KNOWN-SUBSPACE RATE LADDER FOR STREAMING LOW-RANK REGRESSION WITH FAST COEFFICIENT DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study streaming low-rank regression when the relevant subspace is fixed but its coefficients can change abruptly. This setting exposes a plasticity–stability tradeoff: decaying steps learn stationary segments well but recover slowly after jumps, while constant steps react quickly but retain a noise floor. We prove a known-subspace rate ladder for expected dynamic regret under the same linear-Gaussian sample feedback. Monotone decaying-step stochastic gradient suffers linear regret under coefficient jumps, whereas a clipped fixed-window estimator achieves $O(T^{3/4})$ regret when the number of jumps grows as $\sqrt{T}$. An adaptive estimator that combines generalized likelihood ratio (GLR) change-point detection with clipped least squares achieves $O(\sqrt{T}\log T)$ on a well-spaced class with $\Theta(\sqrt{T})$ jumps, and a van Trees argument gives a matching $\Omega(\sqrt{T}\log T)$ lower bound on the same class. Charging each change by its own size, the ratio of the two bounds depends only on the condition number and the signal-to-noise ratio, uniformly in the minimum jump size, whereas a size-independent charge grows as the squared amplitude-to-jump ratio and is linear in T for the smallest jumps the spacing admits. When the subspace is fixed but unknown, running the same detector on half of the samples projected onto a spectral-moment estimate built from the other half keeps the $O(\sqrt{T}\log T)$ rate at an additive cost of order $rd\log^2(dT)$ for learning the subspace when $d \geq \log(dT)$, and a matching lower bound shows that this cost is necessary up to the factor $\log(dT)$ when $d = O(T^{1/4})$; a blockwise Fantope extension covers moving subspaces at the $T^{3/4}$ scale. In synthetic experiments, a fixed window cuts the regret of a theorem-class decaying-step learner 11.3-fold at $T = 50{,}000$; the analyzed GLR rule, run exactly on equally spaced jumps of size at least 0.5 with isotropic Gaussian covariates at high signal-to-noise ratio, grows at close to the oracle's rate and has lower regret at $T \geq 50{,}000$ than every tuned baseline; and a suffix-retaining GLR variant has 17–37% lower regret than the preregistered tuned competitor with the lowest tuning regret at $T \in \{50{,}000, 100{,}000\}$ on equally and unequally spaced change paths with mixed jump sizes, three covariate designs and noise levels 0.1–1.

1 INTRODUCTION

Streaming adaptation under distribution drift is a central challenge in online learning (Hazan, 2016). When the time-varying optimum θ_t^* has low-rank structure, as in adapter-style updates or subspace tracking (Oja, 1982; Hu et al., 2022), a central source of nonstationarity is *fast-switching coefficients* a_t inside a low-dimensional subspace. We study this known-subspace coefficient-drift model and extend it to unknown subspaces. A learner that decays its step size $\eta_t \to 0$ (Robbins & Monro, 1951) converges on stationary segments but recovers slowly from abrupt jumps, whereas a constant η_t tracks jumps at the cost of a persistent noise floor. We isolate the resulting rate ladder, complementing bounds for stationary streams (Hazan, 2016) and path-length drift (Zinkevich, 2003).

Contributions. We prove a *rate ladder* for streaming low-rank regression with fast coefficient drift in the linear-Gaussian known-subspace model, under exact low-rank structure with $\rho_{\max} = 0$

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.

and one feedback model, in which the learner sees the samples or their squared-loss gradients, establishing upper and lower bounds for expected dynamic regret that characterize the cost of each sample-selection rule. Our specific contributions are:

1. **Known-subspace rate ladder with proven upper and lower bounds.** By Theorem 1, monotone-step stochastic gradient suffers $\Omega(T)$ regret; by Theorem 2, clipped fixed-window algorithms achieve $O(T^{3/4})$; by Theorem 4, known-U adaptive windows reach $O(\sqrt{T}\log T)$; and the information-theoretic lower bound of Theorem 8 is $\Omega(\sqrt{T}\log T)$.
2. **Rate-optimal detect-and-restart regression.** Generalized likelihood ratio (GLR) detection with clipped least squares has a finite-sample expected-regret bound in which false alarms, detection delay and post-reset estimation appear as separate terms, with a single-logarithm delay. On separated well-spaced jumps it matches the lower bound up to a factor $O(\kappa\rho_D + LA_{\max}^2/\sigma^2)$, with condition number $\kappa = L/\mu$ and squared amplitude-to-jump ratio $\rho_D = A_{\max}^2/\Delta_0^2$, that does not grow with T, as stated in Corollary 6; charging each change by its own size, as in Theorem 5, replaces $\kappa\rho_D$ by κ^3, uniformly in the minimum jump size. By Corollary 7, at every jump rate $K_T = T^\gamma$ with $\gamma < 1$ it is polynomially below every fixed window, even one tuned with knowledge of γ.
3. **Adaptivity with an unknown subspace.** For a fixed unknown U, GLR on half of the samples, projected onto a spectral-moment estimate from the other half, keeps the adaptive rate at an additive cost $O(rd\log^2(dT) + \log^3(dT))$ free of the number of changes by Theorem 10, and Proposition 11 shows that this cost is necessary up to $\log(dT)$ for $\log(dT) \le d = O(T^{1/4})$; for blockwise moving subspaces, Fantope online gradient descent (OGD) reaches $O(T^{3/4})$ by Theorem 12.
4. **Algorithm-matched synthetic evidence.** A fixed window cuts a theorem-class decaying-step learner's regret 11.3-fold at $T = 50{,}000$; at high signal-to-noise ratio the exact GLR rule grows at nearly the oracle's rate and beats every tuned baseline at $T \ge 50{,}000$; a suffix-retaining variant has 17–37% lower regret than the preregistered tuned competitor with the lowest tuning regret in 28 long-horizon conditions.

Main technical ideas. First, the population gradient is affine, so the *mean* SGD iterate follows noiseless gradient descent exactly and each jump is charged by a dichotomy: either the recovery after it or the stationary rounds before it are slow. Second, a declaration time is a stopping time, so the least-squares risk after a reset is computed for a *ghost* estimator on fresh samples, without conditioning on future non-detection. Third, during a detection delay the mixed-window bias is at most $9\kappa\Delta_k$: a large jump is detected quickly and a small one costs little per round, which makes the transient uniform in the jump size. Fourth, a smooth, compactly supported prior removes the van Trees boundary term and gives a per-segment cost of order $r\sigma^2 \log \frac{T}{K_T+1}$. Finally, for unknown U, splitting the stream makes the regression rounds independent of the estimated subspace, so given it they follow the known-subspace model with a noise level that changes only at the change points, and the detection analysis applies unchanged.

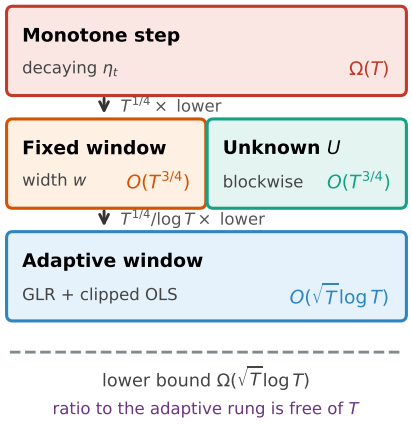


Figure 1: Rate ladder for $K_T = \Theta(\sqrt{T})$ jumps with known U.

Notation. Table 14 in Section B lists all symbols; Table 12 summarizes the guarantees, including those for unknown U.

2 Related Work

Online learning under drift. Dynamic regret was introduced by Zinkevich (2003); later analyses cover variation-budget, path-length and dynamic-comparator settings (Besbes et al., 2015; Hall & Willett, 2013; Jadbabaie et al., 2015), and bandit analogues study switching rewards (Garivier & Moulines, 2011). Closest to our known-U problem, which after projection is r-dimensional online

regression, are adaptive-regret methods. Follow-the-leading-history (FLH) (Hazan & Seshadhri, 2009) and strongly adaptive learning (Daniely et al., 2015) run learners started at different times and control regret on every interval; their exp-concave guarantees assume bounded losses, so unbounded Gaussian responses need truncation. Jacobsen & Cutkosky (2024) analyze discounted online linear regression in dynamic environments with learned discounting, under a protocol where the prediction may use the current feature, and Zhang et al. (2025) bound the dynamic regret of fixed-share exponential weights for mixable losses by comparator path length. Our detect-and-restart learner instead has an expected-regret decomposition into false alarms, delay and post-reset estimation, matched by a lower bound on the same jump class, and §7 compares it empirically with FLH and forgetting least squares.

Low-rank streaming estimation. Oja's algorithm and variants track stationary eigenspaces (Oja, 1982; Balsubramani et al., 2013), also on manifolds (Bonnabel, 2013). The spectral moment $y^2 xx^\top$ of multi-task representation learning (Tripuraneni et al., 2021) is our subspace estimator: for a fixed U we estimate it on half of the rounds and pay a representation cost linear in rd, which we show is necessary; for moving subspaces we track it blockwise with Fantope OGD (Vu et al., 2013) against a moving comparator. Lifelong representation learning (Alquier et al., 2017; Cao et al., 2022) is told where tasks start; here the learner must detect task boundaries. LoRA (Hu et al., 2022), continual-learning regularizers (Kirkpatrick et al., 2017) and online LoRA (Wei et al., 2025) are motivation only; our claims are linear-Gaussian regret bounds.

Change-point detection and cumulative regret. GLR and CUSUM tests are classical tools for abrupt shifts (Lai, 1998); Lepski's method selects locally homogeneous bandwidths (Lepski et al., 1997). Gafni et al. (2026) track a directly observed mean with multiple change points; applied to the unbiased statistics $\sqrt{\pi/2}\, y_t \operatorname{sgn}(z_{t,i})$, their guarantee reaches the order $(K_T{+}1)\log T$ for isotropic designs with known U. We cover unknown anisotropic covariances, where these statistics are biased, with explicit r and κ, a lower bound for the regression class and the unknown-subspace cost; see Section J. Hüselitz et al. (2026) study online jump detection in segmented regression. Among bandit and control analogues, SVUCRL (Dam, 2025), change-point dynamic pricing (Zhao et al., 2026), GLR-CUCB (Zhou et al., 2020) and ISD-linUCB (Lazzaretto et al., 2026) differ in feedback; see Section J. Our bounds are on expected regret; high probability enters only through the detection event.

3 Problem Setting and Rate Ladder Preview

Model. We study streaming low-rank regression in the linear-Gaussian model. At each round $t = 1, \dots, T$, the learner commits to $\theta_t \in \mathbb{R}^d$ using the samples of rounds $1, \dots, t-1$, and then observes $x_t \sim \mathcal{N}(0, \Sigma_t)$ and $y_t = \theta_t^{*\top} x_t + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The time-varying optimum admits a low-rank decomposition:

$$\theta_t^* \;=\; U_t\, a_t \;+\; \rho_t, \qquad U_t \in \mathbb{R}^{d\times r},\; U_t^\top U_t = I_r,\; a_t \in \mathbb{R}^r,\; \|\rho_t\| \le \rho_{\max}, \tag{1}$$

where U_t is a subspace factor, a_t is a fast-switching coefficient vector, piecewise constant with K_T jumps, and ρ_t is a small residual. The main rate ladder in §3–§5 sets $U_t \equiv U$ known and fixed and $\rho_{\max} = 0$. §6 gives a separate blockwise unknown-U extension.

Assumptions. We impose the following regularity conditions on the population loss $L_t(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[\ell(\theta; x, y)]$ and the drift process.

Assumption 1 (μ-strong convexity)**.** *$L_t(\theta) - L_t(\theta') \ge \langle \nabla L_t(\theta'), \theta - \theta' \rangle + \frac{\mu}{2}\|\theta - \theta'\|^2$ for all θ, θ', t.*

Assumption 2 (L-smoothness)**.** *$\|\nabla L_t(\theta) - \nabla L_t(\theta')\| \le L\|\theta - \theta'\|$ for all θ, θ', t. We write $\kappa = L/\mu$.*

Assumption 3 (Sample-gradient feedback)**.** *With $\ell(\theta; x, y) = \frac{1}{2}(y - \theta^\top x)^2$, a gradient-based learner receives $g_t(\theta) = (x_t^\top \theta - y_t)x_t$, which is conditionally unbiased for $\nabla L_t(\theta)$; its variance depends on θ and is not assumed to be bounded by σ^2.*

We additionally use a low-rank-plus-residual decomposition with $1 \le r < d$ in Assumption 6, bounded coefficient variation with minimum jump size $\Delta_0 > 0$ in Assumption 7, bounded coefficients $\|a_t\| \le A_{\max}$ in Assumption 8, and an orthogonal residual $\|\rho_t\| \le \rho_{\max}$ in Assumption 9. Full statements are in Section B.

Regret and distance measures. We measure performance by the expected dynamic regret $\mathrm{Reg}_T = \mathbb{E}\big[\sum_{t=1}^{T}(L_t(\theta_t) - L_t(\theta_t^*))\big]$ and subspace distance by the Grassmann distance $d_{\mathrm{Gr}}(\mathrm{col}(U), \mathrm{col}(U')) = \|\sin\Theta\|_F$, where Θ is the diagonal matrix of principal angles; it is invariant to $U \mapsto UR$ for orthogonal R and equals $\big\|UU^\top - U'U'^\top\big\|_F / \sqrt{2}$.

All seven theorems share this model; all but Theorems 10 and 12 have a known fixed U. The rungs target the same optimum and differ only in which past samples enter the estimate: all of them with a decaying step in Theorem 1, the last w in Theorem 2, or all since the last declaration in Theorem 4.

3.1 Theorem 1: Impossibility for Monotone-Step Learners

A *monotone-step learner* maintains a single iterate $\theta_{t+1} = \theta_t - \eta_t g_t(\theta_t)$ where g_t is a conditionally unbiased stochastic gradient as in Assumption 10, for instance the sample gradient of Assumption 3, $\eta_t = c \cdot t^{-\alpha}$ with $\alpha \in [1/2, 1]$ and $0 < c \le \mu/(2L^2)$. This class *excludes* restart strategies, change-point detectors, expert mixtures, adaptive preconditioners such as Adam and AdaGrad, and any algorithm whose effective step size can increase after a drift event.

Theorem 1 (Impossibility for monotone-step learners)**.** *In the known-U linear-Gaussian model with $\rho_{\max} = 0$ and a design covariance Σ with $\mu I \preceq \Sigma \preceq LI$ and an eigenvector $u \in \mathrm{col}(U)$ of eigenvalue μ, let $\Delta_0 \in (0, 2A_{\max}]$ and $T \ge 64$. For every monotone-step learner and every number of jumps K_T with $\frac{1}{2}T^{1-\alpha} \le K_T \le T/8$, in particular $K_T = \lfloor\sqrt{T}\rfloor$ for every $\alpha \in [\frac{1}{2}, 1]$, the coefficient path that alternates between $\pm\frac{\Delta_0}{2}u$ on K_T+1 equal segments satisfies*

$$\mathbb{E}[\mathrm{Reg}_T] \;\ge\; \frac{3\,\mu\Delta_0^2}{8192}\,T.$$

The proof in Section D follows the mean iterate, which obeys noiseless gradient descent exactly, and charges each jump by the dichotomy of the first idea in Section 1, for any unbiased gradient feedback.

3.2 Theorem 2: Fixed-Window Dual-Timescale Algorithm

Theorem 2 (Fixed-window sufficiency)**.** *Let $x_t \sim \mathcal{N}(0, \Sigma)$ be independent, let the observation noise be independent $\mathcal{N}(0, \sigma^2)$, and let U be a known fixed orthonormal matrix. Assume exact low-rank optima $\theta_t^* = Ua_t$, where $\{a_t\}$ is a deterministic sequence independent of the observations, $\|a_t\| \le A_{\max}$, and there are at most K_T coefficient changes. Write $\Sigma_z = U^\top \Sigma U$ and assume $\mu I \preceq \Sigma_z \preceq LI$. For an integer $r + 1 < w < T$, predict zero for the first w rounds and then use ordinary least squares (OLS) on the previous w observations, projected onto the Euclidean ball of radius $A_{\max}$. Then*

$$\mathbb{E}[\mathrm{Reg}_T] \le \frac{LA_{\max}^2 w}{2} + \frac{L\sigma^2 r(T-w)}{2\mu(w-r-1)} + 2LA_{\max}^2 \min\{T-w, K_T w\}.$$

For fixed $r, A_{\max}, L, \mu, \sigma$ and a feasible window $w \asymp \sqrt{T/(K_T+1)}$, this is $O(\sqrt{T(K_T+1)})$; in particular it is $O(T^{3/4})$ when $K_T = \Theta(\sqrt{T})$.

Clipping controls windows that contain changes, so no minimum segment length is required. The order is tight for this learner family: Proposition 16 in Section E shows that every window width $w \ge r + 2$, clipped or not, incurs $\Omega(\min\{\Delta_0\sigma\sqrt{rTK_T},\ \sigma^2 T,\ \Delta_0^2 T\})$ on isotropic streams with at most K_T jumps of size Δ_0, so fixed-window least squares is $\Theta(T^{3/4})$ when $K_T = \Theta(\sqrt{T})$.

4 Adaptive Known-Subspace Estimation

The adaptive rule replaces the fixed bandwidth of Theorem 2 by a window that expands during stationary periods and resets when a GLR test detects a coefficient jump.

4.1 Algorithm: Adaptive-Window Clipped OLS

Definition 3 (Adaptive-window clipped-OLS learner)**.** *Given a* known, fixed *subspace $U \in \mathbb{R}^{d\times r}$ with orthonormal columns, define projected observations $z_s = U^\top x_s$. The algorithm maintains a window start pointer s_0, initialized to* 1.

1. ***Expanding-window OLS.*** *At each time t, compute the clipped ordinary least-squares estimate on the growing window $\{(z_s, y_s)\}_{s=s_0}^{t-1}$:*

$$\hat{a}_t = \Pi_{\|a\| \le A_{\max}} \left(\hat{\Sigma}_z^{-1} \hat{b}_z \right), \qquad \hat{\Sigma}_z = \frac{1}{w} \sum_{s=s_0}^{t-1} z_s z_s^\top, \quad \hat{b}_z = \frac{1}{w} \sum_{s=s_0}^{t-1} z_s y_s, \tag{2}$$

where $w = t - s_0$ is the current window length and $\Pi_{\|a\| \le A_{\max}}$ denotes Euclidean projection onto $\{a : \|a\| \le A_{\max}\}$. If $w < r + 2$, set $\hat{a}_t = 0$; these are burn-in rounds. The clipping ensures $\|\hat{a}_t\| \le A_{\max}$ always, bounding worst-case per-step regret even when $\hat{\Sigma}_z$ is ill-conditioned.

2. ***GLR change-point detection.*** *Every Δ_c steps, for a fixed cadence $1 \le \Delta_c \le \lceil r \log T \rceil$, and with the additional requirement that the window hold $n = t - s_0 + 1 \ge 2r + 2$ observations, evaluate the GLR statistic on $\{s_0, \dots, t\}$, which compares one least-squares fit on the current window with the best two-segment fit split at a candidate τ. With $\mathrm{RSS}_{a:b} = \min_{a' \in \mathbb{R}^r} \sum_{s=a}^{b} (y_s - z_s^\top a')^2$ the least-squares residual sum of squares on $\{a, \dots, b\}$,*

$$\Lambda(\tau, t) = \frac{1}{2\sigma^2} \big(\mathrm{RSS}_{s_0:t} - \mathrm{RSS}_{s_0:\tau-1} - \mathrm{RSS}_{\tau:t} \big) \tag{3}$$

over all admissible candidate change-points $\tau \in \{s_0 + r + 1, \dots, t - r\}$. If $\max_\tau \Lambda(\tau, t) \ge h$, where the threshold satisfies $h_{\min} \le h \le c_h r \log(2T^3)$ with $h_{\min} = \frac{1}{2}(r + 2\sqrt{ru} + 2u)$, $u = \log(8T^5)$ and $c_h \ge 11/2$, declare a change-point at $\hat{\tau} = \arg\max_\tau \Lambda(\tau, t)$ but reset the estimation window at the declaration time*: $s_0 \leftarrow t + 1$. Thus the next OLS fit uses only fresh post-declaration samples.*

The window is the largest one the GLR test certifies as homogeneous, in the spirit of Lepski's bandwidth selection (Lepski et al., 1997); with prefix sums a scan costs $O(nr^3)$ for a window of length n. The prediction at round t uses $(z_s, y_s)_{s<t}$, and clipping keeps every round's regret below $2LA_{\max}^2$.

4.2 Known-Subspace Adaptive-Window Sufficiency

Theorem 4 (Known-subspace adaptive-window sufficiency)**.** *Let $x_t \sim \mathcal{N}(0, \Sigma)$ be independent with a fixed covariance satisfying Assumptions 1 and 2, let the noise be independent $\mathcal{N}(0, \sigma^2)$, and let the* known, fixed *subspace U, so that $V_s = 0$, carry exact low-rank optima with $\rho_{\max} = 0$ whose coefficient path is deterministic and fixed in advance, with K_T jumps of size $\ge \Delta_0$ at times $\tau_1 < \cdots < \tau_{K_T}$, which split $[1, T]$ into segments $S_0, \dots, S_{K_T}$ of lengths $J_k = |S_k|$, and minimum inter-jump gap $J_{\min} = \min_k J_k$ satisfying*

$$J_{\min} \ge n_0 + 2(D + \Delta_c), \qquad n_0 = C_n(r + \log T),\ C_n = 200,$$
$$D = C_D \left(n_0 + \frac{r\sigma^2 \log T}{\mu \Delta_0^2} \right), \qquad 1 \le \Delta_c \le \lceil r \log T \rceil.$$

Then the adaptive-window clipped-OLS learner of Definition 3 achieves:

$$\mathbb{E}[\mathrm{Reg}_T] \le \underbrace{4LA_{\max}^2 K_T (D + \Delta_c)}_{\text{detection transient}} + \underbrace{(K_T + 1) \cdot 2\kappa\sigma^2 r \cdot \log \tfrac{eT}{(K_T+1)n_0}}_{\text{OLS convergence}}$$
$$+ \underbrace{(K_T + 1) \cdot 2LA_{\max}^2 n_0}_{\text{burn-in}} + \underbrace{O(LA_{\max}^2)}_{\text{rare detection failure}}, \tag{4}$$

where $\kappa = L/\mu$ and $\rho_D = A_{\max}^2/\Delta_0^2$. For $K_T = \Theta(\sqrt{T})$ and $n_0 = O(r \log T)$:

$$\mathbb{E}[\mathrm{Reg}_T] = O\big((\kappa\sigma^2 \rho_D + LA_{\max}^2)\, r\sqrt{T} \log T \big).$$

The bound is on expected regret and is proved in Section F; the complement of the detection event of Lemma 17 has probability at most $1/T$ and is charged by clipping. The algorithm knows $U, \sigma, r, T, A_{\max}$ but *not* Δ_0, K_T, μ. The delay carries a single $\log T$: the tail level $u = \log(8T^5)$ pays for the union over all tested triples, and the noncentral χ^2 lower tail at that level needs only $n \gtrsim r\sigma^2 \log T/(\mu\Delta_0^2)$ post-change samples. The transient term charges every change the worst per-round loss for the delay of the smallest jump; resolving both by the size of each change removes the $A_{\max}^2/\Delta_0^2$ inflation.

Theorem 5 (Jump-size-resolved transient)**.** *Under the assumptions of Theorem 4, let* $\Delta_k = \|a_{\tau_k} - a_{\tau_k - 1}\| \geq \Delta_0$ *and* $D_k = C_D(n_0 + r\sigma^2 \log T/(\mu\Delta_k^2))$*. Then*

$$\mathbb{E}[\mathrm{Reg}_T] \leq \sum_{k=1}^{K_T} \min\{2LA_{\max}^2,\, 81L\kappa^2\Delta_k^2\}\,(D_k + \Delta_c) + 2LA_{\max}^2 + (K_T{+}1)\Big(2LA_{\max}^2 n_0 + 2\kappa\sigma^2 r \log \tfrac{2eT}{(K_T+1)n_0}\Big),$$

and hence, for a universal constant C, $\mathbb{E}[\mathrm{Reg}_T] \leq C(K_T{+}1)\big(\kappa^3\sigma^2 r \log T + LA_{\max}^2(n_0 + \Delta_c)\big)$*, which does not depend on the jump sizes.*

During the delay the least-squares bias is at most $9\kappa\Delta_k$ on a two-sided covariance event, and Lemma 17 gives delay D_k for every jump size; Section F.1 gives the proof.

