settings below. The next result analyzes this trigger with its threshold and window set from the drop Δ. Write $\bar{m}_{\mathrm{rec}}$ and $\bar{m}_{\mathrm{hist}}$ for the averages of the conditional mean rewards $x_t^\top \mu_t$ over the same two windows.

Proposition D.1 (Reward-drop trigger)**.** *Let the assumptions of Theorem 4.1 hold, and let σ_{eff} be the sub-Gaussian scale of the exploitation noise $\xi_t = y_t - x_t^\top \mu_t$. Fix $\Delta > 0$ and assume*

(D1) within every segment, at every armed test, $\bar{m}_{\mathrm{hist}} - \bar{m}_{\mathrm{rec}} \le \Delta/2$;

(D2) after every change, at the first armed test whose recent window lies in the new segment, $\bar{m}_{\mathrm{hist}} - \bar{m}_{\mathrm{rec}} \ge \Delta$.

Let $L := 2\sigma_{\mathrm{eff}}\sqrt{2\log(4T^2/\delta_{\mathrm{FA}})/n}$. Run SPSC-Adaptive with the reward-drop trigger at threshold $\Delta/2 + L$, with $\tau_{\mathrm{burn}} \ge n$ and $n \ge 128\,\sigma_{\mathrm{eff}}^2 \log(4T^2/\delta_{\mathrm{FA}})/\Delta^2$, at probe rate $p \le 1/2$, on segments longer than $m_{\mathrm{relearn}} + (\tau_{\mathrm{burn}} + 2n)/(1-p) + 1$ rounds. Then with probability at least $1 - \delta_{\mathrm{FA}}$ there is no false alarm, and every change is detected within $n/(1-p) + 1$ rounds. Under the remaining assumptions of Proposition 4.3, SPSC-Adaptive with this trigger attains (5) *for T above a threshold of the form of T_0', with $(2R_{\mathcal{A}}S_w + A)K(n/(1-p) + 1)$ as its delay term.*

Proof. Fix a reset time R and a test time t. The probe schedule after a reset is deterministic, so the two windows are determined by (R, t), and $\bar{y}_{\mathrm{hist}} - \bar{y}_{\mathrm{rec}} = (\bar{m}_{\mathrm{hist}} - \bar{m}_{\mathrm{rec}}) + (\bar{\xi}_{\mathrm{hist}} - \bar{\xi}_{\mathrm{rec}})$, where

$\bar{\xi}$ averages the noise over a window. Each ξ_t has conditional mean zero given the past and x_t and is σ_{eff}-sub-Gaussian, so by Azuma–Hoeffding the average over a window of w rounds is within $\sigma_{\text{eff}}\sqrt{2\log(4T^2/\delta_{\text{FA}})/w}$ of zero with probability at least $1-\delta_{\text{FA}}/(4T^2)$. The historical window has $w \geq \tau_{\text{burn}} \geq n$ rounds and the recent one n, so $|\bar{\xi}_{\text{hist}} - \bar{\xi}_{\text{rec}}| \leq L$ for both windows at once. A union bound over the at most T^2 pairs (R,t) makes this hold at every armed test with probability at least $1-\delta_{\text{FA}}$.

No false alarm. Within a segment, (D1) gives $\bar{y}_{\text{hist}} - \bar{y}_{\text{rec}} \leq \Delta/2 + L$, so the trigger does not fire.

Detection. The segment-length condition leaves, after the recovery block and the detection of the previous change, at least $(\tau_{\text{burn}}+n)/(1-p)$ rounds before the next change, so the test is armed when the change arrives. Within $n/(1-p)+1$ rounds after the change, n exploitation rounds have passed and the recent window lies in the new segment. By (D2), $\bar{y}_{\text{hist}} - \bar{y}_{\text{rec}} \geq \Delta - L > \Delta/2 + L$, since $n \geq 128\sigma_{\text{eff}}^2 \log(4T^2/\delta_{\text{FA}})/\Delta^2$ gives $2L \leq \Delta/2$. The trigger fires.

Regret. Parts (iii)–(v) of the proof of Proposition 4.3 (App. B.6) use the detector only through the no-false-alarm event and the delay bound D_k. With $D_k \leq n/(1-p)+1$ they give (18) with $(2R_{\mathcal{A}}S_w + A)K(n/(1-p)+1)$ in place of the W_{det} term, and the same threshold argument gives (5). □

Condition (D2) says that the actions a stale policy keeps playing lose at least Δ in expected reward after a change. Condition (D1) says that within a segment the recent policy is not worse than its own past by more than $\Delta/2$, which holds when the policy improves over the segment. Both are conditions on the trajectory.

Tuning used in experiments. SPSC-Adaptive uses the same probe period as Algorithm 1 in each experiment (Table 11), so its probe rate p is one over that period, not $p^\star$. The other settings are fixed across all datasets and (d,r) cells, with no per-dataset retuning. They are the reward window $n = 50$ exploitation rounds, a burn-in of $\tau_{\text{burn}} = 100$ exploitation rounds, after which the trigger is armed once a full reward window is available, $m_{\text{relearn}} = 30$, and $z_\star = 3.0$.

E EXPERIMENTAL SETUP

This appendix records how each benchmark is built and run, so that every number in the paper can be reproduced. Table 10 shows how each dataset becomes a piecewise low-rank decision stream, and Table 11 lists the dimensions and settings of every experiment. The paragraphs after the tables give the protocol shared by all runs, further details on the benchmarks, and the baselines.

In the reference setting, App. G.2 sweeps the probe period, and App. G.4 uses a probe period of 10 and 20 seeds.

Shared protocol. Every experiment gives all methods the same environment for each seed, and each method draws its own action sets. Results are means with one standard error. Probes are drawn from the scaled sphere and played outside the action set, as the probing-permitted protocol of §2 allows. On MovieLens a probe receives the segment's ridge fit plus noise, since a real rating exists only for movies, and App. F.1 repeats MovieLens with each probe answered by a real rating. The Gaussian noise is clipped at ± 1. This barely changes σ_ε=0.3 but lowers the standard deviation of the stress test's σ_ε=1 noise to about 0.72. SPSC and SW-LinUCB use the same window W, except in the small-d stress test, where SW-LinUCB keeps a window of 200 and SPSC uses 100. Every table reports the costed regret (1) with the realized θ_t in place of μ_t, and the control regret is the same sum without the fee term. Since $\mathbb{E}[\max_x x^\top \theta_t] \geq \max_x x^\top \mu_t$, this dominates (1) in expectation. When θ is fixed within a segment the two agree, except at the first round of a segment. Each probe is charged its forgone reward $\max_{x\in\mathcal{A}_t} x^\top\theta_t - u_t^\top\theta_t$ plus the fee c. On the real-data benchmarks the fee is 2% of the mean per-round reward gap $\max_{x\in\mathcal{A}_t} x^\top\theta_t - \text{mean}_{x\in\mathcal{A}_t} x^\top\theta_t$, which is 0.02 on the UCI, image and MovieLens tasks and 6×10^{-4} on Open Bandit, whose rewards are click rates of order 10^{-2}. The synthetic grid uses c=0.1, a whole round's gap, so its probes are expensive.

Implementation. The runs execute Algorithm 1 with the basis recomputed after every probe, the ridge radius $\sigma_\varepsilon\sqrt{r\log((1+nR_{\mathcal{A}}^2/\lambda)/\delta)} + \sqrt{\lambda}S_w$ with δ=0.05 and n the number of rounds in

Table 10: **Benchmark construction.** How each dataset becomes a piecewise low-rank decision stream.