4.3 Lower-Bound Comparison

Corollary 6 (Rate optimality on separated well-spaced jumps)**.** *Fix* $(r, \sigma, \mu, L, A_{\max})$ *and* $\Delta_0 \leq A_{\max}/2$*, let* $K_T = \Theta(\sqrt{T})$*, and restrict to equipartitions with* $T/(K_T+1) \geq n_0 + 2(D + \Delta_c)$ *and adjacent segment coefficients separated by at least* Δ_0*. The separated construction of Section G.1 keeps the lower bound of Theorem 8 of order* $\sigma^2 r\sqrt{T}\log T$ *on this subclass, so the ratio of the upper bound of Theorem 4 to the minimax lower bound is*

$$O\left(\kappa\rho_D + \frac{LA_{\max}^2}{\sigma^2}\right),$$

which does not depend on T*. Hence on this class the adaptive-window learner is minimax rate-optimal in* T*. By Theorem 5, the ratio is also* $O(\kappa^3 + LA_{\max}^2/\sigma^2)$*, uniformly over every* $\Delta_0 \leq A_{\max}/2$ *for which the class is nonempty.*

At the smallest admissible $\Delta_0 \asymp T^{-1/4}(r\sigma^2 \log T/\mu)^{1/2}$ the bound of Theorem 4 is of order $LA_{\max}^2 T$, while that of Theorem 5 stays $O(\sqrt{T}\log T)$. The factor $LA_{\max}^2/\sigma^2$ comes from the detection floor, cadence and burn-in, low-noise costs that the lower bound does not expose. Under the same protocol, summing follow-the-leading-history interval guarantees after truncating the Gaussian responses gives order $(K_T{+}1)\, r(LA_{\max}^2 + \sigma^2)\log^2 T$, one logarithm more at fixed κ; this compares routes of guarantees, not the actual regret of adaptive-regret methods, as Section F.1 derives. The adaptive rule also separates from fixed windows at every sublinear jump rate:

Corollary 7 (Separation from every fixed window)**.** *Let* $\Sigma = I_d$*,* $\sigma^2 \leq 12A_{\max}^2$*,* $0 < \Delta_0 \leq A_{\max}$*,* $\gamma \in (0,1)$*,* $K_T = \lfloor T^\gamma \rfloor$*, and let* $\mathcal{C}_T$ *be the coefficient paths with at most* K_T *changes of size at least* Δ_0 *spaced at least* $\lfloor T/(K_T{+}1)\rfloor$ *apart. For fixed* $(r, \sigma, A_{\max}, \Delta_0, \gamma)$*, the learner of Definition 3 has* $\sup_{\mathcal{C}_T}\mathbb{E}[\mathrm{Reg}_T] = O(T^\gamma \log T)$ *without knowing* K_T*, whereas fixed-window least squares with any width* $w \geq r + 2$*, even one chosen with knowledge of* γ*, has* $\sup_{\mathcal{C}_T}\mathbb{E}[\mathrm{Reg}_T] = \Omega(T^{(1+\gamma)/2})$*.*

The ratio $T^{(1-\gamma)/2}/\log T$ grows polynomially for every $\gamma < 1$; the proof in Section E combines Theorem 4 and Proposition 16, and Section J contrasts the analysis with the closest methods.

5 Minimax Lower Bound

Theorem 8 (Minimax lower bound for piecewise-constant regression)**.** *Consider the linear-Gaussian model with* known, fixed *subspace* U*:* $y_t = a_t^\top z_t + \varepsilon_t$*,* $z_t \sim \mathcal{N}(0, \Sigma_z)$*,* $\mu I_r \preceq \Sigma_z \preceq LI_r$*,* $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$*, with* a_t *piecewise-constant having* K *change-points, and let* $c_0 = 10r\sigma^2/(\mu A_{\max}^2)$*. Then for any partition with segment lengths* $J_0, \ldots, J_K$*:*

$$\inf_{\mathcal{A}} \sup_{\substack{a_0,\ldots,a_K \in \mathbb{R}^r \\ \|a_k\| \leq A_{\max}}} \mathbb{E}[\mathrm{Reg}_T(\mathcal{A})] \;\geq\; \frac{\sigma^2 r}{2}\sum_{k=0}^{K} \log\left(1 + \frac{J_k}{c_0}\right), \tag{5}$$

where the infimum is over all *online algorithms* $\mathcal{A}$*, even when the partition is revealed. The adversary may choose the equipartition* $J_k = T/(K{+}1)$*; for fixed* $(r, \sigma, \mu, A_{\max})$*,* $K = \Theta(\sqrt{T})$ *and* $T/(K+$

$1) \geq \max\{c_0, 1\}^2$ *the minimax regret over partitions and coefficients is at least* $\frac{\sigma^2 r(K+1)}{4} \log \frac{T}{K+1} = \Omega(\sigma^2 r\sqrt{T}\log T)$.

The same order holds on well-spaced equipartitions whose adjacent coefficients are separated by any $\Delta_0 \leq A_{\max}/2$, the class of Corollary 6; Section G.1 proves this with 640 in place of 10.

Proof idea. Regret splits into $K+1$ per-segment estimation problems; a smooth, compactly supported product prior removes the van Trees boundary term (Van Trees, 1968; Gill & Levit, 1995), and at the n-th round of a segment the Bayes risk is at least $r\sigma^2/(2(n-1+c_0))$, as shown in Section G.

6 Extension to Unknown Subspaces

For isotropic covariates the spectral moment $\mathbb{E}[y_t^2(x_t x_t^\top - I_d)] = 2\theta_t^* \theta_t^{*\top}$ (Tripuraneni et al., 2021) identifies $\mathrm{col}(U)$ once the coefficients have excited it.

Assumption 4 (Cumulative excitation)**.** *There are* t_e *and* $\mu_a > 0$ *such that* $\lambda_{\min}\big(|O_t|^{-1}\sum_{s\in O_t} a_s a_s^\top\big) \geq \mu_a$ *for all* $t \geq t_e$*, with* O_t *the odd rounds before* t*.*

Definition 9 (GLR with a learned subspace, GLR-U)**.** *Odd rounds are subspace rounds and even rounds regression rounds. At the epoch starts* $T_j = 2^j t_s + 1$*, with* $j \geq 0$ *and* t_s *even,* $\hat{U}_j$ *holds the top-*r *eigenvectors of* $\hat{M}_j = |O_{T_j}|^{-1}\sum_{s\in O_{T_j}} y_s^2(x_s x_s^\top - I_d)$*. In epoch* j *the learner predicts* $\hat{U}_j\hat{a}$*, where* $\hat{a}$ *is the prediction of the learner of Definition 3, with* $\bar{\sigma}$ *in place of* σ *in Equation* (3)*, run on all regression rounds with features* $\hat{U}_j^\top x_s$ *and replayed at each epoch start, where all replays together take at most three times as many learner steps as one pass; before* T_0 *it predicts* 0*.*

Theorem 10 (Adaptive learning of a fixed unknown subspace)**.** *Let* $x_t \sim \mathcal{N}(0, I_d)$ *and* $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ *be independent and* $\theta_t^* = U a_t$ *with* U *fixed and unknown, for a deterministic path with* $\|a_t\| \leq A_{\max}$*,* K_T *jumps of size at least* Δ_0 *and Assumption 4. Let* $\bar{\sigma}^2 \geq \sigma^2 + \epsilon^2 A_{\max}^2$ *with* $\epsilon \in (0, \frac{1}{2}]$*,* h *and* Δ_c *as in Definition 3, and let every segment last at least* $2n_0 + 4(\bar{D} + \Delta_c) + 1$ *rounds, where* $\bar{D} = C_D\big(n_0 + r\bar{\sigma}^2 \log T/((1-\epsilon^2)\Delta_0^2)\big)$*. With* $v = A_{\max}^2 + \sigma^2$*,* $u = \log(8dT^3)$*,* $C_U = 208\, r v^2 (d+5) u/\mu_a^2$*,* $R_U = 8\sqrt{r}\, v u^2 (2d+3u)/(3\mu_a)$ *and* $T_\star = \max\{t_s + 1,\, 2t_e,\, 16C_U/\epsilon^2 + 1,\, 8R_U/\epsilon + 1\}$*, GLR-U satisfies, for* $T \geq 4$ *and* $2 \leq t_s \leq T$*,*

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_T] \leq\ & 4A_{\max}^2 K_T\big(\bar{D} + \Delta_c + \tfrac{1}{2}\big) + 4A_{\max}^2 (K_T+1) n_0 + 2(K_T+1)\bar{\sigma}^2 r \log \tfrac{eT}{(K_T+1)n_0} \\ & + 2A_{\max}^2 C_U \log_2 \tfrac{2T}{t_s} + 2A_{\max}^2 \epsilon R_U + 2A_{\max}^2 T_\star + 3A_{\max}^2. \end{aligned} \tag{6}$$

The first line has the form of Equation (4) at $\kappa = 1$ and noise level $\bar{\sigma}$; the second is the cost of learning U. For fixed $(r, \sigma, A_{\max}, \Delta_0, \mu_a, \epsilon)$, $K_T = \Theta(\sqrt{T})$ and $t_e, t_s = O(\sqrt{T})$ the bound is $O\big((\bar{\sigma}^2\rho_D + A_{\max}^2) r\sqrt{T}\log T + rd\log^2(dT) + \log^3(dT)\big)$: the adaptive order of the known subspace plus a cost of learning U, including the warm-up, that is linear in d and free of K_T. GLR-U uses $(\sigma, A_{\max}, r, T)$ but not U, μ_a, t_e, Δ_0 or K_T, and adds $O(d^2)$ work per subspace round and one eigendecomposition per epoch. The proof in Section I combines a matrix Bernstein bound for $\hat{M}_j$ with the Davis–Kahan theorem; given the subspace rounds, the regression rounds seen through $\hat{U}_j$ follow the known-subspace model with a noise level that changes only at the change points, for which Lemma 17 still holds.

Proposition 11 (Cost of learning the subspace)**.** *For every* $d > r$*, even* J *and* $K + 1 \geq 2r$*, with* $T = (K+1)J$*, there is a family of isotropic instances with a fixed unknown* U*,* $\|a_t\| \leq A_{\max}$*, changes of size at least* $A_{\max}/5$ *at the rounds* $kJ + 1$*, and Assumption 4 with* $\mu_a = A_{\max}^2/(250r)$ *and* $t_e = 2rJ + 1$*, hence within the scope of Theorem 10 once* J *meets its spacing condition, on which every algorithm has*

$$\sup \mathbb{E}[\mathrm{Reg}_T] \geq \frac{rd\sigma^2}{2}\log\Big(1 + \frac{J}{c_d}\Big) + \frac{(K+1-r)\, r\sigma^2}{2}\log\Big(1 + \frac{J}{c_r}\Big), \tag{7}$$

where $c_d = 160 rd\sigma^2/A_{\max}^2$ *and* $c_r = 2560 r^2\sigma^2/A_{\max}^2$*.*

Both terms of Equation (6) are thus necessary; for fixed $(r, \sigma, A_{\max})$, $K_T = \Theta(\sqrt{T})$ and $J \geq c_d^2$, hence $d = O(T^{1/4})$, the bounds agree in K_T and T, and in d up to $\log(dT)$. Since Theorem 8 with $r = d$ gives $\Omega(\sigma^2 d\sqrt{T}\log T)$ for unstructured switching regression, a fixed rank-r subspace replaces $d\sqrt{T}\log T$ by $r\sqrt{T}\log T + rd\log^2(dT)$.

Blockwise moving subspaces. If the subspace may move between consolidation blocks $I_m = \{(m-1)B+1, \dots, mB\}$, with $U_t = U_m$ for $t \in I_m$, we pair fixed-window OLS with Fantope OGD on the moment.

Assumption 5 (Block excitation)**.** *Within each block, $\frac{1}{B}\sum_{t\in I_m} a_t a_t^\top \succeq \mu_a I_r$ for some $\mu_a > 0$.*

Theorem 12 (Unknown-subspace fixed-window sufficiency)**.** *Assume isotropic covariates and noise as in Theorem 10, exact low rank $\theta_t^* = U_m a_t$ for $t \in I_m$ with $\|a_t\| \le A_{\max}$, coefficient and subspace paths that are deterministic and independent of the observations, and Assumption 5. Let $2r+2 \le w \le B$. In block m, predict zero for the first w rounds, then use OLS of width w on $z_t = \hat{U}_m^\top x_t$ from the previous w rounds of the block, clipped to radius $A_{\max}$, where $\hat{U}_m$ spans $\hat{P}_m = \mathrm{Top}_r(X_m)$; start from any X_1 in the Fantope*

$$\mathcal{F}_r = \big\{X \in \mathbb{R}^{d\times d} :\ X = X^\top,\ 0 \preceq X \preceq I,\ \mathrm{tr}(X) = r\big\}, \tag{8}$$

and at the block end set $X_{m+1} = \Pi_{\mathcal{F}_r}(X_m + \eta \hat{M}_m)$ with $\hat{M}_m = \frac{1}{2B}\sum_{t\in I_m} y_t^2\big(x_t x_t^\top - \|x_t\|^2 I_d/d\big)$. Let $M = \lfloor T/B \rfloor$, $V_s = \sum_{m=1}^{M-1} \big\|U_{m+1}U_{m+1}^\top - U_m U_m^\top\big\|_F$, $R = r + \sqrt{2r}\,V_s$, $C_G = 14(A_{\max}^4 + \sigma^4)$ and $G^2 = A_{\max}^4 + C_G d^2/B$. Then, for every $\eta > 0$,

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_T] \le \frac{\sigma^2 rT}{2(w-r-1)} + 2A_{\max}^2 K_T w + \frac{A_{\max}^2 Mw}{2} + \frac{8A_{\max}^2 B}{\mu_a}\Big(\frac{R}{\eta} + \frac{\eta}{2}MG^2\Big) \\ + 2A_{\max}^2(T - MB).\end{aligned} \tag{9}$$

With $\eta = \sqrt{2R/(MG^2)}$, $B \asymp \sqrt{T}$, $w \asymp T^{1/4}$, $K_T = \Theta(\sqrt{T})$ and bounded V_s this is $O(T^{3/4})$, the order of Theorem 2, with d only in an $O(d\sqrt{T})$ matrix-noise term; Section H gives the proof and Section C discusses the conditions.

7 Synthetic Validation

Lower rungs: decaying steps versus a fixed window. With $d = 50$, $r = 4$, $\sigma = 0.1$, isotropic x_t and known U, the coefficient is replaced at times k^2 by a random vector of norm 2, over 20 seeds. "Mono" is a member of the class of Theorem 1: steps $a \leftarrow a - ct^{-\alpha}(z_t^\top a - y_t)z_t$; all 12 cells of the grid $c \in \{\frac{1}{2}, \frac{1}{4}, \frac{1}{10}, \frac{1}{20}\}$, $\alpha \in \{\frac{1}{2}, \frac{3}{4}, 1\}$ have higher regret than Fixed at every horizon, and Figure 2 reports the lowest, at $c = \alpha = \frac{1}{2}$. Fixed, with $w = \max\{5, \lfloor T^{1/4} \rfloor\}$ and a 10^{-3} ridge, is a heuristic variant of the rule of Theorem 2. The Mono/Fixed ratio grows from 1.16 to 11.25, with Fixed lower on 20/20 seeds for $T \ge 5{,}000$, and Section A shows that the fixed window's advantage also holds against Adam, AdaGrad and normalized LMS, under Student-t_5 and handwritten-digit covariates (Alpaydin & Kaynak, 1998), and up to $T = 200{,}000$.

Top rung: the analyzed GLR rule, algorithm-matched. All methods receive the same stream $z_t \sim \mathcal{N}(0, I_4)$, clip radius $A_{\max} = 1$ and zero start; $\lfloor \sqrt{T}/2 \rfloor$ equally spaced changes redraw the coefficient on the unit sphere with jumps of at least 0.5; $\sigma = 0.1$, $T \in \{5, \dots, 100\}$k, 20 test and 5 disjoint tuning streams per horizon; the endpoint is $\sum_t \|\hat{a}_t - a_t\|^2/2$. GLR-1 is Definition 3 exactly, untuned, with unregularized RSS, $h = h_{\min}$, discard-all reset and $\Delta_c = 1$. The runs are *algorithm-matched*, not bound-certified: their segments of 138–628 rounds are far shorter than the spacing that the explicit proof constants certify, as Section A.1 quantifies, so they measure the rule's delays and cost directly. FLH (Hazan & Seshadhri, 2009) mixes clipped online ridge experts with ridge 1 or with ridge tuned jointly with the mixing rate, and recursive least squares (RLS) is fixed-discount forgetting least squares; Section A shows that every tuned value is interior to its grid.

Table 1 shows that at $T = 50$k and 100k GLR-1 has the lowest regret after the oracle, and every paired 95% interval of GLR-1 minus a baseline lies below zero; against the strongest baseline, FLH with tuned ridge, the paired difference is -15.6 $[-22.8, -8.5]$ at 50k and -66 $[-74, -58]$ at 100k. Its log–log slope 0.51 is next to the oracle's 0.50, with every slope-difference interval below zero. Its excess over the oracle is a transient of 2.1–2.3 per change; with all jumps of one size $\Delta \in [0.05, 0.8]$ it stays at most 1.2 per change, while for $\Delta \in [0.2, 0.8]$ the delay grows as Δ^{-2}, the pairing behind Theorem 5. Per 100k stream GLR-1 takes 55 s, against 12 s for FLH with tuned ridge and 3 s for RLS, so it is the low-regret end of the regret–time trade-off; checking every second round, $\Delta_c = 2$,

Table 1: Algorithm-matched known-U comparison at $\sigma = 0.1$: mean pseudo-regret $\pm$ s.e. over 20 test streams, paired streams on which GLR-1 is lower, and log–log slope over $T \in \{5, 10, 20, 50, 100\}$k.

Method	$T = 50$k	GLR-1 lower	$T = 100$k	GLR-1 lower	slope
GLR-1	**601.0**$\pm$2.5	—	**861.6**$\pm$4.2	—	**0.51**
Tuned fixed window	751.2$\pm$7.1	20/20	1 134.0$\pm$8.1	20/20	0.62
Theorem 2 fixed window	1 060.8$\pm$10.8	20/20	1 730.0$\pm$14.7	20/20	0.70
FLH, ridge 1	766.1$\pm$4.7	20/20	1 164.5$\pm$5.5	20/20	0.57
FLH, tuned ridge	616.6$\pm$4.2	18/20	927.8$\pm$4.7	20/20	0.56
RLS	631.7$\pm$4.8	18/20	964.8$\pm$6.8	20/20	0.62
Constant-step OGD	742.4$\pm$5.8	20/20	1 166.5$\pm$8.7	20/20	0.60
Periodic restart	3 576.9$\pm$21.1	20/20	5 968.2$\pm$24.5	20/20	0.74
Change-time oracle	348.5$\pm$0.2	—	495.9$\pm$0.2	—	0.50

roughly halves its time at 11% more regret. In Section A.1, with 0–10 changes and ranks 4–16 it has the lowest non-oracle regret whenever the stream changes and declares no change on change-free streams. In Sections A.4 and A.5, with a fixed unknown U, a change every 150 rounds and $\sigma = 0.1$, GLR-U has 26–36$\%$ lower regret than the tuning-selected full-dimensional comparator for $d \in \{50, 100, 200\}$, every paired interval below zero; the gain is realized when changes are frequent enough to amortize the one-time cost of learning U, and Section K lists every registered condition. With $\lfloor\sqrt{T}/2\rfloor$ changes, the sine of the largest principal angle of its subspace estimate falls with T in every condition of Section A.4, from 0.67 to 0.33 at $d = 100$ and $\sigma = 0.1$, and its regret grows with log–log slope 0.32–0.67, below the fixed-window exponent $3/4$. In Section A.3, with fixed or rotating subspaces the learner of Theorem 12 has 29–67$\%$ lower regret than tuned full-dimensional sliding-window ridge at $T \geq 40$k and $d \leq 100$.

Beyond the algorithm-matched generator: suffix retention. GLR-S runs the scan of Definition 3 on at most $1{,}024$ candidate splits with threshold $m\, h_{\min}$ and, at a declaration, keeps the samples from the estimated split on, as in Corollary 20 with $g = 0$; it is evaluated outside Theorems 4 and 5. The preregistered stress test of Section A.2 has $r = 4$ and $\lfloor\sqrt{T}/2\rfloor$ changes, equally spaced with jumps of at least 0.5 on path P1 or at random times with chords in $\{0.1, 0.5, 1\}$ on path P2, three covariate designs and σ from 0.1 to 1; m and all baselines are tuned on 5 streams and tested on 20. GLR-S has 17–29$\%$ lower regret on P1 and 17–37$\%$ on P2 than the registered competitor with the lowest tuning regret, pruned FLH or RLS, in all 28 conditions with $T \in \{50, 100\}$k, with every paired 95% interval below zero and 20/20 wins in Table 5; in Table 4 its mean is also below that of a fixed-share learned-discount forecaster after Jacobsen & Cutkosky (2024) in all eight tested conditions, by 29–40$\%$ on 20/20 streams at $\sigma = 0.5$, with the paired interval below zero in seven conditions; the 17–37$\%$ range refers to the registered competitors. The gain comes from the retained suffix, which replaces zero-prediction burn-in rounds and lengthens clean windows: at an equal multiplier and scan set, GLR-S is 4.8–61.3% below the discard-all rule in all 28 conditions of Tables 6 and 7.

Conclusion and scope. The known-subspace coefficient-drift ladder is $\Omega(T) \to \Theta(T^{3/4}) \to O(\sqrt{T}\log T)$. By Theorem 5 and Corollary 6, the top rung matches the $\Omega(\sqrt{T}\log T)$ lower bound on separated well-spaced jumps up to a factor depending only on the condition number and the signal-to-noise ratio, uniformly in the jump size, and by Corollary 7 it separates polynomially from every fixed window at every sublinear jump rate; by Theorem 10 and Proposition 11, with a fixed unknown subspace the top rung is kept at an additive cost linear in rd that is necessary up to a logarithmic factor. All guarantees are for linear-Gaussian streams with exact low rank, exogenous coefficient paths, well-spaced jumps and known σ and $A_{\max}$, with isotropic design when U is unknown; LoRA is motivation only. In the algorithm-matched experiments, run at horizons where the proof constants are not yet active, the exact GLR rule has lower regret than every tuned baseline at $T \geq 50$k; suffix retention and a learned subspace under frequent changes extend the gains beyond the analyzed generator. Sharper detection constants, an incremental scan that keeps the guarantee, and extensions to an unknown noise level, approximately low-rank targets, anisotropic designs with unknown U and data-dependent change paths are natural next steps.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to formulate questions for surveys or interviews or to transcribe recordings, and support of qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical. Its experiments use synthetically generated data streams and the publicly available handwritten-digits dataset (Alpaydin & Kaynak, 1998) distributed with scikit-learn (Pedregosa et al., 2011); no human subjects are involved, no personal or private data are collected or used, and no new dataset is released. The contribution concerns regret guarantees for streaming low-rank regression under drifting coefficients, and we do not foresee a direct harmful application, dual use, privacy or security risk, or fairness concern specific to this work. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The problem setting and all assumptions are stated in Section 3 and Section B, and the adaptive algorithm is given in Section 4.1. Complete proofs of all results are provided in the appendix, starting with Section D. The experimental setup is described in Section 7, with further details, including data processing, in Section A, and the registered tests are listed in Section K. The code of every learner and experiment, the recorded per-seed outputs, and the scripts that recompute the reported numbers are provided in the anonymous supplementary material.

REFERENCES

Ethem Alpaydin and Cenk Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

Pierre Alquier, The Tien Mai, and Massimiliano Pontil. Regret bounds for lifelong learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 261–269, 2017.

Jushan Bai. Estimating multiple breaks one at a time. *Econometric Theory*, 13(3):315–352, 1997. doi: 10.1017/S0266466600005831.

Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pp. 3174–3182, 2013. URL https://papers.nips.cc/paper/5132-the-fast-convergence-of-incremental-pca.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015. doi: 10.1287/opre.2015.1408.

Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. doi: 10.1109/TAC.2013.2254619.

Xinyuan Cao, Weiyang Liu, and Santosh Vempala. Provable lifelong learning of representations. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6334–6356, 2022.

Tuan Quang Dam. Optimal regret bounds via low-rank structured variation in non-stationary reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=ccHq9OScXW.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1405–1411, 2015.

Tomer Gafni, Garud Iyengar, and Assaf Zeevi. The cost of learning under multiple change points. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=xQLcklRDfM.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory (ALT)*, volume 6925 of *Lecture Notes in Computer Science*, pp. 174–188, 2011. doi: 10.1007/978-3-642-24412-4_16.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: A Bayesian Cramér-Rao bound. *Bernoulli*, 1(1–2):59–79, 1995. doi: 10.2307/3318681.

Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Eric C. Hall and Rebecca M. Willett. Dynamical models and tracking regret in online convex programming. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pp. 579–587, 2013.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3–4):157–325, 2016. doi: 10.1561/2400000013.

Elad Hazan and C. Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 393–400, 2009. doi: 10.1145/1553374.1553425.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

Annika Hüselitz, Housen Li, and Axel Munk. Online jump and kink detection in segmented linear regression: Statistical optimality meets computational efficiency. *Journal of Time Series Analysis*, 47(3):727–748, 2026. doi: 10.1111/jtsa.70035.

Andrew Jacobsen and Ashok Cutkosky. Online linear regression in dynamic environments via discounting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21083–21120, 2024.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *Proceedings of Machine Learning Research*, pp. 398–406, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998. doi: 10.1109/18.737522.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Margherita Lazzaretto, Jonas Peters, and Niklas Pfister. Invariance-based dynamic regret minimization. *arXiv preprint arXiv:2603.03843*, 2026. URL https://arxiv.org/abs/2603.03843.

Oleg V. Lepski, Enno Mammen, and Vladimir G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25(3):929–947, 1997. doi: 10.1214/aos/1069362731.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982. doi: 10.1007/BF00275687.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10434–10443, 2021.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Harry L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley & Sons, New York, 1968.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Vincent Q. Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2670–2678, 2013.

Xiwen Wei, Guihong Li, and Radu Marculescu. Online-LoRA: Task-free online continual learning via low rank adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6634–6645, 2025.

Yi-Ching Yao. Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *The Annals of Statistics*, 15(3):1321–1328, 1987. doi: 10.1214/aos/1176350509.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008.