Benchmark	Actions x_t	Reward y_t	What changes at a change point
Synthetic grid	40 unit-norm vectors per round, coordinate i scaled by $(i{+}1)^{-1.5}$ before normalizing	$x_t^\top \theta^{(k)}$ plus Gaussian noise (σ_ε=0.3), $\theta^{(k)}$ fixed in the segment	a new random subspace $B_k^\star$ and a new $\theta^{(k)} = B_k^\star w_k$ in it
Covertype	40 examples, standardized attributes and their pairwise products cut to $\mathbb{R}^d$, unit norm	$x_t^\top \theta_t$ plus noise, $\theta_t = B_k^\star w_t$ with an AR(1) latent state (ρ=0.99)	the subspace $B_k^\star$, built from class centroids
Pendigits, Satimage	40 examples, standardized attributes and their pairwise products cut to $\mathbb{R}^d$, unit norm	$x_t^\top \theta^{(k)}$ plus noise, $\theta^{(k)}$ a ridge fit to the centered class labels	the set of active classes
MNIST, Fashion-MNIST	40 images, standardized pixels randomly projected to $\mathbb{R}^d$	$x_t^\top \theta^{(k)}$ plus noise, $\theta^{(k)}$ a ridge fit to the centered class labels	the set of active classes
MovieLens-100K	genre indicators, SVD factors of the rating matrix, and genre-pair products, cut to $\mathbb{R}^d$	real centered rating of the played movie (a probe gets the ridge fit plus noise)	the active user cluster (k-means)
Open Bandit (ZOZOTOWN)	logged item features	linear model fitted to real clicks, plus noise	the time slice of the log
Small-d stress test	50 random unit vectors in $\mathbb{R}^2$	$x_t^\top \theta_t$ plus Gaussian noise (σ_ε=1), $\theta_t = B_k^\star w_t$ with an AR(1) latent state (ρ=0.99)	the subspace $B_k^\star$

Table 11: **Settings of every experiment.** The probe period is the number of rounds between probes. All runs use δ=0.05.

Experiment	Where	d	r	K	T	Seeds	λ	Period	W	c
Synthetic grid	Fig. 2, Tab. 12	5–100	1–20	10	5,000	10	0.01	50	400	0.1
Covertype, Pendigits, Satimage	Tabs. 2, 13–15	5, 55, 105	1, 10, 20, 30	10	5,000	10	0.01	10	400	0.02
MNIST, Fashion-MNIST, MovieLens	Tabs. 2, 16–18	55, 105, 200	5, 10, 20	10	5,000	10	0.01	10	400	0.02
Open Bandit	Tab. 19	55, 105	5, 10	10	5,000	10	0.01	10	400	6×10^{-4}
BOSS/Jedra grid	Tab. 3	55, 105, 200	5, 10, 20	10	5,000	10	0.01	50	400	0.1
Small-d stress test	Fig. 3	2	1	4	6,000	30	1.0	30	100	0.1
Reference setting	App. G.2, G.3, G.4	4	1	4	6,000	10	1.0	30	100	0.1

the window, the centering $\widehat{\sigma}^2 = \sigma_\varepsilon^2$, and probes at a fixed period, which the constant-rate remark of App. B.5 covers. SPSC-Adaptive uses the same code and resets on the reward-drop trigger of App. D. App. G.1 checks the conditions of the analysis in these runs.

Benchmark details. In the synthetic grid and the BOSS/Jedra grid, each segment draws a random orthonormal $B_k^\star$ and one latent vector $w_k \sim \mathcal{N}(0, 0.04\, I_r)$, and $\theta_t = B_k^\star w_k$ stays fixed within the segment. The dynamics parameter ρ is not used there. The small-d stress test and the reference setting instead use the AR(1) latent state $w_t = \rho w_{t-1} + \eta_{t-1}$, with ρ=0.99 and $\eta_{t-1} \sim \mathcal{N}(0, 0.04\, I_r)$, a standard deviation of 0.2 per coordinate. Covertype, Pendigits and Satimage come from the UCI repository (Dua & Graff, 2019), MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017)

from OpenML, and MovieLens-100K from GroupLens (Harper & Konstan, 2015). Covertype uses the same AR(1) state with subspaces built from class centroids. The segments have equal length. The Pendigits and Satimage features are standardized attributes and their pairwise products, cut to d coordinates and normalized. The image features are standardized pixels reduced to $\mathbb{R}^d$ by a random projection. MovieLens uses genre indicators, 50 factors from an SVD of the rating matrix with zeros for missing entries, and genre-pair products, cut to d coordinates. Its segments are k-means clusters of users. The reward of a played movie is its real centered rating, a probe receives the segment's ridge fit $x^\top \theta^{(k)}$ plus Gaussian noise at the residual scale, and regret is measured against that fit, $\max_{x \in \mathcal{A}_t} x^\top \theta^{(k)} - x_t^\top \theta^{(k)}$. Open Bandit uses the random-policy ZOZOTOWN slice of the Open Bandit Dataset (Saito et al., 2021). Its reward parameter is a per-segment ridge fit to the real clicks, so it is only approximately low-rank.

Baselines. All baselines use the same λ and δ as SPSC. D-LinUCB uses the discount factor 0.998. SW-LinUCB uses the window W, and Restart-LinUCB restarts every T/K rounds, which are the true boundaries. LinUCB, SW-LinUCB and D-LinUCB also reset at the true boundaries. LinUCB differs from Restart-LinUCB only in its confidence radius, which counts all rounds from the start rather than the rounds since the last reset. LinTS and SW-LinTS sample from $\mathcal{N}(\widehat{\theta}_t, v_t^2 V_t^{-1})$ with $v_t = \sigma_\varepsilon \sqrt{d \log((t+1)/\delta)}$. Oracle-LinUCB knows the true $B_k^\star$ and runs LinUCB in $\mathbb{R}^r$. BOSS and Jedra learn one subspace in their first 200 rounds (App. F.2). The published subspace-aware methods do not apply directly to our setting. LowOFUL (Jun et al., 2019) needs the true subspace. The algorithms of Lu et al. (2021) are either not efficiently implementable or rely on matrix-completion steps that do not fit an action-set bandit. VOFUL is a full-dimensional variance-aware method with no rank structure. We therefore give each of them its own running estimate of the subspace, from the top eigenvectors of reward-weighted outer products. LowOFUL accumulates $\sum_t (y_t x_t)(y_t x_t)^\top$ and re-estimates the top-r eigenspace every 20 rounds after a 30-round warm-up. LowRank-Reward applies the same estimator over the window W, re-estimated every round after a 50-round warm-up, and resets at the true boundaries. VOFUL weights each term by $1/(\|x_t\|^2 + \sigma_\varepsilon^2)$. These adaptations make the baselines stronger, so any remaining gap to SPSC does not come from a failure to use low-rank structure.

Implementation and compute. The experiments use NumPy and SciPy, scikit-learn to prepare the data, and pandas with the `obp` package for Open Bandit. One full (d, r) grid, with 9 cells, 10 seeds, and all methods, takes 6–20 hours on a single 16-core CPU.

F Additional results on the benchmarks

This appendix gives the full results behind §5. It reports every cell of every benchmark, the BOSS and Jedra comparison, and the small-d stress test.

F.1 Per-cell results for every dataset

Intuition. The main text shows a few cells of each benchmark. The full tables let a reader check that the claims of §5 hold on every cell.

Setup. Each table reports the mean costed regret $\pm$ SE over 10 seeds for every method, with the settings of App. E. Green shading and bold mark the cells where an SPSC variant beats every non-oracle baseline, and underlining marks the best non-oracle baseline. The methods that ignore the rank (LinUCB, SW-LinUCB, D-LinUCB, Restart-LinUCB, LinTS, and SW-LinTS) work in the ambient space, so they have the same regret in every row of a d block, except on Covertype, where r changes the instance.

Findings. An SPSC variant beats every non-oracle baseline in 51 of the 58 real-data cells (Tables 13–19). In six of the seven exceptions the winner is Restart-LinUCB, which restarts at the true change points. The seventh is a 2.9% loss to LowOFUL on Fashion-MNIST at $(d, r){=}(200, 5)$. SPSC-Adaptive has lower regret than Algorithm 1 in 41 of the 58 real-data cells, while on the synthetic grid Algorithm 1 is better in 38 of the 40 cells. The paragraphs below take the datasets one at a time.

Synthetic (d, r) grid. Table 12 gives the full grid behind Figure 2. SPSC beats LinUCB on 37 of the 40 cells, and the three losses are all at $d \leq 10$, where the probe cost outweighs the savings from the lower rank.

Table 12: **Synthetic grid (full 40 cells).** Mean costed regret across 10 seeds for 10 methods over $(d, r) \in \{5, 10, 20, 30, 45, 60, 80, 100\} \times \{1, 3, 5, 10, 15, 20\}$ with $r < d$, T=5,000, K=10, probe period 50, window W=400, λ=0.01. **Methods.** SPSC (Algorithm 1, ours), **Adpt** = SPSC-Adaptive (Algorithm 2, ours), LinUCB, **Oracle** = Oracle-LinUCB (given the true subspace), **SW-Lin** = SW-LinUCB, **D-Lin** = D-LinUCB, **Rst** = Restart-LinUCB, **LR-Rew** = LowRank-Reward, LowOFUL, and VOFUL. S/L = SPSC/LinUCB, and S/Best = SPSC/ min{non-oracle, non-Adpt baselines}. Bold marks the 37 cells where SPSC has lower regret than LinUCB (S/L < 1). The 3 cells where LinUCB wins are all in the small-d corner $d \leq 10$, where the probe cost outweighs the dimensionality savings.

d	r	SPSC	Adpt	LinUCB	Oracle	SW-Lin	D-Lin	Rst	LR-Rew	LowOFUL	VOFUL	S/L	S/Best
5	1	355	344	203	8	213	234	206	416	514	553	1.742	1.742
5	3	339	384	221	134	221	254	217	629	930	889	1.529	1.563
10	1	270	289	255	24	260	271	258	254	317	315	1.055	1.060
10	3	**389**	422	399	130	391	428	386	456	629	595	**0.976**	1.008
10	5	**465**	525	487	254	482	523	471	660	800	823	**0.955**	0.988
20	1	**231**	228	251	21	251	253	248	179	238	240	**0.923**	1.290
20	3	**353**	377	456	134	449	467	451	344	449	475	**0.774**	1.026
20	5	**420**	453	565	232	555	581	558	468	575	578	**0.744**	**0.898**
20	10	**631**	699	800	498	780	840	784	804	946	975	**0.789**	**0.809**
20	15	**805**	901	899	736	893	954	883	1011	1235	1222	**0.895**	**0.912**
30	1	**189**	206	220	24	220	222	219	154	198	203	**0.859**	1.225
30	3	**307**	330	407	125	409	413	404	296	355	355	**0.754**	1.038
30	5	**372**	421	544	217	541	553	536	411	498	492	**0.683**	**0.904**
30	10	**572**	615	775	451	763	790	760	690	855	828	**0.738**	**0.829**
30	15	**741**	810	930	663	917	957	913	956	1066	1113	**0.796**	**0.811**
30	20	**890**	960	1018	838	1002	1056	1009	1109	1236	1259	**0.874**	**0.888**
45	1	**165**	169	188	25	187	189	186	130	172	168	**0.881**	1.275
45	3	**271**	296	371	118	371	375	369	273	312	294	**0.730**	**0.992**
45	5	**360**	389	504	212	506	510	502	376	419	431	**0.715**	**0.957**
45	10	**499**	520	675	395	665	684	668	559	646	650	**0.739**	**0.893**
45	15	**651**	684	819	577	812	836	811	736	811	808	**0.796**	**0.885**
45	20	**811**	836	973	758	954	980	952	901	988	989	**0.833**	**0.900**
60	1	**141**	146	165	30	166	166	165	108	134	134	**0.858**	1.311
60	3	**229**	243	296	117	294	296	292	222	249	239	**0.776**	1.033
60	5	**315**	319	414	199	412	416	414	320	366	343	**0.760**	**0.985**
60	10	**441**	479	605	367	600	608	601	536	564	585	**0.729**	**0.822**
60	15	**588**	605	743	524	740	752	738	713	752	736	**0.791**	**0.824**
60	20	**708**	753	853	663	845	868	845	838	882	862	**0.830**	**0.845**
80	1	**110**	125	118	27	118	118	118	91	111	106	**0.931**	1.214
80	3	**231**	250	290	116	292	293	290	224	238	239	**0.796**	1.034
80	5	**279**	295	381	187	379	384	379	293	307	319	**0.733**	**0.954**
80	10	**419**	430	542	345	539	545	538	447	493	473	**0.772**	**0.937**
80	15	**534**	576	671	486	668	671	667	587	638	655	**0.796**	**0.911**
80	20	**640**	663	767	599	761	770	761	691	744	761	**0.835**	**0.927**
100	1	**107**	111	112	28	113	113	113	79	86	93	**0.954**	1.359
100	3	**192**	194	238	107	236	238	236	176	193	201	**0.809**	1.094
100	5	**272**	273	346	183	343	346	344	266	294	290	**0.787**	1.024
100	10	**368**	388	471	317	472	472	470	403	438	440	**0.780**	**0.912**
100	15	**466**	491	573	430	571	576	572	517	530	535	**0.813**	**0.902**
100	20	**609**	634	726	572	725	733	723	661	692	702	**0.839**	**0.921**

Covertype. An SPSC variant wins seven of the nine cells, and at d=105 its regret is 4–16% below LinUCB's (Table 13). The two losses are at d=55 with $r \geq 20$, where Restart-LinUCB is best.

Table 13: **Covertype** per-cell results (T=5,000, K=10). Mean ± SE costed regret over 10 seeds. Green shading and bold mark cells where an SPSC variant beats every non-oracle baseline, and underlining marks the best non-oracle baseline.

d	r	Oracle	**SPSC**	**SPSC-Adp**	LowOFUL	VOFUL	LR-Rew	SW-Lin	D-Lin	Rst-Lin	LinUCB	LinTS	SW-LinTS
5	1	2013 ±138	2660 ±85	**2135** ±44	3748 ±162	3736 ±164	2380 ±136	2651 ±110	2502 ±98	2661 ±124	2694 ±126	3669 ±153	3064 ±106
55	1	718 ±51	**1308** ±22	**1278** ±32	1428 ±34	1432 ±33	1366 ±44	1331 ±20	1346 ±21	1329 ±21	1336 ±21	1460 ±25	1447 ±26
	10	2768 ±29	**4330** ±64	**4178** ±63	5067 ±83	5167 ±130	5123 ±62	4403 ±41	4599 ±57	4348 ±37	4465 ±49	5572 ±74	5490 ±71
	20	4004 ±55	5805 ±59	5803 ±84	7120 ±93	7001 ±72	7538 ±81	5703 ±34	5971 ±40	5669 ±40	5800 ±50	7761 ±42	7561 ±47
	30	5202 ±48	6947 ±84	7030 ±105	8890 ±110	8770 ±93	9527 ±96	6653 ±34	6928 ±30	6594 ±51	6743 ±34	9626 ±87	9368 ±67
105	1	554 ±30	1035 ±29	**1025** ±21	1136 ±40	1120 ±45	1027 ±28	1075 ±32	1082 ±33	1073 ±31	1075 ±33	1103 ±34	1106 ±35
	10	2154 ±27	**3480** ±42	**3421** ±40	3775 ±34	3758 ±48	3869 ±36	3920 ±36	3959 ±36	3909 ±37	3943 ±35	4176 ±40	4149 ±40
	20	3233 ±36	**4606** ±65	**4523** ±46	5297 ±67	5325 ±68	5621 ±43	5302 ±40	5382 ±32	5313 ±30	5356 ±34	5828 ±50	5770 ±46
	30	4165 ±23	**5495** ±70	**5587** ±77	6478 ±34	6556 ±51	6993 ±53	6360 ±38	6473 ±33	6324 ±42	6400 ±37	7069 ±46	7022 ±56

Pendigits. On Pendigits SPSC wins clearly once $r \geq 10$. At d=105 and $r \geq 10$, SPSC-Adaptive's regret is 45–50% below LinUCB's (Table 14). With r=1 and $d \leq 55$ Restart-LinUCB has the lowest regret. At d=5 it has 1,090 against 2,981 for Algorithm 1.