Yu-Jie Zhang, Peng Zhao, and Masashi Sugiyama. Non-stationary online learning for curved losses: Improved dynamic regret via mixability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77044–77068, 2025.

Zifeng Zhao, Feiyu Jiang, Yi Yu, and Xi Chen. High-dimensional dynamic pricing under nonstationarity: Learning and earning with change-point detection. *Management Science*, 2026. doi: 10.1287/mnsc.2023.00889. Articles in Advance; arXiv:2303.07570.

Huozhi Zhou, Lingda Wang, Lav R. Varshney, and Ee-Peng Lim. A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6933–6940, 2020. doi: 10.1609/aaai.v34i04.6176.

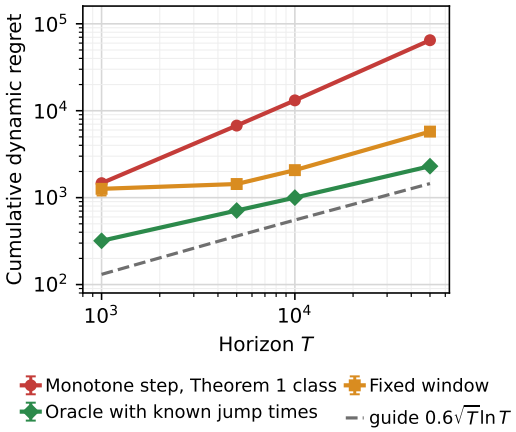
Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pp. 928–936, 2003.

A Additional Experimental Details

Variant map. Table 2 states how every learner used in the experiments relates to the theorems. Throughout, "regret" is the cumulative excess population loss $\sum_t \frac{1}{2}(\hat{a}_t - a_t)^\top \Sigma_z (\hat{a}_t - a_t)$ or its unbiased observed-loss version, as stated per experiment.

Table 2: Relation of each learner in the experiments to the theorems.

Variant	Where reported	Status
GLR-1: Definition 3 with $\Delta_c = 1$, unregularized RSS, $h = h_{\min}$, discard-all reset	Table 1	algorithm of Theorem 4; the proof's constants are not set by the experiment
GLR-1 with cadence $\Delta_c \in \{2, \dots, 32\}$	Section A.1	algorithm of Theorem 4
GLR-1 scanning at most 1,024 equispaced splits	Section A.1	capped scan; not covered by Theorem 4
Full-scan GLR with a 10^{-3} ridge in the RSS, discard-all reset	§7, this appendix	ridge surrogate of Equation (3); diagnostic
Fixed window, clipped OLS with tuned width or $w \asymp \sqrt{T/(K+1)}$	Table 1	algorithm of Theorem 2
Fixed window, OLS with 10^{-3} ridge, unclipped	Figure 2, this appendix	heuristic variant of the rule of Theorem 2 with $w = r + 1$ at $T = 1$k; not covered by its guarantee
Mono: coefficient-only sample-gradient steps on $z_t = U^\top x_t$	Figure 2	update class of Theorem 1; the stream is not its adversarial construction
Implementation Mono: also updates U, normalized steps, QR retraction	robustness checks below	implementation diagnostic
Blockwise Fantope OGD with projected fixed-window OLS	Table 3	algorithm of Theorem 12
FLH with ridge 1 or tuned ridge, RLS, OGD, periodic restart; oracle	Table 1	baselines; oracle knows change times
Spectral-moment tracker with the coefficient rule of Theorem 12	Table 9	low-rank baseline
GLR-S: tuned threshold $m\,h_{\min}$, window restarted at the estimated split	Table 5	modified rule, evaluated empirically; outside Theorems 4 and 5
Discard-all GLR with threshold $m\,h_{\min}$ and the capped scan	Table 6	for $m < 1$ not covered by Theorem 4; comparison arm
Learned-discount forecasters: discounted VAW or ridge experts, exponential weights, optionally fixed-share mixing	Section A.2, Table 4	baselines in the style of Jacobsen & Cutkosky (2024)
Full-dimensional sliding-window ridge and forgetting least squares	Table 8	baselines
Adam and AdaGrad with normalized gradients and clipping; normalized LMS	robustness checks below	baselines



T	changes	Mono	Fixed	Oracle	ratio
1 000	31	1 467	1 263	319	1.16
5 000	70	6 745	1 435	711	4.70
10 000	99	13 155	2 072	1 000	6.35
50 000	223	64 746	5 753	2 301	11.25

Figure 2: Lower rungs. Left: regret versus horizon, mean $\pm$ one standard error over 20 seeds; the dashed line is the visual guide $0.6\sqrt{T}\ln T$, not the bound of Theorem 8. Right: 20-seed mean regret, observable changes, and the ratio Mono/Fixed.

Lower-rung benchmark of Figure 2. The stream has $d = 50$, $r = 4$, $\sigma = 0.1$, isotropic Gaussian x_t and a known fixed U; at every time $k^2 \le T$ the coefficient is replaced by a fresh random vector of norm 2, which sets the norm of the new coefficient, not the jump size, giving 31, 70, 99 and

223 observable changes at $T = 1$k, 5k, 10k and 50k. The same 20 seeds are used at every horizon. Fixed uses $w = \max\{5, \lfloor T^{1/4} \rfloor\}$ and OLS with a 10^{-3} ridge, without clipping, so $w = r + 1$ at $T = 1$k; the oracle runs clipped OLS with radius 4 restarted at the true change times. Mono runs $a \leftarrow a - ct^{-\alpha}(z_t^\top a - y_t)z_t$ from $a = 0$ on $z_t = U^\top x_t$, which is a monotone-step learner of Theorem 1 for the projected problem, where $\mu = L = 1$. Mean regret of every cell of its grid:

	$\alpha = \frac{1}{2}$				$\alpha = \frac{3}{4}$				$\alpha = 1$			
T	$c{=}\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{20}$
1k	1 467	1 683	1 864	1 936	1 897	1 969	2 030	2 033	2 131	2 189	2 124	2 064
5k	6 745	8 125	9 187	9 573	9 427	9 720	9 923	9 972	10 104	10 325	10 285	10 103
10k	13 155	16 094	18 330	19 157	18 988	19 536	19 895	20 005	20 162	20 554	20 611	20 306
50k	64 746	79 730	91 146	95 433	96 097	98 048	99 335	99 833	100 202	101 396	102 370	101 399

Fixed has mean regret 1,263, 1,435, 2,072 and 5,753, below every cell. For the reported cell $c = \alpha = \frac{1}{2}$, Mono minus Fixed has paired mean 5,311 with 95% CI $[5{,}155, 5{,}466]$, 11,083 with $[10{,}826, 11{,}340]$ and 58,993 with $[58{,}450, 59{,}536]$ at $T = 5$k, 10k and 50k, on 20/20 seeds each.

Theorem-matched comparison of Table 1. The coefficient path has $K = \lfloor \sqrt{T}/2 \rfloor$ changes at rounds $1 + kJ$, $J = \lfloor T/(K+1) \rfloor$; each segment coefficient is drawn uniformly on the unit sphere and redrawn until it differs from the previous one by at least 0.5. Test streams use seeds 1000–1019 at every horizon; tuning uses the disjoint seeds 900–904 and selects the grid value with the lowest mean regret. GLR-1 uses candidate splits $\tau \in \{s_0 + r + 1, \ldots, t - r\}$, tests once the window has $n = t - s_0 + 1 \geq 2r + 2$ observations, predicts zero while it has fewer than $r + 2$, and uses $h = h_{\min}$ with $u = \log(8T^5)$. The fixed window of Theorem 2 uses $w = \max\{r + 3, \lceil \sqrt{T/(K+1)} \rceil\}$. FLH follows Hazan & Seshadhri (2009) with lifetime pruning, which keeps $O(\log T)$ live experts; each expert is clipped online ridge with $\lambda = 1$ started at its birth round, and the mixing rate is tuned over $\{0.5, 1, 2, 5, 10, 20, 50, 100\}$. RLS is exponentially forgetting least squares with a fixed discount $\gamma \in \{0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 0.97, 0.98, 0.99, 0.995\}$; it is a fixed-discount baseline, not the learned discounting of Jacobsen & Cutkosky (2024). OGD is projected constant-step gradient descent with $\eta \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.8\}$; periodic restart refits clipped OLS every $P \in \{25, \ldots, 1600\}$ rounds; the tuned fixed window uses $w \in \{6, \ldots, 256\}$. The FLH, RLS and OGD grids were widened after a first tuning round selected a grid endpoint, which can only help these baselines. At $\sigma = 0.1$ the selected values are $\eta = 0.2$, $\gamma \in \{0.6, 0.7, 0.7, 0.7, 0.8\}$ and FLH rate 10 at $T = 5, 10, 20, 50, 100$k, all interior to their grids. The second FLH row of Table 1 tunes the experts' ridge jointly with the mixing rate on the same tuning seeds, over rate $\{0.5, 1, 2, 5, 10, 20, 50, 100\}\times$ ridge $\{0.001, 0.01, 0.03, 0.1, 0.3, 1\}$; it selects rate 10 with ridge 0.03 at $T = 5$k and 0.01 at the other horizons, all interior. Paired differences GLR-1 minus baseline, with 95% t-intervals over 20 streams, at $T = 50$k and 100k are: tuned fixed window -150 $[-166, -134]$ and -272 $[-289, -256]$; FLH -165 $[-174, -156]$ and -303 $[-313, -293]$; FLH with tuned ridge -15.6 $[-22.8, -8.5]$ and -66.3 $[-74.4, -58.2]$; RLS -31 $[-42, -20]$ and -103 $[-117, -89]$; OGD -141 $[-153, -130]$ and -305 $[-322, -288]$. The slope intervals resample whole seed indices 4,000 times, so the dependence created by reusing seed indices across horizons is kept; the 95% intervals of the slope of GLR-1 minus the slope of a baseline are $[-0.123, -0.102]$ for the tuned fixed window, $[-0.202, -0.177]$ for the fixed window of Theorem 2, $[-0.063, -0.052]$ for FLH, $[-0.051, -0.039]$ for FLH with tuned ridge, $[-0.117, -0.096]$ for RLS, $[-0.106, -0.080]$ for OGD and $[-0.239, -0.226]$ for periodic restart.

Detection audit. A change τ_k is the first round of a new segment, and a declaration at round t scans the window $\{s_0, \ldots, t\}$, which contains τ_k iff $s_0 < \tau_k \leq t$. A declaration is a *localized detection* of τ_k if its window contains τ_k and no other change and its split satisfies $|\hat{\tau} - \tau_k| \leq r + 1$; each change can own one localized detection. Its delay is $t - \tau_k + 1$, the number of post-change observations at declaration. For GLR-1 at $\sigma = 0.1$:

T	5k	10k	20k	50k	100k	all
changes over 20 runs	700	1 000	1 400	2 220	3 160	8 480
declarations	700	1 000	1 400	2 220	3 160	8 480
windows with 0 / ≥ 2 changes	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
localized detections	695	988	1 395	2 209	3 137	8 424
delay median / 90% / max	2 / 4 / 15	2 / 5 / 16	2 / 5 / 21	2 / 5 / 14	2 / 5 / 16	2 / 5 / 21

Every declaration's window therefore contains exactly one change, which attributes each declaration to a unique change without a convention for multi-change windows.

Where the regret of GLR-1 is spent. On the test streams of Table 1 we assign each round of GLR-1 to *burn-in* when the window holds fewer than $r + 2$ samples and the prediction is zero, to *transient* when it holds at least $r + 2$ samples and contains a change, and to *estimation* otherwise, and split the oracle's regret into burn-in and estimation in the same way. Mean regret over 20 streams:

		GLR-1				oracle		
T	K	burn-in	transient	estimation	total	burn-in	estimation	transient per change
5k	35	108.0	73.6	3.2	184.9	108.0	3.2	2.10
10k	50	153.0	109.5	4.7	267.2	153.0	4.9	2.19
20k	70	213.0	155.0	7.5	375.5	213.0	7.4	2.21
50k	111	336.0	252.4	12.6	601.0	336.0	12.5	2.27
100k	158	477.0	365.6	18.9	861.6	477.0	18.9	2.31

Burn-in costs 3 per segment for both, six zero predictions of a unit-norm coefficient; GLR-1's estimation rounds cost what the oracle's cost to within 0.2, and the transient is the whole excess over the oracle: 2.10–2.31 per change, rising slowly with T as the threshold $h_{\min}$ does. Across horizons the log–log slopes are 0.49 for burn-in, 0.53 for the transient and 0.60 for estimation, the $(K_T{+}1)\log\frac{T}{K_T+1}$ term. The overall slope 0.51 of Table 1 therefore mostly counts the $K_T \propto \sqrt{T}$ changes at a near-constant cost each; the logarithmic estimation term appears in its own component, and the uniformity in the jump size rests on Theorem 5, since every jump here is at least 0.5 at $\sigma = 0.1$.

Ridge-surrogate full-scan GLR on the benchmark configuration. A full-scan GLR with a ridge surrogate is also run on the stream of Figure 2 with a change every 100 rounds, at $T = 10$k, $d = 50$, $r = 4$ and $\sigma = 0.1$ with 99 changes and 40 seeds; it uses the full-scan statistic $y^\top y - b^\top(G + 10^{-3}I)^{-1}b$ in place of the RSS, threshold $r\log(2T^3)$, unit cadence and discard-all reset. The comparator is the best window of a fixed-window grid chosen separately for each seed after the fact, which favors the comparator. Mean regret $\pm$ s.e. is $1{,}683.2 \pm 12.7$ versus $1{,}971.9 \pm 20.0$; the paired difference is -288.6 with 95% CI $[-328.5, -248.7]$, and GLR is lower on 40/40 seeds. With the event definitions above, the $3{,}960$ changes give $3{,}960$ declarations, none on a window without a change or with two changes, $3{,}946$ localized detections, and delays with median 2 and maximum 33. In a rank-one configuration at the theorem's cadence, with $d = 20$, $r = 1$, $\sigma = 1$, $\|a_t\| = 2$ with alternating sign every $2{,}000$ rounds, $T = 10$k, $\Delta_c = \lceil r \log T \rceil = 10$, threshold $r\log(2T^3)$ and 5 seeds, the same implementation has mean regret 374 versus 529 for the best common fixed window, $w = 20$ among 16 widths, is lower on every seed, and detects all four observable jumps without false alarms.

Unknown-subspace experiment for Theorem 12. Covariates are $x_t \sim \mathcal{N}(0, I_d)$ with $d \in \{20, 50\}$, $r = 4$, $A_{\max} = 2$ and $\sigma = 0.5$; blocks have length $B \in \{100, 200, 400\}$ for $T \in \{10, 40, 160\}$k. Inside each block the coefficient takes r values $A_{\max} s_j R_m e_j$, with R_m random orthogonal and s_j random signs, on pieces of length B/r, so the block second moment equals $(A_{\max}^2/r)I_r$ and $\mu_a = 1$; the logged minimum eigenvalue is 1 in every block. The subspace is fixed, $V_s = 0$, or rotates blockwise in a fixed $2r$-dimensional plane with total path length $V_s = 2$. The algorithm is that of Theorem 12 with $w = \max\{2r + 2, \mathrm{round}(T^{1/4})\}$ and the theorem step size $\eta = \sqrt{2R/(MG^2)}$, started from the true projector, the warm start, from $(r/d)I_d$, the neutral start, or from a projector orthogonal to $\mathrm{col}(U_1)$, the adverse start; the comparator oracle-U runs the same window rule with the true U_m. Table 3 reports 10 test seeds per cell. The subspace distance

decreases with T in all twelve (d, V_s, start) series, including the adverse start, and from the neutral start the regret ratio to oracle-U decreases with T in all four cells. From the warm start with $V_s = 0$ the regret is within 3.4% of oracle-U at every horizon. Over the 10 seeds, the standard errors of the entries of Table 3 are at most 0.036 for the distances, including every adverse-start entry, and 0.025 for the ratios.

Table 3: Blockwise unknown-U experiment with the theorem step size: mean final-quarter subspace distance $\|\sin\Theta\|_F$ at $T = 10\text{k} / 40\text{k} / 160\text{k}$ for three starts, and the regret ratio to oracle-U from the neutral start.

(d, V_s)	warm	neutral	adverse	neutral ratio
$(20, 0)$	.28 / .19 / .14	.46 / .22 / .14	1.63 / .34 / .15	1.41 / 1.21 / 1.10
$(20, 2)$	.73 / .51 / .30	.68 / .50 / .30	.83 / .45 / .30	1.48 / 1.32 / 1.17
$(50, 0)$	.23 / .19 / .15	.72 / .37 / .19	2.00 / 1.97 / .31	1.83 / 1.48 / 1.24
$(50, 2)$	.96 / .78 / .56	.91 / .66 / .53	1.88 / .99 / .46	1.93 / 1.60 / 1.41

Robustness of the lower-rung separation. The following checks use the implementation Mono of Table 2 and the unclipped Fixed of Figure 2. *Longer horizons.* On the first five seeds of the main pool and $T \in \{5, 10, 20, 50, 100, 200\}$k, the Mono/Fixed ratio increases monotonically: 5.71, 7.63, 10.10, 13.74, 16.91, 20.48. The local log–log slopes of the ratio between consecutive horizons are 0.418, 0.406, 0.335, 0.300 and 0.276. *Other learners.* On a $T = 10$k configuration with 99 observable changes and five seeds, where Fixed has regret $2{,}123$, the best known-subspace Adam coefficient learner over step sizes $\{0.01, 0.03, 0.1, 0.3, 1\}$ has $4{,}622$ and the best AdaGrad over step sizes $\{0.1, 0.3, 1, 3\}$ has $4{,}007$; both are fed the normalized gradient $(z_t^\top \hat{a} - y_t) z_t / (\|z_t\|^2 + 10^{-8})$ and clip the coefficient. The best full-dimensional normalized LMS, $\theta \leftarrow \theta - \eta(x_t^\top \theta - y_t) x_t / (\|x_t\|^2 + 10^{-8})$, without the rank-$r$ parameterization, tuned over $\eta \in \{0.05, 0.1, 0.2, 0.5, 1, 1.5\}$, has $15{,}068$. *Covariates.* With variance-normalized clipped Student-t_5 covariates, mean regrets for Mono/Fixed/oracle over 20 seeds are $(1{,}677, 1{,}012, 312)$, $(8{,}111, 1{,}365, 692)$, $(16{,}112, 2{,}050, 983)$ and $(79{,}057, 5{,}697, 2{,}243)$ at $T =$ 1k, 5k, 10k, 50k, so Mono/Fixed is 1.7, 5.9, 7.9 and 13.9. With real covariates from the handwritten-digits data (Alpaydin & Kaynak, 1998) of scikit-learn (Pedregosa et al., 2011), centered, projected to 50 principal coordinates, variance-normalized and resampled online, with a planted regression target, they are $(1{,}675, 1{,}283, 312)$, $(8{,}183, 1{,}351, 697)$, $(16{,}227, 2{,}040, 987)$ and $(79{,}607, 5{,}652, 2{,}262)$, and Mono/Fixed is 1.3, 6.1, 8.0 and 14.1.

Reproducibility. Every number in §7 and this appendix is computed from per-seed outputs of the runs described above. The supplementary material has four parts. The first holds the recorded per-stream outputs, namely the test regret of every arm and seed, the tuning mean of every grid value, per-declaration event records $(t, \hat{\tau}, s_0)$, with Λ for the theorem-matched runs, the true change times and the per-seed outputs of every robustness check above, together with scripts that recompute from them the means, paired intervals, slope bootstrap intervals, detection audits, robustness numbers and registered decisions reported here, together with a compact table of all 43 registered conditions of the suffix-retention study, and every tuning and test run of Section A.4. The second is a script that regenerates a recorded row of the theorem-matched, stress-test, unknown-subspace, jump-size and lower-rung studies from its seed with the final configuration, read from the recorded tuning results, together with a check that regenerates a representative row of each of these studies and of every arm of Table 1 and of the stress test and compares it with the recorded value, and a script that brings regenerated outputs of the A16 analyses into the recorded form. The third is the code of every experiment, including the lower-rung benchmark and the robustness checks above, with a document that lists, per table, the tuning and test runs in registration order and how their outputs are merged. The fourth holds the registered protocols with all amendments.

A.1 Operating Range of GLR-1 at High SNR

The detection audit above describes the regime in which Table 1 is measured: $\sigma = 0.1$, jumps of at least 0.5 in $\|\cdot\|$, so the per-sample detectability $\mu\Delta_0^2/\sigma^2$ is at least 25, and segments of $J \geq 628$ rounds at $T = 100$k. In this regime the median delay is 2 and the 90% quantile at most 5 at every

horizon, far below the segment length, and no scanned window contains two changes. The two terms of Theorem 4 that grow with K_T are therefore a short transient and one burn-in per change, as the theorem's decomposition describes, and GLR-1 runs one estimation window per segment, like the oracle. Its cost is computational: with prefix sums the unit-cadence scan costs $O(nr^3)$ per round and $O(nr^2)$ memory for a window of length n, so a stretch of n rounds without a declaration costs $O(n^2r^3)$ time, and a stream without changes $O(T^2r^3)$ time and $O(Tr^2)$ memory in the worst case; here every window stays below 800 rounds, and the mean single-thread times in seconds are

T	5k	10k	20k	50k	100k
GLR-1	0.7	1.8	4.8	26.4	52.6
pruned FLH	0.4	0.7	1.5	3.9	7.1
RLS / tuned fixed window	0.08 / 0.09	0.15 / 0.16	0.30 / 0.35	0.79 / 0.92	1.40 / 1.66

per stream; the FLH and RLS times are measured with the original tuning grids on the same test streams.

Regret, time and memory across cadences. A cadence $\Delta_c > 1$, allowed by Theorem 4, divides the number of scans by Δ_c and adds up to Δ_c delay rounds per change. On the same test streams, all settings below were timed together in one run, so their times are comparable with each other but not with the table above; memory is the size of the scan's prefix arrays at the largest window, $8\,n_{\max}(r^2 + r + 1)$ bytes, and every estimate uses $O(r^2)$ memory. Mean regret, single-thread time in seconds and prefix memory in MB:

	$T = 50$k			$T = 100$k		
Setting	regret	time	MB	regret	time	MB
GLR-1, $\Delta_c = 1$	601.0	28.2	0.08	861.6	80.4	0.13
GLR-1, $\Delta_c = 2$	670.8	15.2	0.08	959.7	42.2	0.13
GLR-1, $\Delta_c = 4$	829.9	8.4	0.08	1 122.2	23.8	0.13
GLR-1, $\Delta_c = 8$	1 047.8	5.0	0.08	1 443.8	13.5	0.13
GLR-1, $\Delta_c = 16$	1 441.5	3.3	0.08	2 137.5	9.1	0.13
GLR-1, $\Delta_c = 32$	2 256.5	2.4	0.08	3 342.7	6.0	0.13
GLR-1, $\Delta_c = 1$, 1,024-candidate cap	601.0	27.8	0.08	861.6	80.2	0.13
FLH, ridge 1	766.1	8.4	–	1 164.5	19.4	–
RLS	631.7	1.7	–	964.8	3.7	–

Halving the scan frequency roughly halves the time; at $T = 100$k, $\Delta_c = 2$ and $\Delta_c = 4$ cost 42 and 24 s with regret 960 and 1,122, between the unit-cadence rule and the ridge-1 FLH row of Table 1. The 1,024-candidate cap used by GLR-S never binds on these streams, whose windows stay shorter than 1,024, so it returns the same regret.

Regret, time and memory against the strongest baselines. The settings below were timed together on the same test streams, so their times are comparable within this table. State memory is $8(r^2+r+1)$ bytes per stored prefix entry for GLR-1 at the peak window length and per live expert for FLH, which keeps at most 34 experts, $8(r^2 + r)$ bytes for RLS, and the same plus the w stored samples for the fixed window. Mean regret, single-thread time in seconds and state memory in kB:

	$T = 50$k			$T = 100$k		
Setting	regret	time	kB	regret	time	kB
GLR-1, $\Delta_c = 1$	601.0	22.7	82.8	861.6	55.1	130.2
GLR-1, $\Delta_c = 2$	670.8	12.4	82.7	959.7	29.3	130.0
FLH, tuned ridge	616.6	6.4	5.4	927.8	12.3	5.7
FLH, ridge 1	766.1	6.4	5.4	1 164.5	12.2	5.7
RLS	631.7	1.4	0.16	964.8	2.6	0.16
Tuned fixed window	751.2	1.6	0.64	1 134.0	3.0	0.64

The regrets and hyperparameters are those of Table 1, with the hyperparameters selected on the tuning streams. At both horizons the regret–time Pareto set consists of GLR-1 at unit cadence, FLH

with tuned ridge and RLS. GLR-1 is its lowest-regret point: at $T = 100$k it is 66, or 7.1%, below FLH with tuned ridge at 4.5 times its time, and 103 below RLS at 21 times its time; at $T = 50$k the margins are 15.6, or 2.5%, and 30.7, at 3.5 and 16 times the time. Its state is the window's prefix arrays, 0.13 MB at $T = 100$k, against 6 kB for FLH; FLH with tuned ridge and RLS are the cheaper points of the frontier.