Table 14: **Pendigits** per-cell results. T=5,000, K=10, 10 seeds.

d	r	Oracle	**SPSC**	**SPSC-Adp**	LowOFUL	VOFUL	LR-Rew	SW-Lin	D-Lin	Rst-Lin	LinUCB	LinTS	SW-LinTS
5	1	42.9 ±5.2	2981 ±171	4751 ±142	8673 ±234	8872 ±253	6557 ±67	1163 ±9	1357 ±13	1090 ±12	1149 ±9	6631 ±28	4563 ±43
55	1	18.4 ±0.5	4070 ±60	4295 ±158	5569 ±82	5774 ±53	4226 ±102	3719 ±9	3923 ±7	3483 ±10	3541 ±11	4275 ±10	4566 ±12
	10	935 ±3	**2736** ±59	**2488** ±88	3512 ±79	3515 ±53	4130 ±42	3719 ±9	3923 ±7	3483 ±10	3541 ±11	4275 ±10	4566 ±12
	20	1767 ±7	**2762** ±58	**2421** ±49	3715 ±64	3695 ±42	4177 ±28	3719 ±9	3923 ±7	3483 ±10	3541 ±11	4275 ±10	4566 ±12
	30	2408 ±3	**3004** ±63	**2649** ±60	3744 ±48	3685 ±66	4226 ±27	3719 ±9	3923 ±7	3483 ±10	3541 ±11	4275 ±10	4566 ±12
105	1	11.1 ±0.3	**3086** ±80	3305 ±108	4495 ±56	4418 ±90	3242 ±87	3333 ±7	3444 ±6	3219 ±5	3258 ±5	3215 ±11	3626 ±6
	10	673 ±3	**2150** ±58	**1797** ±60	3181 ±80	3053 ±75	3151 ±46	3333 ±7	3444 ±6	3219 ±5	3258 ±5	3215 ±11	3626 ±6
	20	1211 ±5	**2136** ±51	**1637** ±51	2897 ±74	2816 ±72	3166 ±31	3333 ±7	3444 ±6	3219 ±5	3258 ±5	3215 ±11	3626 ±6
	30	1745 ±5	**2311** ±55	**1799** ±64	3032 ±55	2987 ±57	3318 ±15	3333 ±7	3444 ±6	3219 ±5	3258 ±5	3215 ±11	3626 ±6

Satimage. Satimage gives the largest gains among the UCI datasets. At d=105 and $r \geq 10$, SPSC-Adaptive's regret is 54–67% below LinUCB's, and an SPSC variant wins every cell except (d, r)=$(5, 1)$ (Table 15).

Table 15: **Satimage** per-cell results. T=5,000, K=10, 10 seeds.

d	r	Oracle	**SPSC**	**SPSC-Adp**	LowOFUL	VOFUL	LR-Rew	SW-Lin	D-Lin	Rst-Lin	LinUCB	LinTS	SW-LinTS
5	1	34.8 ±3.1	1477 ±87	1715 ±139	2704 ±27	2714 ±25	2694 ±25	638 ±7	750 ±4	626 ±5	667 ±7	2431 ±14	1710 ±15
55	1	17.8 ±0.7	**2519** ±67	**2185** ±130	2874 ±9	2890 ±15	3636 ±65	3046 ±8	3214 ±5	2787 ±8	2854 ±8	3478 ±13	3847 ±6
	10	622 ±4	**1395** ±35	**1104** ±62	2452 ±74	2294 ±99	2795 ±64	3046 ±8	3214 ±5	2787 ±8	2854 ±8	3478 ±13	3847 ±6
	20	1185 ±8	**1541** ±29	**1275** ±36	2487 ±69	2309 ±88	2919 ±28	3046 ±8	3214 ±5	2787 ±8	2854 ±8	3478 ±13	3847 ±6
	30	1719 ±6	**1931** ±33	**1689** ±31	2730 ±62	2754 ±48	2993 ±56	3046 ±8	3214 ±5	2787 ±8	2854 ±8	3478 ±13	3847 ±6
105	1	15.7 ±0.7	**2587** ±82	**2531** ±138	3473 ±15	3495 ±13	4074 ±96	3601 ±8	3723 ±8	3491 ±7	3539 ±9	3631 ±10	3953 ±8
	10	544 ±2	**1370** ±33	**1154** ±43	2303 ±53	2305 ±85	2753 ±62	3601 ±8	3723 ±8	3491 ±7	3539 ±9	3631 ±10	3953 ±8
	20	1101 ±6	**1453** ±36	**1230** ±48	2424 ±58	2263 ±79	2718 ±39	3601 ±8	3723 ±8	3491 ±7	3539 ±9	3631 ±10	3953 ±8
	30	1575 ±6	**1800** ±34	**1630** ±35	2739 ±33	2723 ±61	2789 ±42	3601 ±8	3723 ±8	3491 ±7	3539 ±9	3631 ±10	3953 ±8

Fashion-MNIST and MNIST. On the two image datasets an SPSC variant wins 17 of the 18 cells. The regret of Algorithm 1 is 3–35% below LinUCB's on Fashion-MNIST (Table 16) and 11–30% below on MNIST (Table 17). SPSC-Adaptive is ahead of Algorithm 1 in 9 of the 18 cells and reaches 38% and 34% below LinUCB. The one exception is Fashion-MNIST at (d, r)=$(200, 5)$, where LowOFUL is ahead by 2.9%.

Table 16: **Fashion-MNIST** per-cell results. T=5,000, K=10, 10 seeds.

d	r	Oracle	**SPSC**	**SPSC-Adp**	LowOFUL	VOFUL	LR-Rew	SW-Lin	D-Lin	Rst-Lin	LinUCB	LinTS	SW-LinTS
55	5	510 ±4	**3740** ±88	3853 ±109	4377 ±106	4323 ±56	4768 ±90	4114 ±12	4289 ±12	3781 ±10	3863 ±7	5068 ±18	5261 ±10
	10	935 ±5	**3376** ±71	**3463** ±91	4382 ±74	4441 ±85	4766 ±52	4114 ±12	4289 ±12	3781 ±10	3863 ±7	5068 ±18	5261 ±10
	20	1713 ±6	**3377** ±67	**3147** ±61	4788 ±51	4570 ±69	4941 ±32	4114 ±12	4289 ±12	3781 ±10	3863 ±7	5068 ±18	5261 ±10
105	5	444 ±2	**3128** ±77	**3063** ±101	3339 ±82	3571 ±110	3780 ±53	4106 ±8	4174 ±7	3964 ±9	3993 ±9	4306 ±8	4480 ±13
	10	756 ±5	**2814** ±81	**2665** ±64	3403 ±71	3435 ±60	3736 ±46	4106 ±8	4174 ±7	3964 ±9	3993 ±9	4306 ±8	4480 ±13
	20	1380 ±4	**2745** ±69	**2520** ±52	3380 ±51	3359 ±50	3786 ±56	4106 ±8	4174 ±7	3964 ±9	3993 ±9	4306 ±8	4480 ±13
200	5	385 ±3	2939 ±41	3141 ±101	2855 ±50	2947 ±69	3214 ±54	4056 ±5	4109 ±9	4003 ±6	4035 ±6	4059 ±9	4153 ±10
	10	705 ±5	**2681** ±56	**2681** ±96	2842 ±69	2826 ±71	3080 ±35	4056 ±5	4109 ±9	4003 ±6	4035 ±6	4059 ±9	4153 ±10
	20	1268 ±7	**2625** ±45	**2495** ±69	3152 ±41	3120 ±31	3490 ±40	4056 ±5	4109 ±9	4003 ±6	4035 ±6	4059 ±9	4153 ±10

MovieLens. On MovieLens an SPSC variant beats every non-oracle baseline on all nine cells. SPSC-Adaptive's regret is 8–25% below LinUCB's and Algorithm 1's 1–17% below (Table 18), with margins that grow with d. The result also holds when every response is a real rating. With each probe answered by the real rating of the nearest movie in the round's action set, so that no

Table 17: **MNIST** per-cell results. T=5,000, K=10, 10 seeds.

d	r	Oracle	**SPSC**	**SPSC-Adp**	LowOFUL	VOFUL	LR-Rew	SW-Lin	D-Lin	Rst-Lin	LinUCB	LinTS	SW-LinTS
55	5	709 ±7	**5239** ±146	**5606** ±172	6853 ±192	6862 ±102	7104 ±97	6259 ±22	6559 ±20	5750 ±17	5857 ±12	7542 ±21	7980 ±22
	10	1267 ±8	**4765** ±92	**4943** ±114	6678 ±100	7150 ±98	7083 ±92	6259 ±22	6559 ±20	5750 ±17	5857 ±12	7542 ±21	7980 ±22
	20	2411 ±14	**4955** ±74	**4813** ±84	6736 ±115	7135 ±112	7166 ±96	6259 ±22	6559 ±20	5750 ±17	5857 ±12	7542 ±21	7980 ±22
105	5	653 ±4	**4982** ±118	**5148** ±127	5868 ±125	5636 ±100	6116 ±119	6336 ±12	6448 ±11	6121 ±13	6176 ±10	6251 ±12	6612 ±19
	10	1132 ±10	**4556** ±117	**4562** ±147	5411 ±83	5308 ±105	5904 ±75	6336 ±12	6448 ±11	6121 ±13	6176 ±10	6251 ±12	6612 ±19
	20	2041 ±11	**4503** ±104	**4382** ±74	5386 ±89	5307 ±56	5874 ±62	6336 ±12	6448 ±11	6121 ±13	6176 ±10	6251 ±12	6612 ±19
200	5	586 ±4	**4519** ±103	**4595** ±100	5022 ±90	5257 ±99	4917 ±66	6031 ±10	6069 ±12	5954 ±10	5961 ±8	5868 ±15	6056 ±11
	10	1010 ±7	**4241** ±80	**4052** ±85	4488 ±55	4426 ±59	4834 ±61	6031 ±10	6069 ±12	5954 ±10	5961 ±8	5868 ±15	6056 ±11
	20	1812 ±5	**4163** ±68	**3918** ±110	4743 ±94	4796 ±60	5401 ±69	6031 ±10	6069 ±12	5954 ±10	5961 ±8	5868 ±15	6056 ±11

response comes from the fitted model, an SPSC variant still beats every non-oracle baseline on all nine cells (10 seeds), by 2–13% against the best of them, and SPSC-Adaptive's regret is 3–18% below LinUCB's.

Table 18: **MovieLens** (real ratings, regret against the ridge fit of each segment) per-cell results. T=5,000, K=10, 10 seeds.

d	r	Oracle	**SPSC**	**SPSC-Adp**	LowOFUL	VOFUL	LR-Rew	SW-Lin	D-Lin	Rst-Lin	LinUCB	LinTS	SW-LinTS
55	5	1435 ±22	**3276** ±26	**2830** ±83	5099 ±162	4931 ±151	5527 ±65	3731 ±18	3792 ±19	3664 ±20	3695 ±18	3827 ±22	4015 ±19
	10	2224 ±26	**3458** ±37	**3098** ±74	5112 ±89	4890 ±125	5280 ±51	3731 ±18	3792 ±19	3664 ±20	3695 ±18	3827 ±22	4015 ±19
	20	2824 ±14	3666 ±33	**3417** ±38	4829 ±59	4756 ±112	4816 ±40	3731 ±18	3792 ±19	3664 ±20	3695 ±18	3827 ±22	4015 ±19
105	5	1421 ±30	**3665** ±56	**3260** ±92	5730 ±128	5751 ±90	5797 ±37	4256 ±15	4312 ±19	4230 ±16	4254 ±14	4203 ±18	4328 ±19
	10	2401 ±36	**3842** ±43	**3330** ±59	5247 ±177	5387 ±150	5657 ±60	4256 ±15	4312 ±19	4230 ±16	4254 ±14	4203 ±18	4328 ±19
	20	3147 ±17	**4012** ±36	**3714** ±33	5016 ±101	5148 ±90	5132 ±28	4256 ±15	4312 ±19	4230 ±16	4254 ±14	4203 ±18	4328 ±19
200	5	1451 ±21	**4096** ±60	**3704** ±101	5889 ±102	5962 ±129	5928 ±49	4898 ±20	4934 ±20	4917 ±21	4927 ±18	4636 ±19	4708 ±21
	10	2510 ±43	**4240** ±48	**3681** ±55	5894 ±144	5678 ±141	6021 ±56	4898 ±20	4934 ±20	4917 ±21	4927 ±18	4636 ±19	4708 ±21
	20	3383 ±19	**4453** ±56	**4122** ±31	4815 ±120	5103 ±147	5126 ±59	4898 ±20	4934 ±20	4917 ±21	4927 ±18	4636 ±19	4708 ±21

Open Bandit. On the Open Bandit logs an SPSC variant beats every non-oracle baseline on three of the four cells, by 13% and 7% at d=55 and by 6% at (d, r)=$(105, 5)$ (Table 19). At $(105, 10)$ Restart-LinUCB, which restarts at the true change points, is best, 1% ahead of SPSC-Adaptive. The ambient LinUCB variants are close to one another on these logs, whose segments are consecutive time slices with small shifts, and SPSC's edge is largest at the smaller rank.

Table 19: **Open Bandit (ZOZOTOWN)** per-cell results (T=5,000, K=10, 10 seeds).

d	r	Oracle	**SPSC**	**SPSC-Adp**	LowOFUL	VOFUL	LR-Rew	SW-Lin	D-Lin	Rst-Lin	LinUCB	LinTS	SW-LinTS
55	5	36.3 ±0.5	**81.9** ±2.2	**83.9** ±3.1	121.9 ±2.2	119.5 ±2.6	121.5 ±1.6	95.0 ±0.3	97.2 ±0.3	94.0 ±0.5	94.0 ±0.3	116.5 ±1.0	116.1 ±0.9
	10	69.0 ±0.8	**87.4** ±1.1	**88.7** ±1.8	132.5 ±1.4	129.2 ±1.3	130.8 ±0.9	95.0 ±0.3	97.2 ±0.3	94.0 ±0.5	94.0 ±0.3	116.5 ±1.0	116.1 ±0.9
105	5	32.1 ±0.4	**101.6** ±3.4	**95.8** ±2.7	143.7 ±1.4	142.7 ±2.0	144.1 ±2.6	102.6 ±0.7	103.3 ±0.4	102.1 ±0.5	102.7 ±0.7	140.3 ±1.2	138.2 ±0.8
	10	62.2 ±0.6	104.5 ±1.8	103.4 ±2.2	156.6 ±1.6	155.0 ±1.0	156.6 ±1.2	102.6 ±0.7	103.3 ±0.4	102.1 ±0.5	102.7 ±0.7	140.3 ±1.2	138.2 ±0.8

F.2 BOSS AND JEDRA

Intuition. Stationary low-rank methods use the same rank structure as SPSC but assume one fixed subspace. This comparison asks what tracking a moving subspace adds over learning it once.