Streams with few or no changes, and higher rank. The streams above change every 138–628 rounds. To map where each setting sits when changes are rare, the same arms are run on known-subspace streams of rank $r \in \{4, 8, 16\}$ with $K \in \{0, 2, 10\}$ equally spaced changes, with $z_t \sim N(0, I_r)$, coefficients uniform on the unit sphere, consecutive coefficients at least 0.5 apart, $\sigma = 0.1$, $T \in \{10, 20\}$k, and 3 tuning and 10 test streams per condition. GLR-1 is run untuned at unit cadence, at the coarsest cadence $\Delta_c = \lceil r \log T \rceil$ that Theorem 4 allows, and with the 1,024-candidate cap; the rate and ridge of FLH, the discount of RLS and the width of the fixed window are tuned on the tuning streams. All arms of a condition were timed in one parallel batch, so times are comparable within a row. The $T = 10$k rows, standard errors and selected values are in the supplementary material; at $T = 20$k, mean regret / time in seconds:

K	r	GLR-1	GLR-1, $\lceil r \log T \rceil$	GLR-1, cap	FLH	RLS	fixed	oracle	Pareto
0	4	3.2 / 659	3.2 / 19.0	3.2 / 204	3.2 / 4.2	3.2 / 0.9	28.8 / 1.0	3.2 / 0.8	R
0	8	5.4 / 1 920	5.4 / 22.6	5.4 / 955	6.0 / 6.1	5.4 / 1.0	45.5 / 1.2	5.4 / 0.8	R
0	16	9.7 / 5 171	9.7 / 42.7	9.7 / 3 480	13.3 / 8.8	9.7 / 1.1	64.1 / 1.2	9.7 / 0.9	R
2	4	13.8 / 237	51.0 / 6.8	13.8 / 96.8	19.2 / 4.5	36.7 / 1.0	52.5 / 1.0	9.5 / 0.8	Gk,F,R
2	8	19.8 / 635	91.4 / 7.4	19.8 / 344	29.9 / 5.9	50.2 / 0.9	71.1 / 1.1	16.0 / 0.9	Gk,F,R
2	16	33.2 / 2 202	63.7 / 13.5	33.2 / 1 449	54.2 / 9.2	85.2 / 1.2	111.1 / 1.4	29.1 / 0.9	Gk,F,R
10	4	55.0 / 66.1	232.1 / 2.4	55.0 / 59.6	69.4 / 4.7	96.0 / 0.9	109.1 / 1.1	34.5 / 0.6	Gk,F,R
10	8	80.4 / 124	430.5 / 2.2	80.4 / 132	102.7 / 5.8	129.6 / 1.0	168.4 / 1.1	58.1 / 0.8	G,F,R
10	16	127.2 / 505	862.2 / 4.1	127.2 / 450	182.7 / 9.8	219.9 / 1.1	267.5 / 1.3	105.3 / 1.0	Gk,F,R

The last column lists the regret–time Pareto set among the non-oracle arms, with G for GLR-1, Gk for GLR-1 with the cap, F for FLH and R for RLS. Whenever the stream changes, that is, for $K \in \{2, 10\}$ at both horizons and all ranks, GLR-1 has the lowest regret of all non-oracle arms, 4.4–55.5 or 9.8–38.8% below the best of FLH, RLS and the fixed window, and every run of the three GLR arms makes exactly K declarations. On streams without changes no GLR arm declares a change in any of the 60 runs, and GLR-1 matches the oracle's regret, as does RLS at $\gamma = 1$. The cost of the unit-cadence scan grows with the length of the current window, so it is largest on change-free streams, 659–5,171 s at $T = 20$k for $r = 4$–16 against 66–505 s at $K = 10$; there the cadence $\lceil r \log T \rceil$ returns the same regret 29–170 times faster, while with changes it adds up to $\lceil r \log T \rceil$ rounds of delay per change. The cap returns the regret of the full scan in every condition, with differences below 10^{-13}. In budget terms, unit-cadence GLR-1 buys the lowest regret when changes keep its windows short, at 66–505 s per 20k stream for $K = 10$ against 5–10 s for FLH; with long change-free stretches and a tight time budget, RLS, FLH or GLR-1 at a coarser cadence are the cheaper points of the frontier. This implementation rebuilds the window's prefix arrays at each scan; peak traced memory of one run in MB:

K	r	GLR-1	GLR-1, $\lceil r \log T \rceil$	GLR-1, cap	FLH	RLS	fixed	oracle
0	4	15.52	15.50	8.84	0.02	0.00	0.00	0.00
0	8	55.82	55.68	33.38	0.05	0.00	0.00	0.00
0	16	213.15	210.78	130.10	0.19	0.01	0.01	0.01
2	4	5.18	5.16	3.19	0.02	0.00	0.00	0.00
2	8	18.60	18.61	11.94	0.05	0.00	0.00	0.00
2	16	70.99	70.05	46.39	0.19	0.01	0.01	0.01
10	4	1.41	1.41	1.14	0.02	0.00	0.00	0.00
10	8	5.07	4.98	4.15	0.05	0.00	0.00	0.00
10	16	19.30	19.18	15.95	0.19	0.01	0.01	0.01

The selected values are listed in the supplementary material. RLS selects $\gamma = 1$ on change-free streams; the fixed window selects $w = 2r + 2$ in five conditions, which is the end of its grid only for $r = 4$ with $K = 10$ and $T = 10$k, because the grid also holds widths 16 and 32 below $2r + 2$ for larger r; FLH selects its smallest rate on change-free streams and in one condition with $K = 2$, and its smallest ridge 10^{-3} in three conditions.

Jump sizes down to the detectability boundary. Theorem 5 pairs the delay of each change, of order $\sigma^2 \log T/(\mu\Delta_k^2)$, with its per-round cost, of order Δ_k^2, so the cost per change does not grow as the jumps shrink. To test this pairing with the exact rule, GLR-1 is run untuned at unit cadence, with $h = h_{\min}$ and discard-all reset, on the streams of Table 1 with every jump of size exactly Δ: the first coefficient has norm $\frac{1}{2}$ and each change adds Δ times a uniform unit direction, redrawn until the norm stays at most 1. These streams have $\sigma = 0.1$, $T \in \{5, 10, 20\}$k, $K = \lfloor\sqrt{T}/2\rfloor$ equally spaced changes, segments of $J = 138$, 196 and 281 rounds and the same 20 test seeds 1000–1019. The table reports the per-change excess over the oracle, $(\mathcal{R}_T^{\mathrm{GLR}} - \mathcal{R}_T^{\mathrm{oracle}})/K$ averaged over the 20 streams, and, at $T = 20$k, the declarations per change, the median delay of the first declaration after each change and the median of delay$\times\Delta^2/\sigma^2$:

		excess per change			$T = 20$k		
Δ	Δ^2/σ^2	$T = 5$k	10k	20k	decl./change	median delay	delay$\cdot\Delta^2/\sigma^2$
0.8	64	1.13	1.11	1.19	1.00	4	256
0.4	16	0.74	0.77	0.78	1.00	10	160
0.2	4	0.60	0.67	0.70	1.00	40	160
0.1	1	−0.52	−0.48	−0.30	0.49	150	150
0.05	0.25	−0.64	−0.63	−0.59	0.21	141	35

The half-widths of the 95% intervals of the excess are at most 0.15. For $\Delta \geq 0.2$, 98–100% of the changes are declared inside their segment and 92–100% have a localized detection. From $\Delta = 0.8$ to 0.2 the median delay grows by a factor of 10 to 15, in proportion to Δ^{-2}, since delay$\times\Delta^2/\sigma^2$ stays between 144 and 256 at every horizon, while the excess per change falls from 1.1–1.2 to 0.6–0.7; a cost growing as Δ_0^{-2}, as the transient term of Theorem 4 allows, would multiply it by 16. The excess stays at most 1.19 over the whole grid, a range in which Δ^{-2} varies by a factor 256, and for $\Delta \geq 0.2$ it changes by at most 0.1 across horizons. For $\Delta \leq 0.1$ the jumps lie near the detectability boundary of these segment lengths, and the delays above are conditional on a declaration: 15–49% of the changes are declared inside their segment, the window pools segments whose adjacent coefficients differ by at most 0.1, and GLR-1 has lower regret than the oracle, whose restart at every change pays a burn-in that pooling avoids; the per-change excess is -0.30 to -0.64, with every interval below zero.

Range of the proof constants. The constants of Lemma 17 are explicit sufficient ones: $C_n = 200$, so $n_0 = 200(r + \log T)$, and $C_D = 8(16c_h + 70)$, which is 1,264 at $c_h = 11/2$. The spacing condition of Theorem 4 then asks for

$$J_{\min} \geq (1 + 2C_D)\, n_0 + 2C_D \frac{r\sigma^2 \log T}{\mu\Delta_0^2} + 2\Delta_c,$$

which for $r = 4$ and $T = 5{,}000$ is already about 6×10^6 rounds; with $K_T = \lfloor\sqrt{T}/2\rfloor$ equal segments of length about $2\sqrt{T}$, the condition requires $\sqrt{T}$ to exceed about $(1 + 2C_D) \cdot 100\,(r + \log T)$, i.e. T of order 10^{14} at $r = 4$. The experiments use segments of 138 to 628 rounds. We therefore label them *algorithm-matched*: they run the algorithm of Theorem 4 exactly, with its threshold $h_{\min}$, and measure its delays, localization and cost directly in the detection audit, the decomposition and the jump-size sweep above; they do not certify the finite-sample constants, which the proof does not attempt to optimize.

A.2 Study E5: Suffix-Retention Stress Test

Design. The study was registered before any of its test runs, as listed in Section K. All streams have $r = 4$, $A_{\max} = 1$, unit-norm coefficients, $K = \lfloor\sqrt{T}/2\rfloor$ changes and predictions made before y_t. Path P1 is the generator of Table 1 on fresh seeds. Path P2 draws the K change times uniformly without replacement, so short segments occur, and moves the coefficient along the unit sphere by a chord length drawn uniformly from $\{0.1, 0.5, 1\}$ at each change. Covariates are $\mathcal{N}(0, I_4)$, labeled Gauss.; $\mathcal{N}(0, Q\,\mathrm{diag}(\lambda)Q^\top)$ with geometric λ of condition number 10 and trace 4, labeled cond. 10; or handwritten-digit features (Alpaydin & Kaynak, 1998), centered, reduced to 50 principal coordinates, standardized and resampled online, then projected by a random known 50×4 orthonormal

U with planted responses, labeled digits. The registered grid also contains $T = 10$k and a short-excursion path family P3 of twelve-round excursions away from a fixed coefficient and back; the claims of §7 concern the 28 conditions of Table 5, and the supplementary material tabulates GLR-S against the strongest tuned competitor in all 43 registered conditions. Each condition uses tuning seeds 5000–5004 and fresh test seeds 5100–5119, disjoint from all earlier seed sets.

Methods. GLR-S scans all admissible splits while there are at most 1,024 of them and 1,024 equispaced candidates otherwise, uses $h = m\, h_{\min}$ with $m \in \{0.05, 0.1, 0.2, 0.35, 0.5, 0.75, 1\}$ tuned, and after a declaration restarts the window at the estimated split. Proposition 19 bounds its regret on the event that every retained suffix is homogeneous; GLR-S is evaluated empirically, outside the theorems. The baselines are the tuned fixed window, clipped OLS with $w \in \{6, \dots, 1024\}$; forgetting RLS with $\gamma \in \{0.5, \dots, 0.99999\}$; projected constant-step OGD with $\eta \in \{0.0005, \dots, 0.4\}$; pruned FLH with clipped online-ridge experts and log-domain weights renormalized after every update before the next expert is inserted, as registered in Amendment A11, with mixing rate $\alpha \in \{0.01, \dots, 100\}$ and ridge $\lambda \in \{0.001, \dots, 10\}$; and two learned-discount forecasters in the style of Jacobsen & Cutkosky (2024): discounted Vovk–Azoury–Warmuth experts on the discount grid $\{1 - 2^{-j}\} \cup \{1\}$ combined by exponential weights with tuned ridge and meta rate, with predictions clipped to $|\hat{y}| \le A_{\max} \|z_t\|$, and the same with each expert's coefficient projected onto the $A_{\max}$-ball. Both add a constant ridge to the discounted Gram matrix and combine the experts by one exponential-weights layer, in place of the discount-dependent initial regularization and the adaptive fixed-share combination of the original algorithm; the comparison is therefore with these two forecasters, not with learned discounting as a family. Each baseline's value is the grid point with the lowest mean tuning regret; grids were widened twice where a selection sat at an end, and for FLH and RLS every value selected in Table 5 is interior. The strongest competitor of a condition is the baseline with the lowest mean tuning regret; in all 28 conditions it is also the baseline with the lowest mean test regret and is pruned FLH or RLS. A tuning-stream diagnostic of the learned-discount forecasters is consistent with the current feature entering the discounted Vovk–Azoury–Warmuth Gram matrix and shrinking the prediction when the effective window is short. Times are single-thread wall times; in the row P1, Gauss., $\sigma = 0.1$, $T = 50$k of Table 5 the FLH time was measured separately on the same streams.

Learned discounting with fixed-share mixing. The registered descriptive study A16 adds, after the registered study E5, the fixed-share step $W \leftarrow (1 - s)W + s/E$ after every exponential-weights update to the projected learned-discount forecaster, in two versions: with the discounted Vovk–Azoury–Warmuth experts above, which use z_t in the prediction, and a protocol-adapted version whose experts are discounted ridge regressions on past samples only, so that the combined coefficient is committed before z_t, as for every other learner here. Ridge, meta rate and share rate $s \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ are tuned jointly on the tuning streams of the eight Gaussian conditions of Table 4. The fixed-share step lowers the regret of the projected forecaster in every condition, and the protocol-adapted version is the strongest of all tuned competitors at $\sigma = 0.1$. GLR-S has lower mean regret than both versions in all eight conditions: by 29–40% against the protocol-adapted version at $\sigma = 0.5$, on 20/20 streams, and by 0.9–7.6% at $\sigma = 0.1$, with the paired 95% interval below zero in three of these four conditions. The selected share rates are interior to the grid; the Vovk–Azoury–Warmuth version selects the smallest ridge of its grid in every condition.

Where the gain of GLR-S comes from. GLR-S differs from the analyzed rule in its threshold, its capped scan and its reset. The registered discard-all arm uses the same capped scan and a tuned multiplier; in 25 of the 28 conditions of Table 5 it selects the same m as GLR-S, and in the remaining three conditions it was rerun at GLR-S's m on the same test streams: two Gaussian conditions in Amendment A12, and the condition P2, digits, $\sigma = 0.1$, $T = 50$k in A16, where the two rules have mean regret 161.1 and 390.7. At an equal threshold and scan set, the two rules differ only in what they keep at a declaration, and suffix retention lowers regret by 4.8–61.3% in all 28 conditions, with every paired 95% interval below zero; the largest reductions are on P2, whose short random segments make a discard-all reset expensive. Table 6 gives the Gaussian conditions with the event records. Both rules declare at the same rounds until the first reset, and the medians of their delays agree in Table 6; after the first reset their event sequences differ. GLR-S keeps a suffix whose start is the estimated split, with median distance 2–4 samples to the change. About half of the retained suffixes begin before the change, and at $\sigma = 0.1$ GLR-S makes close to two declarations per change.

Table 4: GLR-S versus learned-discount forecasters with fixed-share mixing in study A16, with Gaussian covariates, 20 test streams and the endpoint of Table 5: mean regret of GLR-S, of the discounted Vovk–Azoury–Warmuth version VAW-FS and of the protocol-adapted version P-FS, the paired difference GLR-S minus forecaster with its 95% t-interval, and the number of the 20 streams on which GLR-S is lower.

				VAW-FS			P-FS		
Path	σ	T	GLR-S	mean	difference [95% CI]	wins	mean	difference [95% CI]	wins
P1	0.1	50k	498.5	1 843.3	$-1\,344.8$ $[-1\,360.2, -1\,329.5]$	20	503.0	-4.5 $[-15.8, 6.7]$	14
P1	0.1	100k	696.5	2 715.5	$-2\,019.0$ $[-2\,040.1, -1\,997.9]$	20	720.5	-24.0 $[-37.7, -10.4]$	15
P1	0.5	50k	968.0	2 369.0	$-1\,401.0$ $[-1\,418.1, -1\,383.8]$	20	1 521.3	-553.3 $[-567.3, -539.3]$	20
P1	0.5	100k	1393.8	3 618.3	$-2\,224.5$ $[-2\,251.4, -2\,197.7]$	20	2 318.8	-925.0 $[-940.3, -909.7]$	20
P2	0.1	50k	160.1	700.1	-540.1 $[-559.4, -520.7]$	20	165.2	-5.1 $[-9.9, -0.4]$	16
P2	0.1	100k	224.4	1 054.1	-829.6 $[-852.4, -806.9]$	20	242.8	-18.4 $[-23.5, -13.2]$	20
P2	0.5	50k	445.1	970.6	-525.5 $[-546.6, -504.4]$	20	627.9	-182.7 $[-191.3, -174.2]$	20
P2	0.5	100k	683.0	1 510.6	-827.5 $[-857.1, -798.0]$	20	985.6	-302.6 $[-317.4, -287.8]$	20

Table 5: GLR-S versus the strongest tuned adaptive competitor on the endpoint $\sum_t \frac{1}{2}(\hat{y}_t - z_t^\top a_t)^2$: mean $\pm$ s.e. over 20 test streams, the competitor and its tuned values, its mean, the paired difference GLR-S minus competitor with its 95% t-interval, the relative reduction in percent, GLR-S's tuned multiplier m, and mean wall times in seconds of GLR-S / competitor. GLR-S is lower on 20/20 streams in every row.

Path	Cov.	σ	T	GLR-S	Competitor	mean	difference [95% CI]	%	m	time
P1	Gauss.	0.1	50k	498.5±5.9	FLH α=10, λ=0.01	626.5	-128.1 $[-139.2, -116.9]$	20.4	0.35	13 / 18
P1	Gauss.	0.1	100k	696.5±10.0	FLH α=10, λ=0.01	925.7	-229.2 $[-247.6, -210.8]$	24.8	0.2	236 / 25
P1	Gauss.	0.5	50k	968.0±8.1	FLH α=3, λ=1	1 277.3	-309.3 $[-327.4, -291.2]$	24.2	0.2	77 / 12
P1	Gauss.	0.5	100k	1 393.8±10.9	FLH α=3, λ=1	1 961.1	-567.3 $[-593.7, -540.9]$	28.9	0.2	249 / 19
P1	Gauss.	1	50k	2 319.4±16.1	FLH α=1, λ=1	2 785.5	-466.1 $[-497.5, -434.7]$	16.7	0.2	80 / 13
P1	Gauss.	1	100k	3 478.4±18.7	FLH α=1, λ=1	4 366.5	-888.2 $[-923.2, -853.1]$	20.3	0.2	274 / 28
P1	cond. 10	0.1	50k	455.0±6.4	FLH α=10, λ=0.03	550.5	-95.6 $[-106.5, -84.7]$	17.4	0.2	95 / 15
P1	cond. 10	0.1	100k	654.3±8.5	FLH α=10, λ=0.01	822.6	-168.3 $[-178.4, -158.3]$	20.5	0.2	192 / 19
P1	cond. 10	0.5	50k	941.3±8.2	FLH α=3, λ=1	1 176.5	-235.2 $[-247.9, -222.5]$	20.0	0.2	101 / 10
P1	cond. 10	0.5	100k	1 385.1±14.6	FLH α=3, λ=1	1 833.7	-448.6 $[-464.3, -432.9]$	24.5	0.2	253 / 25
P1	digits	0.1	50k	490.2±7.2	FLH α=10, λ=0.03	625.3	-135.1 $[-150.2, -119.9]$	21.6	0.2	70 / 13
P1	digits	0.1	100k	682.7±7.5	FLH α=10, λ=0.01	935.2	-252.5 $[-266.5, -238.4]$	27.0	0.2	256 / 25
P1	digits	0.5	50k	979.5±8.3	FLH α=3, λ=0.1	1 291.9	-312.3 $[-333.2, -291.5]$	24.2	0.2	86 / 14
P1	digits	0.5	100k	1 408.8±9.4	FLH α=3, λ=0.1	1 993.4	-584.6 $[-602.5, -566.7]$	29.3	0.2	256 / 20
P2	Gauss.	0.1	50k	160.1±3.1	RLS γ=0.9	212.5	-52.4 $[-57.2, -47.6]$	24.7	0.35	24 / 2
P2	Gauss.	0.1	100k	224.4±3.6	RLS γ=0.9	330.2	-105.8 $[-111.2, -100.3]$	32.0	0.35	324 / 6
P2	Gauss.	0.5	50k	445.1±7.4	FLH α=3, λ=0.1	692.1	-247.0 $[-257.6, -236.4]$	35.7	0.2	196 / 14
P2	Gauss.	0.5	100k	683.0±5.6	FLH α=3, λ=0.1	1 079.8	-396.7 $[-414.2, -379.3]$	36.7	0.2	347 / 19
P2	Gauss.	1	50k	1 206.2±21.4	RLS γ=0.98	1 457.4	-251.2 $[-279.3, -223.2]$	17.2	0.2	197 / 2
P2	Gauss.	1	100k	1 898.8±19.6	FLH α=1, λ=1	2 381.4	-482.6 $[-518.9, -446.3]$	20.3	0.2	461 / 25
P2	cond. 10	0.1	50k	152.7±3.6	RLS γ=0.9	216.5	-63.8 $[-70.3, -57.2]$	29.5	0.2	146 / 2
P2	cond. 10	0.1	100k	214.4±3.7	RLS γ=0.9	333.7	-119.3 $[-125.5, -113.1]$	35.8	0.35	401 / 5
P2	cond. 10	0.5	50k	452.5±6.0	FLH α=3, λ=0.1	671.6	-219.1 $[-229.2, -209.0]$	32.6	0.2	113 / 9
P2	cond. 10	0.5	100k	679.0±6.2	FLH α=3, λ=1	1 015.9	-336.9 $[-350.9, -322.9]$	33.2	0.2	464 / 29
P2	digits	0.1	50k	161.1±3.2	RLS γ=0.9	218.1	-57.0 $[-61.9, -52.1]$	26.1	0.2	109 / 2
P2	digits	0.1	100k	216.1±3.0	RLS γ=0.9	329.4	-113.3 $[-118.3, -108.2]$	34.4	0.35	392 / 5
P2	digits	0.5	50k	453.8±5.5	FLH α=3, λ=0.1	710.9	-257.1 $[-269.4, -244.8]$	36.2	0.2	136 / 15
P2	digits	0.5	100k	687.4±6.0	FLH α=3, λ=0.1	1 091.6	-404.2 $[-418.6, -389.8]$	37.0	0.2	364 / 13

Which rounds carry the gain. The registered descriptive study A16 examines both rules on the 28 conditions and test streams of Table 5 and assigns every round to *burn-in* when the window holds fewer than $r + 2$ samples and the prediction is zero, to *contaminated* when the window holds samples of an earlier coefficient, and to *clean* otherwise. A retained suffix is contaminated if it begins before the latest change it spans. Table 7 gives the Gaussian conditions; the covariance and digit conditions follow the same pattern, and all 28 rows are in the supplementary material. Three quantities carry the gain. The first is burn-in: GLR-S spends 7–287 zero-prediction rounds per stream instead of 388–963, which saves 189–475 in regret in every condition, more than the whole gain at $\sigma = 0.1$ and 53–70% of it at $\sigma = 1$. The second is the clean rounds: at $\sigma \geq 0.5$ the retained samples lower the regret of clean rounds by 88–760. The third is contamination: 42–54% of the retained suffixes begin before the change, holding on average 3–6 earlier samples on P1 and 4–39 on P2; the contaminated rounds cost 44–518 more than under discard-all. At $\sigma = 0.1$ a contaminated suffix is short-lived: the rounds until the next declaration or change average 27–304, with 0.35–0.99 extra declarations

Table 6: Suffix retention at an equal multiplier m and scan set with Gaussian covariates: mean regret of the discard-all rule and of GLR-S, relative reduction in percent, paired 95% interval of GLR-S minus discard-all, wins of GLR-S out of 20, then declarations per change, median delay and median $|\hat{\tau} - \tau_k|$, each as discard-all / GLR-S, and the percentage of retained suffixes that begin before the change they follow.

Path	σ	T	m	discard	GLR-S	%	difference [95% CI]	wins	decl./change	delay	split error	early
P1	0.1	50k	0.35	538.6	498.5	7.4	$[-54.9, -25.3]$	18	1.00 / 1.92	1 / 1	4 / 4	52
P1	0.1	100k	0.2	731.8	696.5	4.8	$[-51.1, -19.4]$	16	1.01 / 1.99	1 / 1	4 / 4	51
P1	0.5	50k	0.2	1 157.0	968.0	16.3	$[-205.8, -172.2]$	20	1.01 / 1.23	5 / 5	2 / 2	54
P1	0.5	100k	0.2	1 679.2	1 393.8	17.0	$[-297.5, -273.4]$	20	1.01 / 1.20	5 / 5	2 / 2	52
P2	0.1	50k	0.35	379.7	160.1	57.8	$[-227.3, -212.0]$	20	0.92 / 1.32	3 / 3	3 / 3	53
P2	0.1	100k	0.35	553.8	224.4	59.5	$[-342.8, -315.9]$	20	0.94 / 1.35	3 / 3	3 / 3	52
P2	0.5	50k	0.2	692.3	445.1	35.7	$[-259.8, -234.5]$	20	0.69 / 0.74	14 / 14	2 / 2	47
P2	0.5	100k	0.2	1 062.1	683.0	35.7	$[-391.8, -366.2]$	20	0.71 / 0.75	15 / 15	2 / 2	44

per change; at $\sigma \geq 0.5$ the per-round regret after a contaminated suffix, 0.003–0.018, is close to that after a clean one, 0.003–0.015. Every paired interval of the burn-in, clean and total differences lies below zero, and every interval of the contaminated difference above zero. The event of Proposition 19 therefore describes the clean suffixes; the measured gain includes the contaminated ones, whose cost the combined burn-in and clean-round savings outweigh in every condition.