Setup. Our adaptations of BOSS (Duong et al., 2024) and Jedra (Jedra et al., 2024) learn a subspace in their first 200 rounds, BOSS from sphere probes and Jedra from random actions, and then run projected LinUCB in it. SPSC-Adaptive detects the change points, and Algorithm 1 is given them. The benchmark is the synthetic grid's model, with $\theta^{(k)}$ fixed within each segment, K=10, T=5,000, σ_ε=0.3, 40 actions, 10 seeds, and probe cost c=0.1, on the eight cells $(55, 5)$, $(55, 10)$, $(105, 5)$, $(105, 10)$, $(105, 20)$, $(200, 5)$, $(200, 10)$, and $(200, 20)$.

Findings. Tracking the subspace pays off on every cell. Both methods lose to SPSC-Adaptive on every cell, by 0.9–13.8% (Table 3). Algorithm 1, which is given the change points, beats them by 5.9–19.8%. The lead is largest at d=55 and narrows at d=200.

F.3 SMALL-d STRESS TEST AGAINST NONSTATIONARY BASELINES

Intuition. Our main benchmarks have large d, so one may ask whether SPSC helps only there. A small-d test with the dimensions used to validate D-LinUCB answers this.

Setup. We use our AR(1) environment with the dimensions of Russac et al. (2019) (d=2, r=1, K=4, T=6,000, σ_ε=1, 50 arms, 30 seeds). D-LinUCB, SW-LinUCB, OFUL, and Reset-LinUCB all reset at the true change points.

Findings. SPSC also helps at small d, with the lowest regret of all methods except Oracle-LinUCB (Figure 3). Its control regret, $2{,}229 \pm 54$, is 46.5% below D-LinUCB's 4,167 and 34.8% below SW-LinUCB's 3,417. OFUL and Reset-LinUCB are near 4,650. Probes raise its costed regret by 0.9%, to 2,249, which is 46.0% and 34.2% below the two (§5.3).

G CHECKS OF THE THEORY AND CONTROLS

The figures of this appendix plot regret without the probe fee unless a caption says otherwise. The fee adds c times the number of probes, the same amount for every SPSC run of a figure.

These studies test the theory in the runs. Table 4 summarizes them with the other experiments. The first checks the conditions of the analysis. The next two vary the probe rate and break the identification conditions one at a time. The fourth tracks the subspace error within and across segments. The last gives SPSC's probes to ambient baselines.

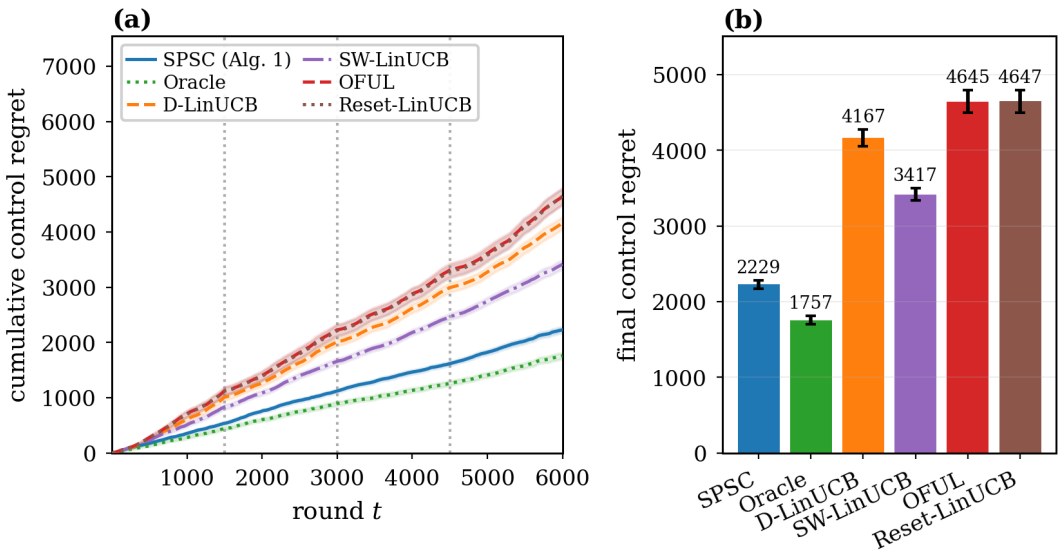


Figure 3: **Small-d stress test** with the dimensions of Russac et al. (2019) (d=2, r=1, K=4, T=6,000, 30 seeds). (a) Cumulative control regret, with dotted lines at the three change points. (b) Final control regret per method.

G.1 The conditions of the analysis in the runs

Intuition. Theorem 4.1 assumes well-conditioned windows, condition (WC), and allows any refresh schedule that uses at least half of the probes, condition (R). We check both in the runs.

Table 20: **Refresh schedule.** Costed regret, mean±SE over 10 paired seeds. The runs recompute the basis after every probe. The last column recomputes it only after the 1st, 2nd, 4th, … probe. The change is relative to the runs, with paired SEs of 0.5 to 1.4 points.

Benchmark	LinUCB	Runs	Refresh after $1, 2, 4, \ldots$
Synthetic, d=30, r=5	545 ± 18	374 ± 9	377 ± 10 (+0.8%)
Synthetic, d=100, r=10	471 ± 14	372 ± 10	372 ± 10 (−0.1%)
Pendigits, d=105, r=10	3258 ± 5	2150 ± 58	2132 ± 59 (−0.9%)

Setup. On the three benchmarks of Table 20 and 10 seeds, we record at every exploitation round with $|\mathcal{W}_t| \geq 2r$ the smallest eigenvalue of the window's projected design $\sum_{s \in \mathcal{W}_t} z_t(x_s) z_t(x_s)^\top$, divided by $|\mathcal{W}_t|$. We also compare the runs, which recompute the basis after every probe, with a version that recomputes it only after the 1st, 2nd, 4th, … probe.

Findings. Both conditions hold in practice, and the refresh schedule makes no difference. The ratio is positive at every such round, with minimum 8.5×10^{-5}, first percentile about 3×10^{-4}, and medians from 8×10^{-4} to 6×10^{-3}, so κ_z is of order 10^{-3}. Recomputing the basis only after the 1st, 2nd, 4th, … probe changes regret by at most 0.9% (Table 20).

G.2 Probe-rate ablation

Intuition. The probe rate is the main tuning choice in SPSC. Rare probes leave the subspace poorly estimated and frequent probes waste rounds, so the bound has an interior optimum in the rate.

Setup. We sweep the probe period over $\{5, 10, 20, 30, 50, 100, 300\}$ at d=4, r=1, K=4, and T=6,000 (10 seeds), and record the final regret and the late-segment subspace error.

Findings. A probe every 10 to 30 rounds works best, as the bound suggests. Regret is lowest at a probe period of 20, and periods of 10 and 30 are within about 5% of it (Figure 4a). For periods of

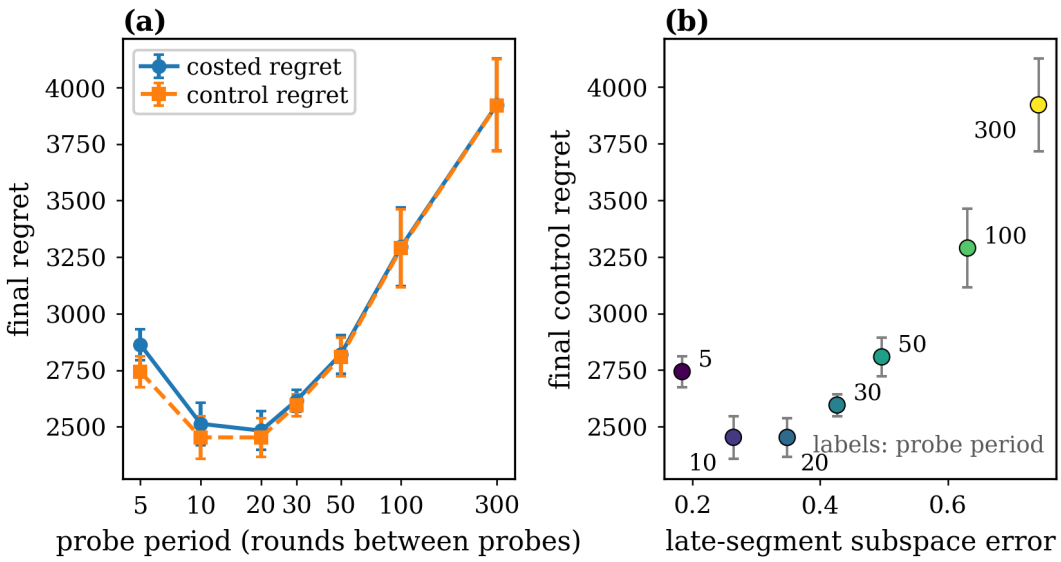


Figure 4: **Probe-rate ablation.** (a) Final regret against the probe period. (b) Final control regret against the late-segment subspace error, with each point labelled by its probe period.

10 or more, a lower late-segment subspace error goes with a lower regret (Figure 4b). At period 5 the error is lowest, but the extra probes cost more than the better estimate returns.

G.3 Robustness and necessity experiments

Intuition. Theorem 2.1 says that under probe independence coverage is what matters, and that the noise variance and the state–noise coupling matter only when the noise depends on the probe direction. We test each condition in turn.

Setup. We break one condition at a time on the reference setting (d=4, r=1, K=4, T=6,000, probe period 30, W=100, 10 seeds) and measure the regret. The centering error of the noise variance goes up to 0.5, more than five times the true σ_ε^2=0.09. The noise is coupled to the reward as $\varepsilon_t \to \varepsilon_t + \epsilon_\times x_t^\top \theta_t$ with $\epsilon_\times$ up to 1. The probes are restricted to one, two, three, or all four directions.

Findings. Only coverage matters in practice, as the theory predicts (Figure 5). A wrong noise variance changes regret by less than 5% (panel a), since a constant centering error adds a multiple of the identity to the lifted estimate and leaves its eigenvectors unchanged. Coupling as strong as the signal itself changes regret by less than 1% (panel b). Restricting the probes to fewer directions raises regret steadily. With one of four directions, SPSC's regret is 4,289, close to LinUCB's 4,463 (panel c), in line with the necessity of coverage in Theorem 2.1.

G.4 Subspace recovery rate and change-point adaptation

Intuition. Theorem 2.1 says the subspace error falls as probes accumulate, and SPSC must rebuild its estimate after every change. We track the error within a segment and across change points.

Setup. We use the reference setting (d=4, r=1, K=4, T=6,000). The within-segment curve probes every 10 rounds over 20 seeds, and the run across change points probes every 30 rounds over 10 seeds.

Findings. The error falls steadily as probes arrive and jumps back up at each change, as the theory describes. It falls from about 0.8 after one probe to about 0.2 after 150 (Figure 6), with a log-log slope of about -0.3 over this range. Corollary B.10 bounds the error for every m by a square-root term plus a finite-sample term of order $d\log(4d/\delta)/m$, clipped at one, and the average slope of

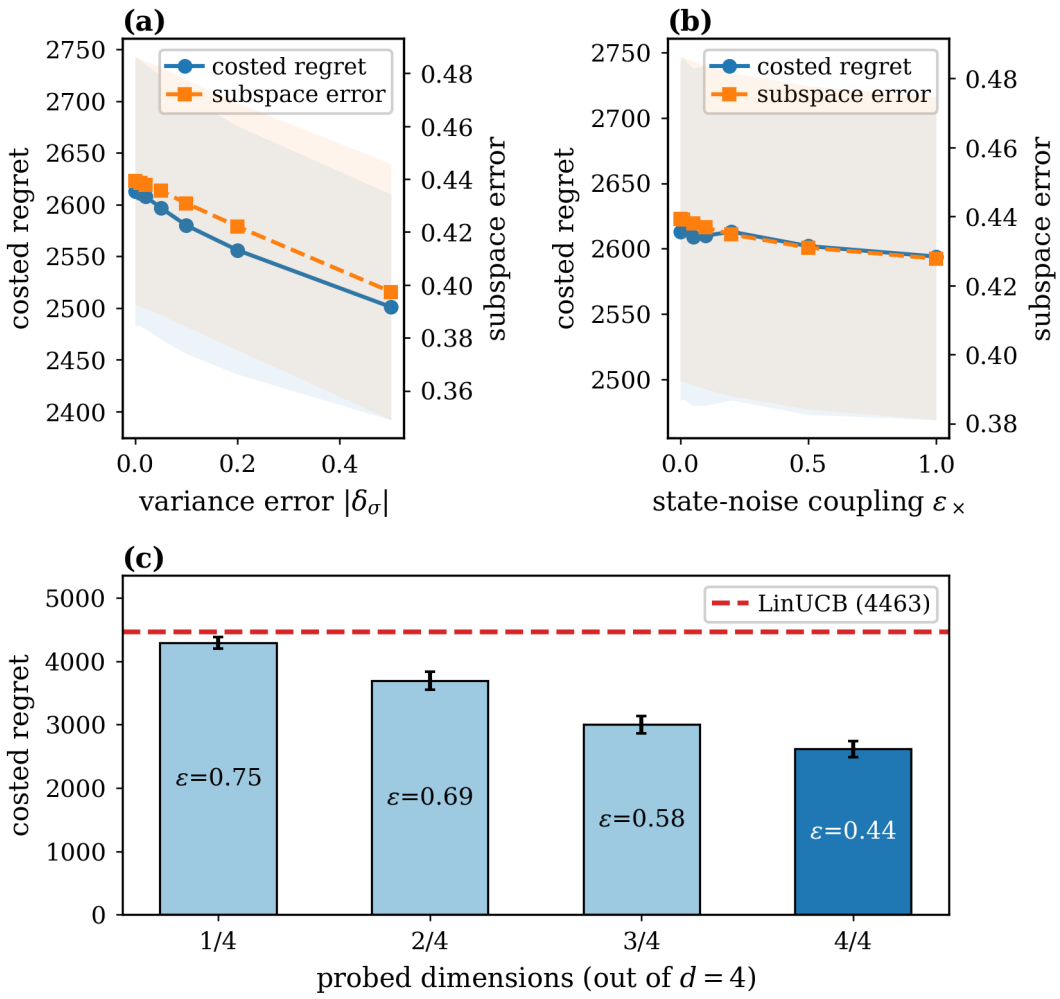


Figure 5: **Robustness and necessity experiments.** Costed regret and subspace error under (a) a wrong noise variance, (b) noise coupled to the reward, and (c) probes restricted to part of the space.

that bound over a range that starts where it is still clipped is shallower than $-1/2$, as observed. The square-root term takes over once m exceeds $d\log(4d/\delta)$, about 23 probes here (d=4, δ=0.05). After each change point the error jumps back to about 0.8 and then falls again as new probes arrive (Figure 7b). SPSC's smoothed regret per round is below LinUCB's in 84% of the rounds, and its control regret over the run is 42% lower (2,596 vs. 4,468, Figure 7a).