Table 7: Round-type decomposition of GLR-S versus discard-all at the same multiplier with Gaussian covariates and 20 test streams: burn-in rounds per stream as GLR-S / discard-all, mean differences GLR-S minus discard-all of the regret on burn-in, contaminated and clean rounds and in total, percentage of retained suffixes that begin before the change, their mean number of earlier samples, and extra declarations per change of GLR-S.

Path	σ	T	burn-in rounds	Δburn-in	Δcontam.	Δclean	Δtotal	contam. %	earlier	extra decl.
P1	0.1	50k	194 / 672	−244.3	205.5	−1.3	−40.1	52	3.5	0.92
P1	0.1	100k	287 / 962	−337.0	302.7	−1.0	−35.3	51	3.7	0.99
P1	0.5	50k	61 / 680	−316.7	266.5	−138.8	−189.0	54	3.4	0.23
P1	0.5	100k	78 / 962	−443.0	366.7	−209.1	−285.4	52	3.1	0.20
P1	1	50k	15 / 679	−330.6	372.0	−511.5	−470.1	44	6.0	0.05
P1	1	100k	17 / 963	−474.7	518.4	−759.5	−715.7	42	5.7	0.03
P2	0.1	50k	93 / 617	−259.7	44.7	−4.6	−219.6	53	4.6	0.38
P2	0.1	100k	127 / 893	−387.1	64.7	−7.0	−329.4	52	4.1	0.40
P2	0.5	50k	15 / 466	−222.4	67.2	−91.9	−247.2	47	20.9	0.04
P2	0.5	100k	17 / 676	−327.4	98.2	−149.8	−379.0	44	28.9	0.03
P2	1	50k	7 / 388	−189.5	122.6	−273.5	−340.4	44	34.5	0.03
P2	1	100k	7 / 571	−287.7	201.2	−456.5	−543.0	43	36.9	0.02

A.3 Study E6: Unknown Subspace Versus Full-Dimensional Regression

This comparison was registered before any of its test runs, as listed in Section K. Covariates are $x_t \sim \mathcal{N}(0, I_d)$ with $d \in \{20, 50, 100\}$, $r = 4$, $A_{\max} = 2$, $\sigma = 0.5$ and blocks of length $B = 4\lfloor \mathrm{round}(\sqrt{T})/4 \rfloor$. Inside each block the coefficient takes the values $c_j s_j R_m e_j$ for B/r rounds each, with $c = (A_{\max}, A_{\max}, A_{\max}, A_{\max})$ for full excitation with $\mu_a = 1$ or $c = (A_{\max}, A_{\max}, A_{\max}, A_{\max}\sqrt{0.1})$ for weak excitation with $\mu_a = 0.1$. The subspace is fixed, rotates blockwise as in Table 3, or rotates by the same total angle continuously, a smooth rotation outside Theorem 12. The learner of Theorem 12 starts from the neutral point $(r/d)I_d$ with η tuned; the full-dimensional baselines are sliding-window ridge regression and forgetting least squares with an initial ridge, both clipped to $\|\theta\| \leq A_{\max}$, with window or discount and ridge tuned on $w \in \{6, \ldots, 1600\}$ or $\gamma \in \{0.5, \ldots, 0.9999\}$ and ridge $\lambda \in \{0.1, \ldots, 1000\}$; tuning seeds are 6000–6004 and test seeds 6100–6119. The endpoint is the pseudo-regret $\sum_t \frac{1}{2} \left\| \hat{\theta}_t - \theta_t \right\|^2$. Table 8 reports the relative reduction of the unknown-subspace learner against each baseline; every paired 95% interval lies below zero and the learner is lower on 20/20 streams in every cell. The selected window lies inside its grid in every cell; the forgetting-least-squares selection is the grid end $\gamma = 0.9999$, $\lambda = 1000$ in 15 of the 36 cells, with the grid kept as registered. Forgetting least squares is implemented by the covariance-form recursion. Study A16 also evaluates the 150 tuning cells of condition, γ and λ with

$d = 100$ and $\gamma \in \{0.5, 0.7\}$ on the same tuning streams in the mathematically identical information form $G_t = \gamma G_{t-1} + x_t x_t^\top$, $G_0 = \lambda I$, $b_t = \gamma b_{t-1} + x_t y_t$; in each condition the best of them has 1.7–2.2 times the tuning mean of the selected configuration, so every selection and test result of Table 8 stands, and at every selected configuration the two forms agree up to rounding error.

Table 8: Unknown-subspace learner of Theorem 12 from the neutral start versus full-dimensional regression: relative regret reduction in percent against sliding-window ridge / forgetting least squares, 20 paired test streams per cell.

T	d	excitation	fixed U	blockwise rotation	smooth rotation
40k	20	full	42 / 50	38 / 47	38 / 47
40k	20	weak	33 / 46	29 / 43	29 / 43
40k	50	full	49 / 60	45 / 57	44 / 56
40k	50	weak	40 / 53	36 / 50	36 / 50
40k	100	full	44 / 50	40 / 47	40 / 47
40k	100	weak	35 / 43	31 / 39	31 / 39
160k	20	full	50 / 45	47 / 42	47 / 42
160k	20	weak	45 / 42	42 / 39	42 / 39
160k	50	full	64 / 71	61 / 68	60 / 68
160k	50	weak	58 / 69	55 / 66	55 / 66
160k	100	full	67 / 74	64 / 71	63 / 71
160k	100	weak	62 / 70	58 / 67	58 / 67

Table 9: Rotating subspaces: relative regret reduction in percent of the Fantope learner against the tuned spectral-moment tracker feeding the same coefficient rule, and final-quarter subspace distance $\|\sin\Theta\|_F$ as Fantope / tracker; 20 paired test streams per cell, every paired 95% interval below zero.

			blockwise rotation		smooth rotation	
T	d	excitation	%	distance	%	distance
40k	20	full	6.2	0.42 / 0.65	6.0	0.42 / 0.65
40k	20	weak	4.6	0.46 / 0.65	4.9	0.46 / 0.66
40k	50	full	2.6	0.60 / 0.70	2.5	0.62 / 0.71
40k	50	weak	1.6	0.68 / 0.75	1.7	0.67 / 0.74
40k	100	full	0.9	0.76 / 0.80	0.8	0.76 / 0.79
40k	100	weak	0.6	0.82 / 0.85	0.6	0.82 / 0.85
160k	20	full	9.5	0.28 / 0.34	9.4	0.27 / 0.33
160k	20	weak	10.5	0.29 / 0.38	10.6	0.30 / 0.38
160k	50	full	8.5	0.41 / 0.64	8.5	0.41 / 0.64
160k	50	weak	5.7	0.47 / 0.64	5.8	0.47 / 0.64
160k	100	full	4.5	0.49 / 0.66	4.6	0.49 / 0.67
160k	100	weak	3.5	0.52 / 0.67	3.6	0.52 / 0.67

Low-rank representation and the Fantope update. The registered study also runs a spectral-moment tracker that feeds the same coefficient rule: it estimates the projector as the top-r eigenspace of an exponentially weighted average of the trace-centered moments $\hat{M}_m$, with forgetting factor tuned on $\{0, 0.5, 0.8, 0.9, 0.95, 0.99\}$ together with the undiscounted running average, and with the same block length, window, rank, clipping and tuning seeds as the Fantope learner. Against the full-dimensional baselines this tracker is 25–67% below sliding-window ridge and 32–74% below forgetting least squares in the 36 cells of Table 8, so the gap of Table 8 is carried by the online low-rank representation, which both learners share. With a fixed subspace the tuned tracker uses the undiscounted average, which is the stationary moment that the Fantope iterate also estimates. When the subspace rotates, the Fantope update of Theorem 12 has lower regret than the tuned tracker in all 24 rotating cells, with every paired 95% interval below zero and 20/20 wins, and a smaller final-quarter subspace distance in Table 9. Mean single-thread times per stream at $T \geq 40$k are 0.8–33 s for both low-rank learners and 0.5–39 s for the full-dimensional baselines; the Fantope state is a $d \times d$ matrix, so its memory is $O(d^2)$ although the leading regret terms depend on r.

A.4 Study A18: GLR with a Learned Subspace on a Fixed Unknown Subspace

Design. The study was registered before any of its runs; its entry is in Section K. Streams have $x_t \sim \mathcal{N}(0, I_d)$, a Haar-random $U \in \mathbb{R}^{d\times 4}$ per stream, the coefficient path of Table 1, with $\lfloor\sqrt{T}/2\rfloor$ equally spaced changes, coefficients on the unit sphere and consecutive coefficients at least 0.5 apart, and $y_t = x_t^\top U a_t + \sigma\,\mathcal{N}(0,1)$, for $d \in \{20, 50, 100\}$, $\sigma \in \{0.1, 0.5\}$ and $T \in \{20, 50, 100\}$k, with 20 test streams per condition. GLR-U is Definition 9 exactly with $t_s = 64$, the scan of GLR-1 with unregularized RSS, $h = h_{\min}$, $\Delta_c = 1$ and discard-all reset, and $\bar\sigma = \sqrt{2}\,\sigma$, that is, $\epsilon = \sigma/A_{\max}$; it is untuned. The reference is GLR-1 given the true U. The endpoint is $\sum_t \frac{1}{2}\left\|\hat\theta_t - \theta_t^*\right\|^2$; the subspace part is $\sum_t \frac{1}{2}\left\|(I - \hat P_{j(t)})\theta_t^*\right\|^2$, the second term of Equation (12). As for GLR-1, the runs are algorithm-matched: the warm-up $T_\star$ certified by Theorem 10 exceeds these horizons by orders of magnitude.

Results. Table 10 reports the regret of GLR-U, its subspace part and the largest principal angle of the last subspace estimate. The subspace error falls with T in every condition, at $d = 100$ and $\sigma = 0.1$ from 0.67 to 0.33; the regret grows with log–log slope 0.32–0.67 over T, below the fixed-window exponent $3/4$ in all six (d, σ) pairs, which is check A18-2; and the subspace part accounts for 24–68% of the regret and rises with d; this is the projection loss, the second term of Equation (12), which the subspace terms of Equation (6) bound, and at these horizons it is of the same order as the rest. Table 15 in Section K lists the registered comparison with the tuned full-dimensional baselines in all 18 conditions. GLR-U takes 6–123 s per stream, between 0.4 and 2.2 times GLR-1 with the true U on the same streams.

Table 10: GLR-U of Definition 9 on a fixed unknown subspace with $r = 4$ and $A_{\max} = 1$: mean regret $\pm$ s.e. over 20 streams, its subspace part sub., and the mean sine $\sin$ of the largest principal angle between the last subspace estimate and $\mathrm{col}(U)$; log–log slope of the mean regret over T; GLR-1 with the true U at $T = 100$k for reference.

		$T = 20$k			$T = 50$k			$T = 100$k				known U
σ	d	regret	sub.	sin	regret	sub.	sin	regret	sub.	sin	slope	100k
0.1	20	3 743±87	2 122	0.31	4 917±128	2 714	0.22	6 228±85	3 339	0.14	0.32	859
0.1	50	6 565±89	3 891	0.52	9 071±167	5 577	0.34	10 878±163	6 774	0.23	0.32	863
0.1	100	9 006±61	5 573	0.67	13 851±195	8 677	0.47	17 593±259	11 428	0.33	0.42	859
0.5	20	6 906±43	2 448	0.38	12 172±85	3 396	0.27	17 847±92	4 241	0.18	0.59	4 071
0.5	50	8 269±56	4 453	0.57	15 078±101	6 845	0.41	21 803±135	8 493	0.28	0.61	4 116
0.5	100	9 325±47	6 290	0.79	18 063±98	10 529	0.56	27 210±172	14 267	0.39	0.67	4 100

A.5 Study A19: GLR-U with Frequent Changes

Design. The form of Theorem 10 suggests where learning U pays off: its representation term does not grow with the number of changes, whereas a full-dimensional learner re-estimates d coordinates after every change. The registered study A19, listed in Section K, therefore uses the streams of Section A.4 with a change every $J \in \{150, 300\}$ rounds, that is, $K_T = 665$ or 332 at $T = 100$k, with $d \in \{50, 100, 200\}$ and $\sigma \in \{0.1, 0.5\}$. GLR-U is run untuned, exactly as in Section A.4; the full-dimensional baselines, sliding-window ridge with $w \in \{30, \ldots, 500\}$ and $\lambda \in \{1, 10, 30, 100\}$ and forgetting least squares with $\gamma \in \{0.9, \ldots, 0.99\}$ and $\lambda \in \{1, 10\}$, are tuned on 5 streams and tested on 20 fresh ones; the comparator is the one with the lower tuning regret.

Results. At $\sigma = 0.1$, the setting of check A19-1, GLR-U has 26–36% lower regret than the tuned full-dimensional comparator with a change every 150 rounds for every $d \in \{50, 100, 200\}$, and 16–23% lower with a change every 300 rounds for $d \in \{100, 200\}$, with every paired 95% interval below zero and $20/20$ wins in Table 11; Table 16 in Section K lists all 12 registered conditions. The projection loss is 39–62% of the regret of GLR-U in these cells. Because the comparator is full-dimensional, these reductions measure GLR-U as a whole, the learned representation together with the GLR tracker run on it. In every cell of Table 11 the selected window and ridge of the comparator are interior to their grids.

Table 11: GLR-U against the tuned full-dimensional comparator with frequent changes, with $r = 4$, $A_{\max} = 1$, $\sigma = 0.1$, $T = 100$k and 20 paired test streams: mean regret $\pm$ s.e., projection loss sub. of GLR-U, comparator, where window is sliding-window ridge and RLS is forgetting least squares, paired difference GLR-U minus comparator with 95% interval, relative reduction in percent, wins, and GLR-1 with the true U for reference.

J	K_T	d	GLR-U	sub.	comparator	regret	difference [95% CI]	%	wins	known U
150	665	50	15 914±95	6 233	window	23 529±61	−7 615 [−7 810, −7 420]	32.4	20/20	3 612
150	665	100	21 532±141	10 706	window	33 554±47	−12 023 [−12 357, −11 689]	35.8	20/20	3 606
150	665	200	30 497±153	17 268	window	41 214±28	−10 716 [−11 024, −10 408]	26.0	20/20	3 601
300	332	100	18 295±135	10 773	window	23 619±75	−5 324 [−5 657, −4 991]	22.5	20/20	1 799
300	332	200	28 165±284	17 488	window	33 468±69	−5 303 [−5 960, −4 645]	15.8	20/20	1 799

B Assumptions and Theorem Alignment

Table 12: The adaptation rate ladder for streaming low-rank regression with $K_T = \Theta(\sqrt{T})$ coefficient jumps of size at least Δ_0, ordered from the slowest to the fastest rate. All rates are for expected dynamic regret Reg_T in the exact linear-Gaussian model with $\rho_{\max} = 0$ as $T \to \infty$ with $(r, \sigma, \mu, L, A_{\max}, \Delta_0)$ fixed; each lower bound is attained by a path with equally spaced jumps.

Rung	**Algorithm class**	**Rate**	**Reference**
1	Monotone-step, decaying η_t	$\Omega(T)$	Thm. 1
2	Fixed-window least squares of width w	$\Theta(T^{3/4})^{\ddagger}$	Thm. 2, Prop. 16
3	Known-U adaptive window, GLR + clipped OLS	$O(\sqrt{T}\log T)$	Thms. 4, 5
—	Lower bound, all algorithms	$\Omega(\sqrt{T}\log T)$	Thm. 8
ext.	GLR-U, fixed unknown U	$O(\sqrt{T}\log T)^{\S}$	Thm. 10, Prop. 11
ext.	Fixed window + Fantope OGD, blockwise U	$O(T^{3/4})^{\dagger}$	Thm. 12

‡ Upper bound for clipped OLS with $w \asymp T^{1/4}$; lower bound for every $w \geq r + 2$, clipped or unclipped. § Plus an additive $O(rd\log^2(dT) + \log^3(dT))$ cost of learning U, necessary up to the factor $\log(dT)$ for $\log(dT) \leq d = O(T^{1/4})$; isotropic design and cumulative excitation. † Same order as rung 2 with known U; isotropic design, block excitation and bounded block path length.

Assumption 6 (Low-rank decomposition)**.** *The decomposition* (1) *holds with* $1 \leq r < d$*.*

Assumption 7 (Bounded variation)**.** *The subspace path length* $V_s(T) = \sum_{t=1}^{T-1} d_{\mathrm{Gr}}(\mathrm{col}(U_t), \mathrm{col}(U_{t+1}))$ *and the coefficient variation* $V_f(T) = \sum_{k=1}^{K_T} \left\| a_{\tau_k} - a_{\tau_{k-1}} \right\|$ *are finite, with minimum jump size* $\Delta_0 > 0$*.*

Assumption 8 (Bounded coefficients)**.** $\|a_t\| \leq A_{\max}$ *for all* t*.*

Assumption 9 (Residual budget)**.** *The residual* ρ_t *is an orthogonal approximation error:* $U_t^\top \rho_t = 0$ *and* $\|\rho_t\| \leq \rho_{\max}$*. This decomposition permits discussing approximation error, but all proved rates here set* $\rho_{\max} = 0$*; extending the unknown-subspace moment argument to nonzero residuals requires controlling cross terms, not simply adding a loss term.*

Assumption 10 (Conditionally unbiased gradient oracle)**.** *Let* $\mathcal{F}_s = \sigma(x_1, y_1, \ldots, x_{s-1}, y_{s-1})$ *be the natural filtration, so that the iterate* θ_s *of a monotone-step learner is* $\mathcal{F}_s$*-measurable. The stochastic gradient satisfies* $\mathbb{E}[g_s(\theta_s) \mid \mathcal{F}_s] = \nabla L_s(\theta_s)$ *almost surely, and* $\nabla L_s(\theta_s^*) = 0$*.*

Assumption 10 is used only by Theorem 1, whose learner class is defined through stochastic-gradient steps; it makes explicit the unbiasedness stated in prose in §3, and the step where it enters is spelled out in the proof of Lemma 13. All results are stated for $1 \leq r < d$, as in Assumption 6.

Table 13 summarizes which assumptions each theorem requires and the observation model assumed. All seven theorems share the same linear-Gaussian observation model $y_t = \theta_t^{*\top} x_t + \varepsilon_t$. Theorems 1, 4, 5 and 8 are exact-low-rank known fixed-U results; Theorem 2 is also known fixed-U and exact-low-rank, and the two unknown-U theorems assume isotropic design: Theorem 10 a fixed subspace with cumulative excitation, and Theorem 12 blockwise-constant subspaces.

Table 13: Assumption–theorem alignment. All results use the linear-Gaussian model with μ-SC/L-smooth population loss. "Known U" means the subspace is given to the algorithm; "Unk. U" means it must be estimated. In the Res column, 0 means exact low-rank. The two unknown-U theorems additionally require isotropic design, paths independent of the observations and an excitation condition, Assumption 4 or Assumption 5.

	1 SC	2 Sm	3 Noise	6 LR	7 Var	8 Bdd	9 Res
Thm 1, mono-step LB	✓	✓	✓	✓	✓	✓	0
Thm 2, fixed-w UB	✓	✓	✓	✓	✓	✓	0
Thm 4, adaptive UB	✓	✓	✓	✓	✓	✓	0
Thm 5, jump-resolved UB	✓	✓	✓	✓	✓	✓	0
Thm 8, minimax LB	✓	✓	✓	✓	—	✓	0
Thm 10, fixed unk. U	✓	✓	✓	✓	✓	✓	0
Thm 12, block unk. U	✓	✓	✓	✓	✓	✓	0

Reading the ladder with a residual. The theorems set $\rho_{\max} = 0$. For a residual that is orthogonal in the design metric, $U^\top \Sigma \rho_t = 0$, which for isotropic Σ is Assumption 9, and a residual path fixed in advance, two facts follow directly. First, for every $\theta = Ua$,

$$L_t(Ua) - L_t(\theta_t^*) = \tfrac{1}{2}(a - a_t)^\top \Sigma_z (a - a_t) + \tfrac{1}{2}\rho_t^\top \Sigma \rho_t,$$

because the cross term $(a-a_t)^\top U^\top \Sigma \rho_t$ vanishes; every learner whose iterates lie in $\mathrm{col}(U)$ therefore pays the additional $\frac{1}{2}\sum_t \rho_t^\top \Sigma \rho_t$, which is linear in T when the residual persists. Second, the projected responses are $y_t = a_t^\top z_t + (\rho_t^\top x_t + \varepsilon_t)$, and $\rho_t^\top x_t$ is Gaussian and uncorrelated with, hence independent of, z_t; the effective noise is independent of the design with variance at most $\bar{\sigma}^2 = \sigma^2 + L\rho_{\max}^2$. In the proof of Theorem 2, $\mathbb{E}[\|\tilde{a}_t - a_t\|^2 \mid G] \le \bar{\sigma}^2 \mathrm{tr}(G^{-1})$ still holds, so the bound of Theorem 2 holds with σ^2 replaced by $\bar{\sigma}^2$, plus the term of the first fact. The lower bounds, proved with $\rho_t = 0$, remain lower bounds on the larger class that allows residuals.

Notation. Table 14 lists the symbols used in the theorems and proofs, with the place where each is introduced.

C Discussion of Theorem 12 Conditions

Assumption 5 requires that coefficients excite all r subspace directions within each block. It must be checked through the smallest eigenvalue of the within-block coefficient second moment: a jump count alone does not ensure it, since repeated collinear coefficients remain rank deficient, and for piecewise-constant coefficients with $r > 1$ each block must contain at least r linearly independent coefficient values. It fails when coefficients are rank-deficient, for example when they are always aligned to a single direction. In the degenerate case $\mu_a \to 0$, the subspace is unidentifiable and the slow-track bound degrades gracefully via $1/\mu_a$. This assumption therefore scopes the unknown-subspace theorem to a persistently excited coefficient regime; it is not implied by the piecewise-constant coefficient-only drift construction used for the known-subspace rate ladder when $r > 1$. The blockwise-constant condition is a proof assumption, not a claim about arbitrary continuously drifting subspaces. Continuous within-block drift would add a perturbation term comparing $U_t a_t$ to the block representative $U_m a_t$, which the theorem does not include. The explicit dependence on d in the gradient-noise term reflects the cost of estimating a $d \times d$ projector from finite samples; for $B = \Theta(\sqrt{T})$ and bounded V_s, this term is $O(d\sqrt{T})$, which is lower-order than the $T^{3/4}$ leading terms when $d = o(T^{1/4})$. Every block also pays its first w prediction rounds, the explicit $A_{\max}^2 M w/2$ term, which is $O(T^{3/4})$ for $B \asymp \sqrt{T}$ and $w \asymp T^{1/4}$.

D Proof of Theorem 1: Necessity

Roadmap. The construction fixes an alternating path; Lemma 13 reduces the expected loss to the mean iterate; Lemma 14 charges each jump to a disjoint window; Steps 1–3 sum the charges.

Table 14: Notation.

Symbol	Meaning	Introduced
$d,\ r,\ T$	ambient dimension, rank, horizon	§3
$U,\ z_t = U^\top x_t$	known orthonormal basis; projected covariate	§3, Definition 3
$a_t,\ \theta_t^* = Ua_t,\ \rho_t$	coefficient, optimum, residual; $\rho_{\max} = 0$ in all theorems	Equation (1)
$\Sigma,\ \Sigma_z = U^\top \Sigma U$	design covariance and its projection	§3
$\mu,\ L,\ \kappa = L/\mu$	strong-convexity and smoothness constants, condition number	Assumptions 1 and 2
$\sigma,\ A_{\max}$	noise level; coefficient radius, also the clip radius	§3, Assumption 8
$K_T,\ \tau_k,\ S_k,\ J_k$	number of jumps, jump times, segments, segment lengths	Theorem 4
$J_{\min}$	minimum segment length, the inter-jump gap	Theorem 4
$\Delta_0,\ \Delta_k$	minimum jump size; size of the k-th jump	Assumption 7, Theorem 5
$\rho_D = A_{\max}^2/\Delta_0^2$	squared amplitude-to-jump ratio	Theorem 4
$\mathrm{Reg}_T,\ \mathcal{R}_T$	expected dynamic regret; realized pseudo-regret	§3, Proposition 19
$s_0,\ n$	current window start; number of samples in the window	Definition 3
$\mathrm{RSS}_{a:b},\ \Lambda(\tau, t)$	least-squares residual sum of squares; GLR statistic	Equation (3)
$h,\ h_{\min},\ u,\ c_h$	threshold, its minimum, tail level $\log(8T^5)$, threshold cap	Definition 3
Δ_c	cadence: rounds between two scans, $1 \le \Delta_c \le \lceil r \log T \rceil$	Definition 3
$n_0 = C_n(r + \log T)$	burn-in length, $C_n = 200$	Theorem 4, Lemma 17
$D,\ D_k,\ C_D$	delay bound for the smallest and the k-th jump; $C_D = 8(16c_h + 70)$	Lemma 17, Theorem 5
$\mathcal{E}_{\mathrm{cov}},\ \mathcal{E}_{\mathrm{cov}}^{\pm}$	every interval of length $\ge n_0$ has well-conditioned Gram matrix	Lemma 17, Section F.1
$\mathcal{E}_{\mathrm{GLR}},\ \mathcal{E}_{\mathrm{det}}$	uniform χ^2 tail event; no false alarm and delay at most $D + \Delta_c$	Lemma 17
$\tilde{a}_t$	ghost estimator: clipped OLS from a stopping-time start, defined whether or not a later test declares	Section F, Part 3
w	window width of fixed-window least squares	Theorem 2
$B,\ M,\ I_m$	block length, number of blocks, block index set	Theorem 12
$\mu_a,\ V_s,\ \mathcal{F}_r$	block excitation, projector path length, Fantope	Assumption 5, Equation (8)
$\hat{M}_m,\ \eta$	trace-centered spectral moment of block m; Fantope step size	Theorem 12
$t_e,\ O_t$	excitation time; odd rounds before t	Assumption 4
$T_j,\ t_s,\ \hat{U}_j$	epoch starts $2^j t_s + 1$; first epoch length; subspace estimate of epoch j	Definition 9
$\bar{\sigma},\ \epsilon,\ \bar{D}$	noise level of GLR-U with $\bar{\sigma}^2 \ge \sigma^2 + \epsilon^2 A_{\max}^2$; its delay bound	Theorem 10
$C_U,\ T_\star$	subspace-error constant; warm-up length	Theorem 10
$\mathrm{OLS}_{\mathrm{harm}}(g)$	harmonic estimation charge with guard g	Corollary 20

Throughout, $x_t \sim \mathcal{N}(0, \Sigma)$ with $\mu I \preceq \Sigma \preceq LI$ and $\Sigma u = \mu u$ for a unit vector $u \in \mathrm{col}(U)$, so that

$$L_t(\theta) - L_t(\theta_t^*) = \tfrac{1}{2}(\theta - \theta_t^*)^\top \Sigma (\theta - \theta_t^*).$$

The learner is $\theta_{t+1} = \theta_t - \eta_t g_t(\theta_t)$ with $\eta_t = ct^{-\alpha}$, $\alpha \in [\frac{1}{2}, 1]$, $0 < c \le \mu/(2L^2)$, arbitrary θ_1, and any gradient with $\mathbb{E}[g_t(\theta_t) \mid \mathcal{F}_t] = \nabla L_t(\theta_t) = \Sigma(\theta_t - \theta_t^*)$ as in Assumption 10; the sample gradient $g_t(\theta) = x_t x_t^\top (\theta - \theta_t^*) - \varepsilon_t x_t$ is one example.

Construction. Let $J = \lfloor T/(K_T + 1) \rfloor$ and $\tau_k = kJ + 1$ for $k = 1, \ldots, K_T$. The optimum is $\theta_t^* = (-1)^k \frac{\Delta_0}{2} u$ on the segment $S_k = \{\tau_k, \ldots, \tau_{k+1} - 1\}$, with $\tau_0 = 1$ and $\tau_{K_T+1} = T + 1$, so every jump has size Δ_0 and $\|\theta_t^*\| = \Delta_0/2 \le A_{\max}$. Every segment has length at least J.

Lemma 13 (Mean dynamics). *Let $m_t = \mathbb{E}\theta_t$ and $e_t = \langle m_t - \theta^\circ, u \rangle$ for a vector θ° equal to θ_s^* for all s in a stationary stretch $\{t_0, \ldots, t_1\}$. Then for $t_0 \le t \le t_1$,*

$$e_{t+1} = (1 - \mu\eta_t)\, e_t,$$
$$\mathbb{E}\big[L_t(\theta_t) - L_t(\theta_t^*)\big] \ge \frac{\mu}{2}\, e_t^2.$$

Proof. Taking expectations of the update and using conditional unbiasedness,

$$\begin{aligned} m_{t+1} &= m_t - \eta_t \, \mathbb{E}\big[\mathbb{E}[g_t(\theta_t) \mid \mathcal{F}_t]\big] \\ &= m_t - \eta_t \Sigma(m_t - \theta^\circ), \end{aligned}$$

because the population gradient is affine in θ. Taking the inner product with u and using $\Sigma u = \mu u$ gives the recursion. For the second claim, the excess loss is a convex quadratic, so Jensen's inequality and $\Sigma \succeq \mu I$ give

$$\begin{aligned} \mathbb{E}\big[\tfrac{1}{2}(\theta_t - \theta^\circ)^\top \Sigma(\theta_t - \theta^\circ)\big] &\geq \tfrac{1}{2}(m_t - \theta^\circ)^\top \Sigma(m_t - \theta^\circ) \\ &\geq \tfrac{\mu}{2} \left\| m_t - \theta^\circ \right\|^2 \\ &\geq \tfrac{\mu}{2} \, e_t^2. \end{aligned}$$

□

Since $\mu\eta_t \leq \mu c \leq \mu^2/(2L^2) \leq \frac{1}{2}$, every factor $1 - \mu\eta_t$ lies in $[\frac{1}{2}, 1)$. No property of the gradient noise beyond conditional unbiasedness is used, so the argument covers sample gradients of any variance.

Lemma 14 (Jump charging dichotomy)**.** *Fix a jump $\tau = \tau_k$ from θ^- to θ^+, let $N_\tau = \lfloor 1/(2\mu\eta_\tau) \rfloor$ and $W = \min\{N_\tau, \lfloor J/2 \rfloor\}$. Then at least one of the two windows $\{\tau - W, \ldots, \tau - 1\}$ and $\{\tau, \ldots, \tau + W - 1\}$ has expected regret at least $\mu\Delta_0^2 W/32$.*

Proof. Write $e_t^- = \langle m_t - \theta^-, u \rangle$ and $e_t^+ = \langle m_t - \theta^+, u \rangle = e_t^- \mp \Delta_0$.

Case 1: $|e_\tau^-| \leq \Delta_0/2$. Then $|e_\tau^+| \geq \Delta_0/2$. For $0 \leq n < W$ the rounds $\tau, \ldots, \tau + n - 1$ lie in the segment of θ^+, and $\eta_s \leq \eta_\tau$ for $s \geq \tau$, so Lemma 13 and Bernoulli's inequality give

$$\begin{aligned} |e_{\tau+n}^+| &\geq (1 - \mu\eta_\tau)^n \, |e_\tau^+| \\ &\geq (1 - n\mu\eta_\tau) \, \frac{\Delta_0}{2} \\ &\geq \frac{\Delta_0}{4}, \end{aligned}$$

using $n \leq N_\tau \leq 1/(2\mu\eta_\tau)$. Each round of the post-jump window therefore costs at least $\frac{\mu}{2}(\Delta_0/4)^2 = \mu\Delta_0^2/32$.

Case 2: $|e_\tau^-| > \Delta_0/2$. The rounds $\tau - W, \ldots, \tau - 1$ lie in the segment of θ^- because $W \leq J/2$. For s in this window, $e_\tau^- = \prod_{v=s}^{\tau-1}(1 - \mu\eta_v) \, e_s^-$ with every factor in $(0, 1]$, so $|e_s^-| \geq |e_\tau^-| > \Delta_0/2$, and each pre-jump round costs at least $\mu\Delta_0^2/8$. □

Completing the proof of Theorem 1. *Step 1: disjoint windows.* The pre-jump window of τ_k is contained in the last $\lfloor J/2 \rfloor$ rounds of S_{k-1} and the post-jump window in the first $\lfloor J/2 \rfloor$ rounds of S_k. Since $|S_{k-1}| \geq J$, the post-jump window of τ_{k-1} and the pre-jump window of τ_k do not overlap, so the windows charged by Lemma 14 are disjoint across jumps and their charges add.

Step 2: window length. Assume $T \geq 8$ and $\frac{1}{2}T^{1-\alpha} \leq K_T \leq T/8$. Then $K_T + 1 \leq T/4$, so $T/(K_T + 1) \geq 4$ and

$$\begin{aligned} J &\geq \frac{T}{K_T + 1} - 1 \geq \frac{3}{4} \, \frac{T}{K_T + 1}, \\ \left\lfloor \frac{J}{2} \right\rfloor &\geq \frac{J}{3} \geq \frac{T}{4(K_T + 1)}. \end{aligned}$$

For $k \geq K_T/2$,

$$\tau_k \geq \frac{K_T}{2} \cdot \frac{3}{4} \, \frac{T}{K_T + 1} \geq \frac{3T}{16}.$$

Because $2\mu c \le \mu^2/L^2 \le 1$, $N_{\tau_k} \ge \lfloor \tau_k^\alpha \rfloor \ge \tau_k^\alpha/2$, and with $(3/16)^\alpha \ge 3/16$ and $T^\alpha \ge T/(2K_T)$,

$$\begin{aligned} N_{\tau_k} &\ge \frac{3}{32}\, T^\alpha \\ &\ge \frac{3}{64}\, \frac{T}{K_T} \\ &\ge \frac{3}{64}\, \frac{T}{K_T+1}. \end{aligned}$$

Hence $W_k \ge \frac{3}{64}\, T/(K_T+1)$ for every $k \ge K_T/2$.

Step 3: summation. There are at least $K_T/2$ jumps with $k \ge K_T/2$. Summing the charges of Lemma 14 over them,

$$\begin{aligned} \mathbb{E}[\mathrm{Reg}_T] &\ge \frac{K_T}{2} \cdot \frac{\mu\Delta_0^2}{32} \cdot \frac{3}{64}\, \frac{T}{K_T+1} \\ &\ge \frac{3\,\mu\Delta_0^2}{8192}\, T, \end{aligned}$$

using $K_T/(K_T+1) \ge \frac{1}{2}$. For $K_T = \lfloor \sqrt{T} \rfloor$ the conditions hold for every $\alpha \in [\frac{1}{2}, 1]$ once $T \ge 64$, since $\lfloor \sqrt{T} \rfloor \ge \sqrt{T}/2 \ge T^{1-\alpha}/2$ and $\sqrt{T} \le T/8$. □

Remark 15 (Role of the window cap)**.** *Capping the charging windows at $\lfloor J/2 \rfloor$ makes the argument uniform in α: for $\alpha = 1$ the natural recovery length $N_\tau \asymp \tau$ exceeds the segment length, and the cap simply charges half a segment per jump. The bound depends on the jumps only through Δ_0^2 and holds for every $\Delta_0 \in (0, 2A_{\max}]$.*

E Proof of Theorem 2 and the Fixed-Window Lower Bound

Roadmap. The upper bound splits rounds into initial, contaminated and clean windows; Proposition 16 lower-bounds clean rounds in Step 1 and transient rounds in Step 2 for every width and chooses the path after the width in Step 3; Corollary 7 combines it with Theorem 4.

Proof of Theorem 2. With $\theta_t = U\hat{a}_t$ and $\theta_t^* = Ua_t$, the excess loss is

$$\begin{aligned} L_t(\theta_t) - L_t(\theta_t^*) &= \tfrac{1}{2}(\hat{a}_t - a_t)^\top \Sigma_z (\hat{a}_t - a_t) \\ &\le \tfrac{L}{2} \left\| \hat{a}_t - a_t \right\|^2. \end{aligned}$$

We split the rounds into three groups.

Initial window, $t \le w$: $\hat{a}_t = 0$, so each loss is at most $LA_{\max}^2/2$, in total $LA_{\max}^2 w/2$.

Contaminated windows, $t > w$ and $a_s \ne a_t$ for some s in the window $W_t = \{t-w, \dots, t-1\}$: then a change time τ, one with $a_\tau \ne a_{\tau-1}$, lies in $\{t-w+1, \dots, t\}$, and each change time is charged by at most the w rounds $t \in \{\tau, \dots, \tau+w-1\}$. Hence there are at most $\min\{T-w, K_T w\}$ such rounds, whether a window contains one change or several. Clipping and $\|a_t\| \le A_{\max}$ give $\|\hat{a}_t - a_t\| \le 2A_{\max}$, so each loss is at most $2LA_{\max}^2$.

Clean windows, all other $t > w$: $a_s = a_t$ for every $s \in W_t$, so $y_s = a_t^\top z_s + \varepsilon_s$ with $z_s = U^\top x_s \sim \mathcal{N}(0, \Sigma_z)$ i.i.d. and independent of the noise; the index set W_t is deterministic because the coefficient path is fixed in advance. With $G = \sum_{s \in W_t} z_s z_s^\top$, which is Wishart with $w > r+1$ degrees of freedom and invertible almost surely, the unclipped OLS estimate satisfies

$$\tilde{a}_t - a_t = G^{-1} \sum_{s \in W_t} z_s \varepsilon_s,$$

hence

$$\mathbb{E}[\|\tilde{a}_t - a_t\|^2 \mid G] = \sigma^2 \mathrm{tr}(G^{-1}).$$

The inverse-Wishart mean $\mathbb{E}[G^{-1}] = \Sigma_z^{-1}/(w-r-1)$ gives

$$\mathbb{E}\left\|\tilde{a}_t - a_t\right\|^2 = \frac{\sigma^2 \mathrm{tr}(\Sigma_z^{-1})}{w-r-1} \le \frac{\sigma^2 r}{\mu(w-r-1)}.$$

Euclidean projection onto the ball of radius $A_{\max}$, which contains a_t, does not increase the distance to a_t. Each of the at most $T - w$ clean rounds therefore costs at most $L\sigma^2 r/(2\mu(w - r - 1))$ in expectation.

Adding the three groups proves the bound. For the rate, a window $w \asymp \sqrt{T/(K_T + 1)}$ makes T/w, $K_T w$ and w all $O(\sqrt{T(K_T + 1)})$, which is $O(T^{3/4})$ when $K_T = \Theta(\sqrt{T})$. □

Finite risk and the window condition. For unclipped OLS the clean-window risk $\sigma^2 \mathrm{tr}(\Sigma_z^{-1})/(w - r - 1)$ is exact, so the condition $w > r + 1$ cannot be dropped: at $w = r + 1$ the expected risk of unclipped OLS is infinite. Clipping bounds every loss by $2LA_{\max}^2$, and a ridge term $\lambda > 0$ bounds the estimator's second moment; the ridge variant used in Figure 2 therefore has finite risk for every w.

Proposition 16 (Lower bound for every fixed window)**.** *Let $\Sigma = I_d$, so that $\mu = L = 1$, let U be known, $\Delta_0 \le A_{\max}$ and $\sigma^2 \le 12A_{\max}^2$, and let $1 \le K \le T/(2w_0) - 1$ with $w_0 = 32r + 128$. Consider the learner of Theorem 2 with any window $w \ge r + 2$, with or without clipping to the $A_{\max}$-ball. Then some coefficient path alternating between $\pm\frac{\Delta_0}{2}u$, $u \in \mathrm{col}(U)$, with at most K changes satisfies*

$$\mathbb{E}[\mathrm{Reg}_T] \;\ge\; c\,\min\{\Delta_0\sigma\sqrt{rTK},\; \sigma^2 T,\; \Delta_0^2 T\}, \qquad c = 9 \cdot 10^{-7}.$$

For $K = \lfloor\sqrt{T}\rfloor$ and fixed Δ_0, σ, r this is $\Omega(T^{3/4})$, matching Theorem 2.

Proof. Write Π for the Euclidean projection onto the $A_{\max}$-ball, which is the identity for the unclipped learner, and $a^{\pm} = \pm\frac{\Delta_0}{2}u$, so $\|a^{\pm}\| \le A_{\max}/2$. Two facts are used throughout. First, for any p and any a with $\|a\| \le A_{\max}/2$,

$$\|\Pi(p) - a\| \ge \min\{\|p - a\|, A_{\max}/2\},$$

since either $\Pi(p) = p$ or $\|\Pi(p)\| = A_{\max}$. Second, for a $j \times r$ matrix Z with i.i.d. $\mathcal{N}(0,1)$ entries, Gordon's inequality with Gaussian concentration of the 1-Lipschitz maps $Z \mapsto \sigma_{\max}(Z), \sigma_{\min}(Z)$ (Gordon, 1985; Vershynin, 2018) gives, for $s \ge 0$,

$$\Pr\big(\sigma_{\max}(Z) > \sqrt{j} + \sqrt{r} + s\big) \le e^{-s^2/2},$$
$$\Pr\big(\sigma_{\min}(Z) < \sqrt{j} - \sqrt{r} - s\big) \le e^{-s^2/2}.$$

Step 1: clean rounds. Let $t > w$ be a round whose window W_t lies in one segment with coefficient a, and let Z be the $w \times r$ design and $G = Z^\top Z$. Given Z, $\tilde{a}_t - a = G^{-1}Z^\top\varepsilon$ is $\mathcal{N}(0, \sigma^2 G^{-1})$, so $\|\tilde{a}_t - a\|^2 \ge \sigma^2\xi/\lambda_{\max}(G)$ with $\xi \sim \chi_r^2$ independent of Z. By the second fact with $s = 2$, $\lambda_{\max}(G) \le (\sqrt{w} + \sqrt{r} + 2)^2 \le 16w$ with probability at least $1 - e^{-2} \ge 0.86$, and the Paley–Zygmund inequality with $\mathbb{E}\xi = r$ and $\mathbb{E}\xi^2 = r^2 + 2r \le 3r^2$ gives $\Pr(\xi \ge r/3) \ge \frac{4}{9} \cdot \frac{1}{3} \ge 0.148$. By independence, $\|\tilde{a}_t - a\|^2 \ge \sigma^2 r/(48w)$ with probability at least 0.127. Since $\sigma^2 r/(48w) \le \sigma^2/48 \le A_{\max}^2/4$, the first fact gives

$$\mathbb{E}\big[L_t(\theta_t) - L_t(\theta_t^*)\big] \ge \frac{1}{2} \cdot 0.127 \cdot \frac{\sigma^2 r}{48w} \ge \frac{\sigma^2 r}{756\, w}.$$

Step 2: transient rounds. Suppose $w \ge w_0$, let a change from a^- to a^+ occur at τ, and let $t \in \{\tau + 1, \dots, \tau + \lfloor w/2 \rfloor\}$ with W_t inside the two adjacent segments. The window has $j \ge w/2$ pre-change rows Z_- with Gram matrix G_-, and

$$\begin{aligned}\tilde{a}_t - a^+ &= G^{-1}G_-(a^- - a^+) + G^{-1}Z^\top\varepsilon \\ &=: b + \zeta.\end{aligned}$$

By the second fact, $\lambda_{\max}(G) \le 16w$ with probability at least $1 - e^{-2}$, and, with $s = \sqrt{j}/4$ and $j \ge 16r$, $\lambda_{\min}(G_-) \ge j/4$ with probability at least $1 - e^{-j/32} \ge 1 - e^{-2}$ because $j \ge w_0/2 \ge 64$. On both events,

$$\|b\| \ge \frac{\lambda_{\min}(G_-)\,\Delta_0}{\lambda_{\max}(G)} \ge \frac{(j/4)\Delta_0}{16w} \ge \frac{\Delta_0}{128}.$$

Given Z, ζ is symmetric, and $\|b + \zeta\|^2 + \|b - \zeta\|^2 = 2\|b\|^2 + 2\|\zeta\|^2$, so $\|b + \zeta\| \ge \|b\|$ with conditional probability at least $\frac{1}{2}$. With the first fact and $\Delta_0/128 \le A_{\max}/2$, $\|\Pi(\tilde{a}_t) - a^+\| \ge \Delta_0/128$

with probability at least $\frac{1}{2}(1-2e^{-2}) \geq 0.36$, so each such round costs at least $\frac{1}{2} \cdot 0.36\,(\Delta_0/128)^2 \geq \Delta_0^2/91{,}200$.

Step 3: choice of path. Case A: $w_0 \leq w \leq T/(K+1)$. Use K equally spaced changes; segments have length $J = \lfloor T/(K+1) \rfloor \geq w$. At most $(K+1)w$ rounds are initial or have a window meeting a change, so at least $T-(K+1)w$ rounds are clean, and each change has $\lfloor w/2 \rfloor \geq w/3$ transient rounds of Step 2. If $(K+1)w \leq T/2$,

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_T] &\geq \frac{\sigma^2 rT}{1{,}512\,w} + \frac{Kw\Delta_0^2}{273{,}600} \\ &\geq 2\sqrt{\frac{\sigma^2 rTK\Delta_0^2}{1{,}512 \cdot 273{,}600}} \\ &\geq \frac{\Delta_0 \sigma\sqrt{rTK}}{10{,}170}.\end{aligned}$$

If $(K+1)w > T/2$, then $Kw \geq T/4$ and the transient rounds alone give at least $T\Delta_0^2/1{,}094{,}400$. *Case B:* $w > T/(K+1)$. If $w > T/4$, the w initial zero predictions cost $w\,\Delta_0^2/8 \geq T\Delta_0^2/32$. Otherwise use segments of length w, i.e. $K' = \lfloor T/w \rfloor - 1 \leq K$ changes and $K' \geq T/w - 2 \geq T/(2w)$ because $T/w \geq 4$; the transient rounds give at least $K'(w/3)\Delta_0^2/91{,}200 \geq T\Delta_0^2/547{,}200$. *Case C:* $r+2 \leq w < w_0$. Use K equally spaced changes; since $(K+1)w_0 \leq T/2$, at least $T/2$ rounds are clean, and Step 1 gives $\mathbb{E}[\mathrm{Reg}_T] \geq \sigma^2 rT/(1{,}512\,w_0) \geq \sigma^2 T/241{,}920$ because $r/w_0 \geq 1/160$.

Every case is at least $c\min\{\Delta_0\sigma\sqrt{rTK}, \sigma^2 T, \Delta_0^2 T\}$ with $c = 9 \cdot 10^{-7} \leq 1/1{,}094{,}400$. □

The proposition applies to the learner of Theorem 2 itself, so together they give $\Theta(\sqrt{TK_T})$ for fixed-window least squares with any width $w \geq r+2$; no premise about a "steady-state variance" or a "recovery cost" is assumed.

Proof of Corollary 7. *Step 1: the paths of Proposition 16 lie in* $\mathcal{C}_T$. Let $K = K_T = \lfloor T^\gamma \rfloor$ and $J = \lfloor T/(K+1) \rfloor$. In Cases A and C of the proof of Proposition 16 the K changes are equally spaced, at the rounds $1 + kJ$, so consecutive changes are J apart and the last segment has length $T - KJ \geq J$. In Case B with $w \leq T/4$ the segments have length $w > T/(K+1) \geq J$, and with $w > T/4$ the bound uses only the initial zero predictions, so the equally spaced path serves. Every jump has size exactly Δ_0 and every coefficient has norm $\Delta_0/2 \leq A_{\max}$. Hence, for every $w \geq r+2$, some path in $\mathcal{C}_T$ has

$$\mathbb{E}[\mathrm{Reg}_T] \geq c\min\{\Delta_0\sigma\sqrt{rTK},\ \sigma^2 T,\ \Delta_0^2 T\},$$

provided $K \leq T/(2w_0) - 1$, which holds for all large T because $\gamma < 1$. For fixed (r, σ, Δ_0) and $K \geq T^\gamma/2$ the first entry is of order $T^{(1+\gamma)/2}$ and the other two of order T, so the minimum is $\Omega(T^{(1+\gamma)/2})$. The width may depend on T, K_T and γ; the path is chosen after the width.

Step 2: the adaptive learner on $\mathcal{C}_T$. With $\Sigma = I_d$, $\mu = L = \kappa = 1$. Every path of $\mathcal{C}_T$ has at most K jumps of size at least Δ_0 and minimum gap at least $J \geq T^{1-\gamma}/2 - 1$, whereas $n_0 + 2(D + \Delta_c) = O(\log T)$ for fixed (r, σ, Δ_0) and $\Delta_c = 1$. Hence the spacing condition of Theorem 4 holds for all large T, uniformly over $\mathcal{C}_T$, and Equation (4) gives

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_T] &\leq 4A_{\max}^2 K(D+1) + 2(K+1)\,\sigma^2 r \log\frac{eT}{(K+1)n_0} \\ &\quad + 2(K+1)A_{\max}^2 n_0 + O(A_{\max}^2) \\ &= O(T^\gamma \log T)\end{aligned}$$

for every path of $\mathcal{C}_T$. The learner uses neither K_T nor γ.

Step 3: ratio. Dividing the two orders, $T^{(1+\gamma)/2}/(T^\gamma \log T) = T^{(1-\gamma)/2}/\log T \to \infty$. □

F Proof of Theorem 4: Adaptive Window

Roadmap. Lemma 17 builds the good event $\mathcal{E}_{\mathrm{det}} = \mathcal{E}_{\mathrm{cov}} \cap \mathcal{E}_{\mathrm{GLR}}$, on which there is no false alarm and every delay is at most $D + \Delta_c$, from Lemma 18 and χ^2 tails; the proof of Theorem 4 then charges,

per segment, the delay in Part 1, the burn-in in Part 2, the ghost least-squares risk in Part 3 and the complement of $\mathcal{E}_{\mathrm{det}}$ in Part 4.

Theorem (Restatement of Theorem 4)**.** *Under the exact low-rank linear-Gaussian model with known, fixed subspace U and $\rho_{\max} = 0$, K_T jumps of size $\geq \Delta_0$, and minimum inter-jump gap $J_{\min} \geq n_0 + 2(D + \Delta_c)$, the adaptive-window clipped-OLS learner achieves:*

$$\mathbb{E}[\mathrm{Reg}_T] = O\Big((\kappa\sigma^2\rho_D + LA_{\max}^2)\, r\sqrt{T}\log T\Big) \quad \textit{for } K_T = \Theta(\sqrt{T}).$$

The proof relies on two technical lemmas.