G.5 The probes given to ambient baselines

Intuition. SPSC sees probe responses that an ambient learner never sees. We ask whether its gains come from those observations themselves or from what SPSC does with them.

Setup. Probe-LinUCB is LinUCB given exactly the probes of Algorithm 1. It probes on the same rounds in the same directions, observes the same responses, pays the same forgone reward and fee, and adds each probe to its ambient ridge regression as an ordinary observation. Like LinUCB it resets at the true change points. Probe-SW-LinUCB is SW-LinUCB with probes at SPSC-Adaptive's base rate, kept in its window like any other observation, and it is not told the change points. Both use the confidence radius of the baseline they extend. Table 21 compares each SPSC variant with its probe-fed counterpart on every cell of the synthetic grid and the seven real-data benchmarks, with 10 seeds and costed regret.

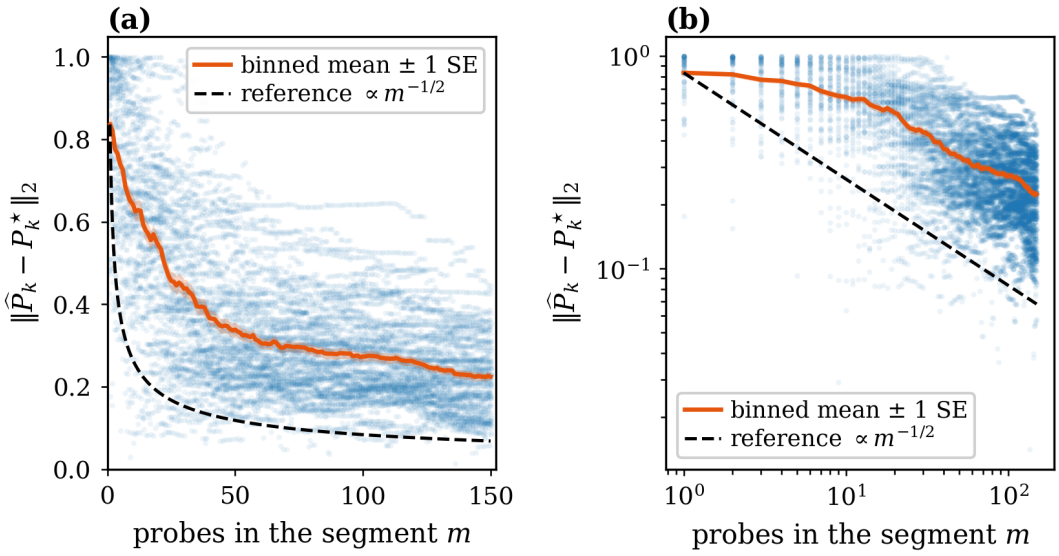


Figure 6: **Subspace recovery** (d=4, r=1, 20 seeds). (a) Projector error against the number of probes in the current segment, for individual runs and the binned mean. (b) The same on a log-log scale, with an $m^{-1/2}$ reference line.

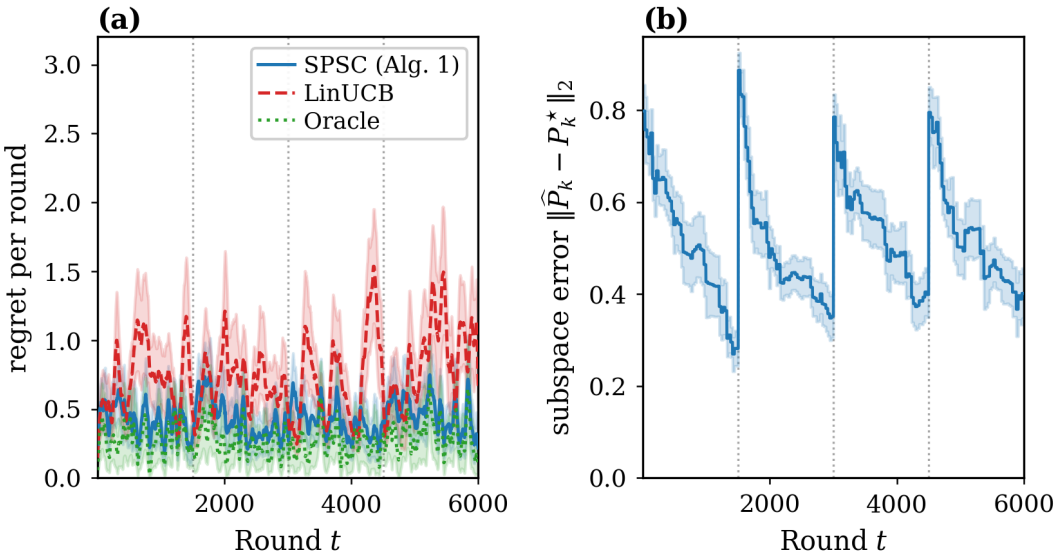


Figure 7: **Change-point adaptation.** (a) Instantaneous control regret, smoothed over 60 rounds. (b) Mean subspace error, with dotted lines at the change points.

Findings. Giving an ambient learner the same probes does not close the gap (Table 21). An SPSC variant beats both probe-fed baselines in 82 of the 98 cells, and in 49 of the 50 cells with $d \geq 60$. Algorithm 1 has lower regret than Probe-LinUCB in 75 cells, and SPSC-Adaptive lower regret than Probe-SW-LinUCB in 86. The probes help an ambient learner most when a segment's probes nearly span $\mathbb{R}^d$. The 50 probes of a real-data segment span 50 of the 55 dimensions at d=55 but under half of them at $d \geq 105$, and the 10 probes of a grid segment span at most half of them for $d \geq 20$. At d=55 the probes lower LinUCB's regret by 36% on Satimage and 25% on Pendigits, and an SPSC variant leads in 15 of the 23 real-data cells. The margin grows with d, as the crossover of Theorem 4.1 predicts, to 31 of the 32 real-data cells with $d \geq 105$ and all 35 grid cells with $d \geq 20$. SPSC turns the same probes into an r-dimensional basis and exploits inside it, so its exploitation cost scales with r, while the probe-fed baselines still learn in d dimensions.

Table 21: **SPSC against ambient baselines given the same probes** (costed regret, 10 seeds). Ratios are medians over the cells of each benchmark, and wins count the cells where the SPSC variant has the lower regret. The last column counts the cells where the better SPSC variant beats both probe-fed baselines.

Benchmark	Cells	Probe-LinUCB /LinUCB	SPSC /Probe-LinUCB	Wins	SPSC-Adp /Probe-SW-LinUCB	Wins	Beats both
Synthetic grid	40	1.01	0.81	35	0.84	38	35
Covertype	9	0.99	0.96	7	0.90	9	9
Pendigits	9	0.89	1.04	3	0.95	5	5
Satimage	9	0.80	0.84	6	0.78	6	6
MNIST	9	0.95	0.78	7	0.77	8	8
Fashion-MNIST	9	0.94	0.75	6	0.77	8	7
Open Bandit	4	1.04	0.92	4	0.89	4	4
MovieLens	9	0.99	0.92	7	0.83	8	8
All	98	0.99		75		86	82