Lemma 17 (Uniform GLR detection event)**.** *Under the assumptions of Theorem 4, namely known U, independent designs with fixed covariance, independent Gaussian noise, a deterministic coefficient path fixed in advance, jumps of size at least Δ_0, $\rho_{\max} = 0$ and the stated minimum spacing, and for any threshold $h \in [h_{\min},\, c_h r\log(2T^3)]$ with $c_h \geq 11/2$, where $h_{\min} = \frac{1}{2}(r + 2\sqrt{ru} + 2u)$ and $u = \log(8T^5)$, the GLR rule satisfies, with probability at least $1 - 1/T$ over the full horizon:*

1. *no false alarms are declared;*
2. *every true jump is detected after delay at most $D + \Delta_c$, where $D = C_D(n_0 + r\sigma^2\log(T)/(\mu\Delta_0^2))$ with a universal constant $C_D \geq 8(16c_h + 70)$.*

Proof. **Stopping-time conditioning and uniform GLR event.** The estimation-window start s_0 is initialized to 1 and, after a detection declared at time t, is reset to $t + 1$. Hence every realized s_0 is a stopping time with respect to the natural filtration $\mathcal{F}_t = \sigma(z_1, y_1, \ldots, z_t, y_t)$. The retrospective maximizer $\hat{\tau} = \arg\max_\tau \Lambda(\tau, t)$ is *not* used as the next window start. Therefore, conditionally on $\mathcal{F}_{s_0-1}$, all observations used by subsequent OLS fits and GLR tests are fresh samples from the current segment until the next jump.

Fixed-triple GLR law. Fix a deterministic admissible triple (s, τ, t) and condition on its design. Let Z_L, Z_R be the pre/post design matrices, with Gram matrices $G_L = Z_L^\top Z_L$ and $G_R = Z_R^\top Z_R$. Stack the responses into y. Let P_0 be the orthogonal projector onto the pooled regression space $\{(Z_L a, Z_R a) : a \in \mathbb{R}^r\}$ and P_1 the projector onto the split space $\{(Z_L a_L, Z_R a_R) : a_L, a_R \in \mathbb{R}^r\}$. Since the pooled space is contained in the split space,

$$2\Lambda(s, \tau, t) = \sigma^{-2} y^\top (P_1 - P_0) y.$$

The matrix $P_1 - P_0$ is an orthogonal projector of rank at most r. Hence, conditionally on the design, 2Λ is central χ_q^2 under no change, with $q \leq r$, and is noncentral $\chi_q^2(\lambda)$ under a true split, where

$$\begin{aligned}\lambda &= \sigma^{-2}\left\|(P_1 - P_0)m\right\|^2 \\ &= \sigma^{-2}\delta^\top (G_L^{-1} + G_R^{-1})^{-1}\delta,\end{aligned}$$

with $\delta = a_R - a_L$. The last identity is the standard two-sample least-squares identity obtained by solving the pooled and split normal equations.

Let $\mathcal{E}_{\mathrm{cov}}$ be the event that every interval of length at least $n_0 = C_n(r + \log T)$ has empirical covariance lower bounded by $\mu/4$. We take $C_n = 200$. Lemma 18 with $\delta = 1/(4T^4)$ needs a length of at least

$$\begin{aligned}16r + 32\log(4T^4) &\leq 16r + 45 + 128\log T \\ &\leq 200(r + \log T) = n_0,\end{aligned}$$

so a union bound over at most T^2 intervals gives $\Pr(\mathcal{E}_{\mathrm{cov}}^c) \leq T^2\delta = 1/(4T^2)$. On $\mathcal{E}_{\mathrm{cov}}$, if the left and right sides have lengths $m \geq n \geq n_0$, then $G_L \succeq m\mu I/4$ and $G_R \succeq n\mu I/4$, so

$$\begin{aligned}(G_L^{-1} + G_R^{-1})^{-1} &\succeq \frac{\mu mn}{4(m+n)} I_r \\ &\succeq \frac{\mu n}{8} I_r.\end{aligned}$$

Thus $\lambda \geq \mu n \left\|\delta\right\|^2/(8\sigma^2)$.

Finally, fix the single tail level $u = \log(8T^5)$ and let $\mathcal{E}_{\mathrm{GLR}}$ be the event on which, for every deterministic admissible triple, the central upper-tail bound for triples without a change and the

noncentral lower-tail bound for triples split at a true change used below hold at level u. Each bound fails with conditional probability at most e^{-u} given the design, and there are at most T^3 triples and two bounds per triple, so

$$\Pr(\mathcal{E}_{\mathrm{GLR}}^c) \le 2T^3 e^{-u} = \frac{1}{4T^2}.$$

Every triple checked by the adaptive algorithm is one of these deterministic triples. We use the following bounds, all valid for $T \ge 4$:

$$\begin{aligned} u &\le 3\log(2T^3), \\ u &\le 7\log T, \\ \log(2T^3) &\le 4\log T, \\ \log T &\ge 1. \end{aligned}$$

Detection delay on the no-false-alarm history. Because Definition 3 scans all admissible split points, the true change-point τ^* is included in the GLR maximization once both sides contain at least $r+1$ samples. On the no-false-alarm event, the previous declaration reset occurs only after the previous true jump. The strengthened minimum-gap assumption then ensures that the pre-change side has at least $\max\{n_0, n\}$ clean samples for all $n \le D + \Delta_c$ considered below. For $n \ge n_0$ post-change observations, the fixed-triple calculation gives $2\Lambda(\tau^*, t) \sim \chi^2_q(\lambda)$ with $q \le r$ and $\lambda \ge \mu n \Delta_0^2/(8\sigma^2)$ on $\mathcal{E}_{\mathrm{cov}}$. On $\mathcal{E}_{\mathrm{cov}}$ both Gram matrices are invertible, so $q = r$, and the standard noncentral χ^2 lower-tail bound

$$\Pr\Big(\chi^2_r(\lambda) \le (r+\lambda) - 2\sqrt{(r+2\lambda)u}\Big) \le e^{-u}$$

holds at the level u of $\mathcal{E}_{\mathrm{GLR}}$. Since

$$\begin{aligned} 2\sqrt{(r+2\lambda)u} &\le 2\sqrt{ru} + 2\sqrt{2\lambda u} \\ &\le r + 5u + \lambda/2, \end{aligned}$$

on $\mathcal{E}_{\mathrm{GLR}}$ the statistic satisfies $2\Lambda \ge \lambda/2 - 5u$, which exceeds $2h$ once $\lambda \ge 4h + 10u$. Because $h \le c_h r \log(2T^3)$ gives

$$4h + 10u \le (16c_h + 70)\, r \log T,$$

it suffices that

$$n \ge \frac{8(16c_h + 70)\, r\sigma^2 \log T}{\mu \Delta_0^2}$$

and

$$n \ge n_0,$$

which holds for $n \ge C_D(n_0 + r\sigma^2 \log(T)/(\mu\Delta_0^2))$ with $C_D \ge 8(16c_h + 70)$. A single logarithm suffices: the tail level $u = O(\log T)$ already pays for the union over all triples, so no further logarithmic factor enters the delay.

No false alarms on the good event. Under the null hypothesis H_0 of no change-point, the fixed-triple calculation gives $2\Lambda(\tau, t) \sim \chi^2_q$ with $q \le r$, hence its upper tail is bounded by that of χ^2_r. The Laurent–Massart bound (Laurent & Massart, 2000) is

$$\Pr\big(\chi^2_r \ge r + 2\sqrt{ru} + 2u\big) \le e^{-u},$$

and $r + 2\sqrt{ru} + 2u = 2h_{\min} \le 2h$. The interval of admissible thresholds is nonempty because

$$\begin{aligned} 2h_{\min} &\le 2r + 3u \\ &\le 11 r \log(2T^3), \end{aligned}$$

i.e. $h_{\min} \le c_h r \log(2T^3)$ for $c_h \ge 11/2$. Hence on $\mathcal{E}_{\mathrm{GLR}}$ no triple without a change crosses the threshold, i.e. there are no false alarms.

Failure probability. The same tail level u is used for both tails, so the failure probability is accounted for once:

$$\Pr(\mathcal{E}_{\mathrm{cov}}^c) + \Pr(\mathcal{E}_{\mathrm{GLR}}^c) \le \frac{1}{4T^2} + \frac{1}{4T^2} \le \frac{1}{T},$$

which gives the stated probability $1 - 1/T$. □

Lemma 18 (OLS concentration)**.** *For $w \geq 16r + 32\log(1/\delta)$, the sample covariance satisfies $\lambda_{\min}(\hat{\Sigma}_z) \geq \mu/4$ with probability $\geq 1-\delta$.*

Proof. By Gordon's inequality (Gordon, 1985); see also Vershynin (2018), the minimum singular value of the $w \times r$ Gaussian data matrix satisfies $\Pr(\sigma_{\min}(G/\sqrt{w}) \leq 1/2) \leq e^{-w/32}$ when $w \geq 16r$. The result follows from the eigenvalue factorization $\lambda_{\min}(\hat{\Sigma}_z) \geq \mu \cdot \lambda_{\min}(\hat{\Sigma}_g)$. □

Proof of Theorem 4. Partition $[1,T]$ into K_T+1 segments $S_0,\dots,S_{K_T}$ delimited by jump times $\tau_1 < \cdots < \tau_{K_T}$, with $|S_k| = J_k \geq J_{\min}$. In each segment k:

Part 1: detection transient. Let $\mathcal{E}_{\text{det}}$ be the event from Lemma 17. On $\mathcal{E}_{\text{det}}$, there are no false alarms and every true jump is detected within $D+\Delta_c$, where $D = C_D(n_0 + r\sigma^2\log T/(\mu\Delta_0^2))$. During these $D+\Delta_c$ steps, clipping and $\|a_t\| \leq A_{\max}$ bound the per-step regret by

$$\begin{aligned}\tfrac{1}{2}\|\hat{a}_t - a_t\|^2_{\Sigma_z} &\leq \tfrac{L}{2}(2A_{\max})^2 \\ &= 2LA^2_{\max} \\ &\leq 4LA^2_{\max}.\end{aligned}$$

Cost per jump: $4LA^2_{\max}(D+\Delta_c)$.

Part 2: burn-in. After reset, the algorithm requires $n_0 = C_n(r+\log T)$ steps before Lemma 18 guarantees $\lambda_{\min}(\hat{\Sigma}_z) \geq \mu/4$. Cost: $2LA^2_{\max}n_0$.

Part 3: OLS convergence, without conditioning on the future. Let $s_0 = 1$ and, for $k \geq 1$, let $s_k - 1$ be the first declaration time in $\{\tau_k,\dots,\tau_k + D + \Delta_c - 1\}$, or $\tau_k + D + \Delta_c - 1$ if there is none. Each s_k is a stopping time, so the samples with indices $\geq s_k$ are independent of $\mathcal{F}_{s_k-1}$. For $s_k \leq t < \tau_{k+1}$ define the *ghost* estimate $\tilde{a}_t$ as clipped OLS on $\{s_k,\dots,t-1\}$; it is defined whether or not the algorithm later declares. On $\mathcal{E}_{\text{det}}$ there is no false alarm, so the algorithm's estimate at such t equals $\tilde{a}_t$, and therefore

$$\mathbb{E}\big[(L_t(\theta_t) - L_t(\theta_t^*))\mathbf{1}_{\mathcal{E}_{\text{det}}}\big] \leq \mathbb{E}\big[L_t(U\tilde{a}_t) - L_t(Ua_t)\big].$$

Given $\mathcal{F}_{s_k-1}$, $\tilde{a}_t$ is clipped OLS on $n = t - s_k$ fresh homogeneous samples, so for $n \geq n_0 \geq 2r+2$ the inverse-Wishart mean $\mathbb{E}[G_n^{-1}] = \Sigma_z^{-1}/(n-r-1)$ and the nonexpansiveness of clipping give

$$\begin{aligned}\mathbb{E}\big[L_t(U\tilde{a}_t) - L_t(Ua_t) \;\big|\; \mathcal{F}_{s_k-1}\big] &\leq \frac{L\sigma^2\mathrm{tr}(\Sigma_z^{-1})}{2(n-r-1)} \\ &\leq \frac{2\kappa\sigma^2 r}{n}.\end{aligned}$$

Summing over the segment,

$$\sum_{n=n_0}^{J_k} \frac{2\kappa\sigma^2 r}{n} \leq 2\kappa\sigma^2 r \log\frac{eJ_k}{n_0},$$

and, by concavity of $\log$, the sum over the K_T+1 segments is at most $(K_T+1)\,2\kappa\sigma^2 r\log\frac{eT}{(K_T+1)n_0}$. Selecting a window by the non-detection of later tests would bias the noise; the ghost estimate avoids this because its window start is a stopping time and its end is deterministic.

Part 4: rare detection failure. On $\mathcal{E}_{\text{det}}^c$, clipping gives worst-case regret $O(LA^2_{\max})$ per step, so the contribution to expected regret is

$$O(LA^2_{\max}T)\Pr(\mathcal{E}_{\text{det}}^c) = O(LA^2_{\max}).$$

Summation. Adding the four terms over all segments:

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_T] \leq{}& K_T \cdot 4LA^2_{\max}(D+\Delta_c) + (K_T+1)\cdot 2LA^2_{\max}n_0 \\ &+ (K_T+1)\cdot 2\kappa\sigma^2 r\log\tfrac{eT}{(K_T+1)n_0} + O(LA^2_{\max}).\end{aligned}$$

Substituting $D = O(n_0 + r\sigma^2\log T/(\mu\Delta_0^2))$, $n_0 = O(r\log T)$, and $K_T = \Theta(\sqrt{T})$, the signal-dependent detection component contributes

$$\frac{4C_D K_T L A^2_{\max} r\sigma^2\log T}{\mu\Delta_0^2} = O\big(\kappa\sigma^2\rho_D \cdot r\sqrt{T}\log T\big),$$

where $\rho_D = A_{\max}^2/\Delta_0^2$. The detection floor, set by n_0 and Δ_c, and burn-in contribute $O(LA_{\max}^2 r\sqrt{T}\log T)$, and OLS convergence contributes $O(\kappa\sigma^2 r\sqrt{T}\log T)$, which is dominated by the first term because $\rho_D \geq 1/4$. Combining,

$$\mathbb{E}[\mathrm{Reg}_T] = O\big((\kappa\sigma^2\rho_D + LA_{\max}^2)\, r\sqrt{T}\log T\big).$$

□

F.1 Proof of Theorem 5: Jump-Size-Resolved Transient

Roadmap. Step 1 strengthens the covariance event to two sides; Step 2 gives each jump its own delay D_k; Step 3 bounds the mixed-window bias by $9\kappa\Delta_k$; Step 4 sums the noise as in Part 3 of Section F; Steps 5–6 add the terms and remove the jump sizes.

The proof reuses the events of Lemma 17 and changes only how the rounds between a change and its declaration are charged.

Step 1: two-sided covariance event. Let $\mathcal{E}_{\mathrm{cov}}^{\pm}$ be the event that every interval I of length $|I| \geq n_0$ satisfies

$$\begin{aligned}\frac{\mu}{4}|I|\, I_r &\preceq G_I \\ &\preceq \frac{9L}{4}|I|\, I_r,\end{aligned}$$

where $G_I = \sum_{s\in I} z_s z_s^\top$. Write $G_I = \Sigma_z^{1/2} W^\top W \Sigma_z^{1/2}$ with W an $|I| \times r$ standard Gaussian matrix. For $|I| \geq 16r$, $\sqrt{r} \leq \sqrt{|I|}/4$, and Gordon's inequality with deviation $\sqrt{|I|}/4$ gives

$$\begin{aligned}\frac{\sqrt{|I|}}{2} &\leq \sigma_{\min}(W) \\ &\leq \sigma_{\max}(W) \\ &\leq \frac{3\sqrt{|I|}}{2}\end{aligned}$$

outside probability $2e^{-|I|/32}$; the displayed bounds follow from $\mu I \preceq \Sigma_z \preceq LI$. A union bound over at most T^2 intervals gives $\Pr((\mathcal{E}_{\mathrm{cov}}^{\pm})^c) \leq 1/(4T^2)$ with the same $C_n = 200$, because

$$\begin{aligned}32\log(8T^4) &\leq 67 + 128\log T \\ &\leq 200(r + \log T) = n_0\end{aligned}$$

gives $2e^{-n_0/32} \leq 1/(4T^4)$ for every interval of length at least n_0. We use $\mathcal{E}_{\mathrm{cov}}^{\pm}$ in place of $\mathcal{E}_{\mathrm{cov}}$ in Lemma 17; this changes no statement of Theorem 4. Let $\mathcal{E}_{\mathrm{det}}$ denote the resulting detection event, with $\Pr(\mathcal{E}_{\mathrm{det}}^c) \leq 1/T$.

Step 2: jump-specific delay. The event $\mathcal{E}_{\mathrm{GLR}}$ of Lemma 17 controls the noncentral lower tail for *every* triple split at a true change, whatever the size of that change. The delay argument of Lemma 17 applied to the k-th change uses only $\lambda \geq \mu n\Delta_k^2/(8\sigma^2)$, so on $\mathcal{E}_{\mathrm{det}}$ the k-th change is declared at some $t_k \leq \tau_k + D_k + \Delta_c$. The pre-change side contains at least $\max\{n_0, n\}$ samples for every $n \leq D_k + \Delta_c \leq D + \Delta_c$ by the spacing assumption.

Step 3: bias during the delay. Fix k and a prediction round t with $\tau_k \leq t \leq t_k$, and work on $\mathcal{E}_{\mathrm{det}}$. The window $\{s_0, \ldots, t-1\}$ starts after the previous declaration, hence after τ_{k-1}, so it consists of $m \geq n_0$ samples with coefficient $a^- = a_{\tau_k - 1}$ followed by $n = t - \tau_k \geq 0$ samples with coefficient $a^+ = a_{\tau_k}$. Let G_L, G_R be the Gram matrices of the two parts, $G = G_L + G_R$, and let $\xi = G^{-1}\sum_{s=s_0}^{t-1} z_s\varepsilon_s$. The unclipped estimate is

$$\begin{aligned}\tilde{a}_t &= G^{-1}(G_L a^- + G_R a^+) + \xi \\ &= a^+ + G^{-1}G_L(a^- - a^+) + \xi.\end{aligned}$$

On $\mathcal{E}_{\mathrm{cov}}^{\pm}$, $\lambda_{\min}(G) \geq \mu(m+n)/4$ and $\lambda_{\max}(G_L) \leq 9Lm/4$, so

$$\begin{aligned}\big\|G^{-1}G_L\big\| &\leq \frac{9\kappa m}{m+n} \\ &\leq 9\kappa,\end{aligned}$$

and therefore

$$\left\|\tilde{a}_t - \xi - a^+\right\| \le 9\kappa\Delta_k.$$

Projection onto the $A_{\max}$-ball does not increase the distance to a^+, which lies in the ball, so with $\hat{a}_t$ the clipped estimate,

$$\left\|\hat{a}_t - a^+\right\| \le \min\{2A_{\max},\, 9\kappa\Delta_k + \|\xi\|\}.$$

Using the elementary bounds

$$\begin{aligned}\tfrac{1}{2}\|v\|_{\Sigma_z}^2 &\le \tfrac{L}{2}\|v\|^2,\\ (x+y)^2 &\le 2x^2 + 2y^2,\\ \min\{a, b+c\} &\le \min\{a,b\} + c \quad (c \ge 0),\end{aligned}$$

we obtain

$$L_t(U\hat{a}_t) - L_t(Ua^+) \le \min\{2LA_{\max}^2,\, 81L\kappa^2\Delta_k^2\} + L\|\xi\|^2.$$

Step 4: noise. The vector ξ is the noise part of least squares on the window and does not depend on the coefficient path. As in Part 3 of the proof of Theorem 4, the window start is a stopping time and the window consists of fresh samples; for a window of $N \ge n_0 \ge 2r+2$ fresh samples,

$$\begin{aligned}\mathbb{E}[L\|\xi\|^2] &= L\sigma^2\mathbb{E}\mathrm{tr}(G^{-1})\\ &= \frac{L\sigma^2\mathrm{tr}(\Sigma_z^{-1})}{N-r-1}\\ &\le \frac{2\kappa\sigma^2 r}{N}.\end{aligned}$$

On clean rounds the excess loss is $\frac{1}{2}\|\hat{a}_t - a_t\|_{\Sigma_z}^2 \le \frac{L}{2}\|\xi\|^2$, which obeys the same bound. Dropping the nonnegative indicator of the event that the run has not yet been declared, the noise contribution of the run started at a reset is at most

$$\sum_{N=n_0}^{N_{\max}} \frac{2\kappa\sigma^2 r}{N} \le 2\kappa\sigma^2 r \log\frac{eN_{\max}}{n_0},$$

where the deterministic cap

$$\begin{aligned}N_{\max} &\le J + D + \Delta_c\\ &\le 2J\end{aligned}$$

uses the length J of the segment in which the run starts and the spacing assumption. Summing over the K_T+1 runs and applying Jensen's inequality to $\sum_j \log J_j$ gives

$$(K_T+1)\,2\kappa\sigma^2 r\log\frac{2eT}{(K_T+1)n_0}.$$

Step 5: summation. On $\mathcal{E}_{\mathrm{det}}$ each change contributes at most $D_k + \Delta_c$ delay rounds, each charged $\min\{2LA_{\max}^2, 81L\kappa^2\Delta_k^2\}$ beyond its noise term; each of the K_T+1 runs pays at most n_0 burn-in rounds at $2LA_{\max}^2$; the noise terms are bounded in Step 4. On $\mathcal{E}_{\mathrm{det}}^c$ clipping bounds every round by $2LA_{\max}^2$, which contributes at most $2LA_{\max}^2 T\Pr(\mathcal{E}_{\mathrm{det}}^c) \le 2LA_{\max}^2$. This proves the first display.

Step 6: jump-size-free form. For the second display, apply $\min\{a,b\}(x+y) \le ax + by$ with

$$\begin{aligned}x &= C_D n_0 + \Delta_c,\\ y &= \frac{C_D r\sigma^2\log T}{\mu\Delta_k^2}.\end{aligned}$$

Because

$$L\kappa^2\Delta_k^2 \cdot \frac{r\sigma^2\log T}{\mu\Delta_k^2} = \kappa^3\sigma^2 r\log T,$$

each change contributes at most

$$2LA_{\max}^2(C_D n_0 + \Delta_c) + 81C_D\kappa^3\sigma^2 r\log T.$$

The remaining terms are at most $(K_T+1)(2LA_{\max}^2 n_0 + 6\kappa\sigma^2 r\log T)$ for $T \ge e^2$, and $\kappa \ge 1$. □

Comparison with adaptive-regret guarantees. Under the same protocol, in which the coefficient is chosen before z_t is revealed, the guarantee available for follow-the-leading-history is obtained by summing interval bounds over the $K_T + 1$ stationary segments. For squared loss, the exp-concavity modulus and the base-learner bound scale with the squared response range Y^2, and truncating Gaussian responses at their high-probability level gives $Y^2 \asymp (LA_{\max}^2 + \sigma^2)\log T$. The summed guarantee is therefore of order $(K_T+1)\, r\, (LA_{\max}^2 + \sigma^2)\log^2 T$ with one expert per start time, and one $\log T$ more with pruning. On the well-spaced class, Theorem 5 gives $(K_T+1)\, r\, (\kappa^3\sigma^2 + LA_{\max}^2)\log T$, one logarithm fewer at fixed κ. This compares guarantees, not the algorithms' actual regret.

F.2 Conditional Guarantee for Suffix Retention

The exact theorem above resets at the declaration time. A suffix-retention variant instead keeps the samples after the retrospective split selected by the same GLR statistic. The next result formalizes what this can buy once localization is assumed correct; it is conditional on a localization event and complements Lemma 17.

Proposition 19 (Conditional suffix-retention bound)**.** *Consider clipped expanding-window OLS in the known-U exact-low-rank Gaussian model with a deterministic piecewise-constant coefficient sequence. For $\delta \in (0,1)$ put*

$$u_\delta = \log(T^2/\delta),$$
$$b_\delta = r + 2\sqrt{r u_\delta} + 2u_\delta,$$

and let $\mathcal{E}_\delta$ be the event that, simultaneously for every homogeneous interval I with $n = |I|$ and $G_I = \sum_{s\in I} z_s z_s^\top \succeq n\mu I_r/4$, the clipped OLS estimate on I satisfies

$$L_t(U\hat{a}_I) - L_t(Ua_I) \le \frac{2\kappa\sigma^2 b_\delta}{n}$$

at every prediction round t with the same coefficient as I; then $\Pr(\mathcal{E}_\delta) \ge 1-\delta$, even if the interval is subsequently selected using all of its responses. Let $\mathcal{E}_{\mathrm{suf}}$ be an event on which all retained suffixes are homogeneous, each jump contributes at most D_k contaminated prediction rounds, and all clean intervals of length at least $n_0 = C_n(r+\log T)$ have empirical covariance at least $\mu I_r/4$. Write R for the number of clean restarts, B_j for their retained lengths, and M_j for the subsequent clean prediction-run lengths; set $B_0 = 0$ for the initial run. Let $H(0,q) = 0$ and $H(M,q) = 1 + \log(1+M/q)$ for $M > 0$. On $\mathcal{E}_{\mathrm{suf}} \cap \mathcal{E}_\delta$, the realized pseudo-regret $\mathcal{R}_T = \sum_t [L_t(\theta_t) - L_t(\theta_t^*)]$ obeys

$$\mathcal{R}_T \le 4LA_{\max}^2 \sum_k D_k + 2LA_{\max}^2 \sum_{j=0}^{R} (n_0 - B_j)_+ + 2\kappa\sigma^2 b_\delta \sum_{j=0}^{R} H(M_j, \max\{n_0, B_j\}). \tag{10}$$

If $\Pr(\mathcal{E}_{\mathrm{suf}}^c) \le p$, expected regret is bounded by the expectation of this nonnegative right-hand side, plus $2LA_{\max}^2 T(p+\delta)$. No probability bound on $\mathcal{E}_{\mathrm{suf}}$ is supplied by this proposition.

Proof. The simultaneous event. Condition on the complete covariate sequence. For each fixed homogeneous interval I with invertible G_I, unrestricted OLS satisfies $G_I^{1/2}(\tilde{a}_I - a_I)/\sigma \sim \mathcal{N}(0, I_r)$, so the Laurent–Massart bound gives

$$\Pr\Big(\Big\|G_I^{1/2}(\tilde{a}_I - a_I)\Big\|^2 > \sigma^2 b_\delta \;\Big|\; \{z_s\}\Big) \le e^{-u_\delta} = \frac{\delta}{T^2}.$$

A union bound over at most T^2 intervals needs no independence between overlapping intervals. On the resulting event, $G_I \succeq n\mu I_r/4$ implies

$$\|\tilde{a}_I - a_I\|^2 \le \frac{4\sigma^2 b_\delta}{n\mu};$$

Euclidean projection onto the coefficient ball cannot increase the distance to a_I, so the excess loss is at most

$$\frac{L}{2}\left\|\hat{a}_I - a_I\right\|^2 \le \frac{2\kappa\sigma^2 b_\delta}{n}.$$

Because the event holds for all such intervals at once, it also holds for any interval later selected from the responses.

The pathwise bound. Clipping gives excess loss at most $2LA_{\max}^2$ on every round, so the displayed contaminated-round charge is conservative. A clean run initially has B_j observations; at most $(n_0 - B_j)_+$ predictions precede a sample count of n_0. Thereafter the simultaneous event applies even though the same responses selected the retained interval. For $q \ge 1$ and $M > 0$,

$$\sum_{m=0}^{M-1} \frac{1}{q+m} \le 1 + \log\Big(1 + \frac{M}{q}\Big).$$

Summing these bounds proves the pathwise claim. On the complement of $\mathcal{E}_{\mathrm{suf}} \cap \mathcal{E}_\delta$ use the uniform clipping bound and a union bound. In particular, conditioning only on a response-selected suffix being clean does not justify the unselected OLS expectation $\sigma^2 r/n$. □

Corollary 20 (Guarded suffix retention under localization)**.** *After a declaration at t_k with retrospective split $\hat{\tau}_k$, retain observations starting at $\hat{\tau}_k + g$ for a deterministic guard $g \ge 0$; if this is later than t_k, discard all and resume from $t_k + 1$. Put $\tau_{K_T+1} = T + 1$. Suppose an event $\mathcal{E}_g$ ensures three conditions. The ordering condition: there are no false alarms and each change τ_k receives exactly one declaration t_k, which comes before the next change, $\tau_k \le t_k < \tau_{k+1}$, with delay $d_k = t_k - \tau_k + 1$. The localization condition: $|\hat{\tau}_k - \tau_k| \le g$. The clean-interval covariance condition. On $\mathcal{E}_g \cap \mathcal{E}_\delta$,*

$$\begin{aligned}\mathcal{R}_T \le\ & 4LA_{\max}^2 \sum_k d_k + 2LA_{\max}^2 n_0 \\ & + 2LA_{\max}^2 \sum_k \big(n_0 - (d_k - 2g)_+\big)_+ + \mathrm{OLS}_{\mathrm{harm}}(g),\end{aligned}$$

where $\mathrm{OLS}_{\mathrm{harm}}(g)$ *is the last term in Equation* (10)*, with $B_0 = 0$ and $B_k = (d_k - 2g)_+$. If $\Pr(\mathcal{E}_g^c) \le p$, add $2LA_{\max}^2 T(p+\delta)$ after taking expectations.*

Proof. Homogeneity of retained suffixes. By the localization condition, $\hat{\tau}_k + g \ge \tau_k$, so the retained interval $\{\hat{\tau}_k + g, \dots, t_k\}$ starts no earlier than τ_k; by the ordering condition it ends at $t_k < \tau_{k+1}$. Hence it lies inside the segment $\{\tau_k, \dots, \tau_{k+1} - 1\}$ and is homogeneous; if it is empty the learner restarts from $t_k + 1 \le \tau_{k+1}$, which is also a clean start. Its length is at least

$$\big[t_k - (\hat{\tau}_k + g) + 1\big]_+ \ge (d_k - 2g)_+.$$

Contaminated rounds. By the ordering condition, the run started at the kth restart predicts from samples of segment k only until round $\tau_{k+1} - 1$; the rounds $\tau_{k+1}, \dots, t_{k+1}$ are the only rounds at which its window mixes two segments, and there are d_{k+1} of them. No other rounds are contaminated, because there are no false alarms and every change is declared before the following one. *Conclusion.* Thus the event $\mathcal{E}_{\mathrm{suf}}$ of Proposition 19 holds with $D_k = d_k$, and substituting the lower bounds $(d_k - 2g)_+$ for B_k is valid because increasing a retained length can only decrease both displayed estimation charges. The initial burn-in term $2LA_{\max}^2 n_0$, the term with $j = 0$, remains present. The ordering condition is what the event of Lemma 17 provides for discard-all resets when every delay is below the minimum spacing; for a retaining rule it is part of the assumed event. □

Interpreting the conditional bound. For discard-all resets, each declaration pays a fresh n_0-sample burn-in. In Equation (10), a retained clean suffix of length B_j changes this cost to $(n_0 - B_j)_+$. Thus a detector with declaration delay can turn part of that delay into a warm start: if $B_j \ge n_0$, the local burn-in term disappears; if $B_j < n_0$, it is reduced by B_j samples. Classical split-estimator theory (Yao, 1987; Bai, 1997) suggests a guard of order $g = O(\sigma^2 r \log T/(\mu\Delta_0^2))$ for Corollary 20; a uniform localization event for the online GLR argmax at that guard, together with the ordering condition for the retaining rule, would make the corollary unconditional.

G Proof of Theorem 8: Minimax Lower Bound

Roadmap. The regret splits into independent per-segment estimation problems; a smooth compactly supported product prior and the van Trees inequality lower-bound the Bayes risk of each round; summing over rounds and segments gives Equation (5).

Theorem (Restatement of Theorem 8)**.** *For the linear-Gaussian model with known subspace, a partition of* $[1, T]$ *into* $K+1$ *segments of lengths* $J_0, \dots, J_K$ *revealed to the learner,* $\mu I_r \preceq \Sigma_z \preceq L I_r$ *and* $c_0 = 10r\sigma^2/(\mu A_{\max}^2)$*,*

$$\inf_{\mathcal{A}} \sup_{\|a_k\| \le A_{\max}} \mathbb{E}[\mathrm{Reg}_T] \;\ge\; \frac{r\sigma^2}{2} \sum_{k=0}^{K} \log\Big(1 + \frac{J_k}{c_0}\Big).$$

Proof. **Step 1: smooth prior construction.** Define the univariate density

$$q_B(t) = \frac{15}{16B}\Big(1 - \frac{t^2}{B^2}\Big)^2 \mathbf{1}_{\{|t|<B\}},$$

with $B = A_{\max}/\sqrt{r}$. This is C^1 with $q_B(\pm B) = 0$, ensuring the van Trees boundary term vanishes, and its Fisher information is $\int (q_B')^2/q_B = 10/B^2$. The product prior $\pi(a) = \prod_{i=1}^r q_B(a_i)$ is compactly supported in $\{a : \|a\| \le A_{\max}\}$ and has Fisher information $I_\pi = (10r/A_{\max}^2) I_r$. Draw the segment coefficients independently from π and reveal the partition to the learner; this can only help it.

Step 2: van Trees inequality. Fix a segment and its n-th round. Data from other segments are independent of this segment's coefficient a, so the Fisher information about a comes from the $n-1$ earlier rounds of the segment and has expectation $(n-1)\Sigma_z/\sigma^2$. The multivariate van Trees inequality (Van Trees, 1968; Gill & Levit, 1995) gives

$$\mathbb{E}_\pi \mathbb{E}_a[(\hat{a}_n - a)(\hat{a}_n - a)^\top] \succeq M_n^{-1},$$

where

$$M_n = I_\pi + \frac{(n-1)\Sigma_z}{\sigma^2},$$

hence

$$\mathbb{E}_\pi \mathbb{E}_a\big[\|\hat{a}_n - a\|_{\Sigma_z}^2\big] \ge \mathrm{tr}(\Sigma_z M_n^{-1}).$$

Since Bayes risk under any prior lower-bounds the supremum over coefficients, this yields a minimax lower bound, and the per-round excess loss is exactly $\frac{1}{2}\|\hat{a}_n - a\|_{\Sigma_z}^2$.

Step 3: per-segment cost. In the eigenbasis of Σ_z, whose eigenvalues satisfy $\lambda_i \ge \mu$,

$$\begin{aligned}\mathrm{tr}(\Sigma_z M_n^{-1}) &= \sum_{i=1}^{r} \frac{\sigma^2}{n - 1 + 10r\sigma^2/(A_{\max}^2 \lambda_i)} \\ &\ge \frac{r\sigma^2}{n - 1 + c_0}.\end{aligned}$$

Since $1/(n-1+c_0) \ge \int_{n-1}^{n} dx/(x + c_0)$,

$$\sum_{n=1}^{J} \frac{r\sigma^2}{2(n-1+c_0)} \;\ge\; \frac{r\sigma^2}{2} \log\Big(1 + \frac{J}{c_0}\Big).$$

Step 4: summation over segments. Summing the per-segment bounds proves the display for every partition. No Jensen step is used: the adversary may simply choose the equipartition $J_k = T/(K+1)$, which gives

$$\frac{r\sigma^2(K+1)}{2} \log\Big(1 + \frac{T}{(K+1)c_0}\Big).$$

For fixed $(r, \sigma, \mu, A_{\max})$ and $T/(K+1) \ge \max\{c_0, 1\}^2$ this is at least

$$\frac{r\sigma^2(K+1)}{4} \log \frac{T}{K+1},$$

i.e. $\Omega(\sigma^2 r \sqrt{T} \log T)$ for $K = \Theta(\sqrt{T})$. No amplitude condition growing with T is required; a large amplitude only makes c_0 small. □

G.1 Separated Well-Spaced Lower-Bound Class

The lower bound can be instantiated inside the detectable-jump class used by Theorem 4. Let $v = e_1$ and suppose $\Delta_0 \le A_{\max}/2$. For an equipartition $J = T/(K+1)$ satisfying $J \ge n_0 + 2(D + \Delta_c)$, alternate the segment prior centers between $+A_{\max}v/2$ and $-A_{\max}v/2$. Around each center use the same smooth product prior as above with $B_{\text{prior}} = A_{\max}/(8\sqrt{r})$. Every sampled coefficient remains inside the $A_{\max}$ ball, and any adjacent pair from opposite parity segments has distance at least

$$\begin{aligned} A_{\max} - 2\sqrt{r}\, B_{\text{prior}} &= \frac{3A_{\max}}{4} \\ &\ge \Delta_0. \end{aligned}$$

The van Trees argument is segment-local and translation invariant; the prior Fisher information becomes $(640r/A_{\max}^2)I_r$, so Steps 3–4 hold with c_0 replaced by

$$c_0' = \frac{640 r \sigma^2}{\mu A_{\max}^2}.$$

Thus, for fixed parameters, the separated well-spaced subclass has minimax regret at least $\frac{r\sigma^2(K+1)}{4} \log \frac{T}{K+1}$ once $T/(K+1) \ge \max\{c_0', 1\}^2$, which is $\Omega(\sigma^2 r\sqrt{T}\log T)$ for $K = \Theta(\sqrt{T})$. The restriction $\Delta_0 \le A_{\max}/2$ is needed: for $r = 1$ and $\Delta_0 = 2A_{\max}$, feasible coefficients must alternate between the two endpoints, leaving no fresh parameter to estimate.

H Proof of Theorem 12: Unknown Subspace

Roadmap. The regret splits into the fast-track window terms of Theorem 2 on the estimated subspace and a subspace term; Lemma 22 keeps the block-m projector independent of the block's data, Lemma 23 bounds the moment noise, the Fantope OGD regret bounds the moving-comparator error, and Lemma 21 converts it into a projector error.

Proof idea. For isotropic covariates, the Fantope $\mathcal{F}_r$ is the convex hull of rank-r projectors and the block objective $\langle -M_m, X\rangle$ is linear in X, so OGD on $\mathcal{F}_r$ applies without Riemannian curvature complications. The observable matrix $\hat{M}_m$ is unbiased for the trace-centered block moment; because the iterate is chosen before the block and has fixed trace, both the centering and the estimation error drop out of the expected linear regret. Under Assumption 5, the eigengap μ_a converts the linear regret into a squared-error bound on $\|X_m - Q_m\|_F^2$, rounding to $\hat{P}_m = \mathrm{Top}_r(X_m)$ costs a factor 2, and exact Gaussian orthogonality gives the projected-regression noise. The full proof follows.

Theorem (Restatement of Theorem 12)**.** *In the isotropic exact-low-rank linear-Gaussian model with blockwise-constant unknown subspaces $U_t = U_m$ for $t \in I_m$, deterministic coefficient and subspace paths independent of the observations, and the block excitation of Assumption 5, let $2r + 2 \le w \le B$. For every $\eta > 0$ the dual-timescale algorithm of Theorem 12 satisfies Equation* (9)*. In particular, for fixed $d, r, A_{\max}, \sigma, \mu_a$, bounded V_s, $B \asymp \sqrt{T}$, $w \asymp T^{1/4}$ and $K_T = \Theta(\sqrt{T})$, $\mathbb{E}[\mathrm{Reg}_T] = O(T^{3/4})$.*

Throughout, isotropy gives $\mu = L = 1$ and $L_t(\theta) - L_t(\theta_t^*) = \frac{1}{2}\|\theta - \theta_t^*\|^2$; we write $\beta_t = U_m a_t = \theta_t^*$ for $t \in I_m$, $M_m = B^{-1}\sum_{t\in I_m}\beta_t\beta_t^\top$, $\mathcal{F}_{m-1}$ for the data of blocks $1, \dots, m-1$, and $P_0(H) = H - \mathrm{tr}(H)I_d/d$. The proof combines three technical lemmas.

Lemma 21 is the special case, for the block moment, of the curvature lemma of the Fantope relaxation (Vu et al., 2013, Lemma 3.1); we include its short proof for completeness.

Lemma 21 (Eigengap curvature, isotropic design)**.** *Under Assumption 5 and the blockwise model $U_t = U_m$ in block m, the linear block objective $f_m(P) = \langle -M_m, P\rangle$ satisfies for all $P \in \mathcal{F}_r$,*

$$f_m(P) - f_m(Q_m) \ge \frac{\mu_a}{2}\|P - Q_m\|_F^2.$$

Proof. Let $M_m = U_m A_m U_m^\top$ where $A_m = \frac{1}{B}\sum_{t\in I_m} a_t a_t^\top \succeq \mu_a I_r$ by Assumption 5. The eigenvalues of M_m are $\lambda_1 \ge \cdots \ge \lambda_r \ge \mu_a$ on $\mathrm{col}(U_m)$ and $\lambda_{r+1} = \cdots = \lambda_d = 0$. Write the spectral decomposition $M_m = \sum_{i=1}^r \lambda_i v_i v_i^\top$ where $v_1, \dots, v_r$ span $\mathrm{col}(U_m)$, so $Q_m = \sum_{i=1}^r v_i v_i^\top$.

For any $P \in \mathcal{F}_r$, that is, $0 \preceq P \preceq I$ and $\mathrm{tr}(P) = r$,

$$\begin{aligned} f_m(P) - f_m(Q_m) &= \langle -M_m, P - Q_m \rangle \\ &= \sum_{i=1}^{r} \lambda_i \langle v_i v_i^\top, Q_m - P \rangle \\ &= \sum_{i=1}^{r} \lambda_i (1 - v_i^\top P v_i). \end{aligned}$$

Since $0 \preceq P \preceq I$, we have $0 \le v_i^\top P v_i \le 1$, so every summand is nonnegative.

Next,

$$\|P - Q_m\|_F^2 = \mathrm{tr}(P^2) - 2\mathrm{tr}(PQ_m) + \mathrm{tr}(Q_m^2).$$

Since Q_m is a projector, $Q_m^2 = Q_m$ and $\mathrm{tr}(Q_m^2) = r$. Since $0 \preceq P \preceq I$, $\mathrm{tr}(P^2) \le \mathrm{tr}(P) = r$. And $\mathrm{tr}(PQ_m) = \sum_{i=1}^{r} v_i^\top P v_i$. Thus

$$\begin{aligned} \|P - Q_m\|_F^2 &\le 2r - 2\sum_{i=1}^{r} v_i^\top P v_i \\ &= 2\sum_{i=1}^{r} (1 - v_i^\top P v_i). \end{aligned}$$

Combining the two displays,

$$\begin{aligned} f_m(P) - f_m(Q_m) &= \sum_{i=1}^{r} \lambda_i (1 - v_i^\top P v_i) \\ &\ge \lambda_r \sum_{i=1}^{r} (1 - v_i^\top P v_i) \\ &\ge \frac{\mu_a}{2} \|P - Q_m\|_F^2 . \end{aligned}$$

□

Lemma 22 (Past-measurable projection reduction)**.** *Fix block m and condition on $\mathcal{F}_{m-1}$, which determines an orthonormal $\hat{U}_m$ with $\hat{P}_m = \hat{U}_m \hat{U}_m^\top$. For $t \in I_m$ the projected features $z_t = \hat{U}_m^\top x_t$ are i.i.d. $\mathcal{N}(0, I_r)$ and*

$$\begin{aligned} y_t &= b_t^\top z_t + \xi_t, \\ b_t &= \hat{U}_m^\top \theta_t^*, \end{aligned}$$

where ξ_t is mean-zero Gaussian, independent of z_t, with variance $\sigma^2 + e_t^2$ and

$$e_t^2 = \left\| (I - \hat{P}_m)\theta_t^* \right\|^2 .$$

Consequently, at a round t whose previous w rounds lie in I_m and carry the same coefficient as t, the clipped projected OLS predictor satisfies

$$\mathbb{E}\big[L_t(\theta_t) - L_t(\theta_t^*) \;\big|\; \mathcal{F}_{m-1} \big] \le \frac{\sigma^2 r}{2(w - r - 1)} + e_t^2.$$

Proof. The Gaussian decomposition $x_t = \hat{U}_m z_t + \hat{U}_{m,\perp} q_t$ has $z_t \sim \mathcal{N}(0, I_r)$ and $q_t \sim \mathcal{N}(0, I_{d-r})$ independent, conditionally on $\mathcal{F}_{m-1}$, because $\hat{U}_m$ is determined before the block and the block data are independent of earlier blocks. Hence $y_t = b_t^\top z_t + \xi_t$ with $\xi_t = \theta_t^{*\top} \hat{U}_{m,\perp} q_t + \varepsilon_t$, which is Gaussian, independent of z_t, and has variance $\sigma^2 + e_t^2$. On a window of w rounds with the same coefficient, b_t and e_t^2 are constant, so, as in Section E with $\Sigma_z = I_r$, the unclipped OLS estimate satisfies

$$\mathbb{E} \left\| \tilde{b}_t - b_t \right\|^2 = \frac{(\sigma^2 + e_t^2) r}{w - r - 1}.$$

Since

$$\|b_t\| \le \|\theta_t^*\| \le A_{\max},$$

clipping does not increase this distance. Orthogonality gives

$$\left\|\hat{U}_m \hat{b}_t - \theta_t^*\right\|^2 = \left\|\hat{b}_t - b_t\right\|^2 + e_t^2,$$

so the conditional excess loss is at most

$$\frac{\sigma^2 r}{2(w-r-1)} + \frac{e_t^2}{2}\Big(1 + \frac{r}{w-r-1}\Big) \le \frac{\sigma^2 r}{2(w-r-1)} + e_t^2,$$

where the last step uses $w \ge 2r+2$. □

Lemma 23 (Gradient concentration)**.** *Let $d \ge 2$, $C_G = 14(A_{\max}^4 + \sigma^4)$ and $\hat{M}_m = \frac{1}{2B}\sum_{t\in I_m} y_t^2 P_0(x_t x_t^\top)$ as in Theorem 12. Then $\mathbb{E}\hat{M}_m = M_m^\circ := P_0(M_m)$,*

$$\mathbb{E}\|\hat{M}_m - M_m^\circ\|_F^2 \le C_G d^2/B,$$
$$\mathbb{E}\|\hat{M}_m\|_F^2 \le A_{\max}^4 + C_G d^2/B = G^2,$$

and, whenever $\mathrm{tr}X = \mathrm{tr}Q$,

$$\langle M_m^\circ, X - Q\rangle = \langle M_m, X - Q\rangle.$$

The uncentered matrix is not the target:

$$\|\mathbb{E}\hat{M}_m - M_m\|_F^2 = (\mathrm{tr}M_m)^2/d$$

does not vanish as B grows.

Proof. Step 1: mean. Isserlis' identity gives

$$\mathbb{E}[y_t^2 x_t x_t^\top] = (\|\beta_t\|^2 + \sigma^2) I_d + 2\beta_t\beta_t^\top,$$

and P_0 annihilates multiples of I_d, so

$$\mathbb{E}\big[\tfrac{1}{2} y_t^2 P_0(x_t x_t^\top)\big] = P_0(\beta_t\beta_t^\top).$$

Averaging proves $\mathbb{E}\hat{M}_m = M_m^\circ$, and

$$\mathbb{E}\hat{M}_m - M_m = -\mathrm{tr}(M_m) I_d/d$$

gives the last claim.

Step 2: one summand. For a single summand $H = \frac{1}{2} y^2 P_0(xx^\top)$ with $b = \|\beta\|$,

$$\left\|P_0(xx^\top)\right\|_F^2 = (1 - 1/d)\left\|x\right\|^4.$$

By rotation invariance take $\beta = be_1$, drop odd moments of the noise, and use

$$\mathbb{E}[x_1^4 \left\|x\right\|^4] = 3(d+4)(d+6),$$
$$\mathbb{E}[x_1^2 \left\|x\right\|^4] = (d+2)(d+4),$$
$$\mathbb{E}\left\|x\right\|^4 = d(d+2)$$

to obtain

$$\mathbb{E}\|H\|_F^2 = \frac{1-1/d}{4}\big\{3b^4(d+4)(d+6) + 6b^2\sigma^2(d+2)(d+4) + 3\sigma^4 d(d+2)\big\}.$$

For $d \ge 2$,

$$(d+4)(d+6) \le 12d^2,$$
$$(d+2)(d+4) \le 6d^2,$$
$$d(d+2) \le 2d^2,$$
$$2b^2\sigma^2 \le b^4 + \sigma^4,$$

so

$$\begin{aligned}\mathbb{E}\|H\|_F^2 &\le (13.5b^4 + 6\sigma^4)d^2 \\ &\le C_G d^2.\end{aligned}$$

Step 3: block average. The summands are independent within the block, so

$$\begin{aligned}\mathbb{E}\|\hat{M}_m - M_m^\circ\|_F^2 &= B^{-2}\sum_t \mathbb{E}\|H_t - \mathbb{E}H_t\|_F^2 \\ &\le B^{-2}\sum_t \mathbb{E}\|H_t\|_F^2 \\ &\le \frac{C_G d^2}{B}.\end{aligned}$$

Moreover

$$\begin{aligned}\|M_m^\circ\|_F &\le \|M_m\|_F \\ &\le \mathrm{tr} M_m \\ &\le A_{\max}^2,\end{aligned}$$

so

$$\begin{aligned}\mathbb{E}\|\hat{M}_m\|_F^2 &= \|M_m^\circ\|_F^2 + \mathbb{E}\|\hat{M}_m - M_m^\circ\|_F^2 \\ &\le G^2.\end{aligned}$$

The trace identity follows from $\langle I_d, X - Q\rangle = \mathrm{tr}(X - Q) = 0$. □

Proof of Theorem 12. **Step 1: Fantope OGD with moving comparators.** Since $Q_m \in \mathcal{F}_r$ and Euclidean projection onto $\mathcal{F}_r$ is nonexpansive,

$$\begin{aligned}\|X_{m+1} - Q_m\|_F^2 \le \|X_m - Q_m\|_F^2 + 2\eta\langle \hat{M}_m, X_m - Q_m\rangle \\ + \eta^2 \left\|\hat{M}_m\right\|_F^2.\end{aligned}$$

Rearranging and summing over $m = 1, \dots, M$,

$$\begin{aligned}\sum_{m=1}^M \langle \hat{M}_m, Q_m - X_m\rangle \le \frac{1}{2\eta}\sum_{m=1}^M \left(\|X_m - Q_m\|_F^2 - \|X_{m+1} - Q_m\|_F^2\right) \\ + \frac{\eta}{2}\sum_{m=1}^M \left\|\hat{M}_m\right\|_F^2.\end{aligned}$$

For $X, Y \in \mathcal{F}_r$, whose eigenvalues lie in $[0, 1]$ and which satisfy $\langle X, Y\rangle \ge 0$,

$$\begin{aligned}\|X - Y\|_F^2 &\le \|X\|_F^2 + \|Y\|_F^2 \\ &\le 2r,\end{aligned}$$

so $\mathrm{diam}(\mathcal{F}_r) \le \sqrt{2r}$. Rewriting the telescoping sum, the first term is at most $\|X_1 - Q_1\|_F^2 \le 2r$ plus

$$\begin{aligned}\sum_{m<M} \left(\|X_{m+1} - Q_{m+1}\|_F^2 - \|X_{m+1} - Q_m\|_F^2\right) \\ \le 2\sqrt{2r}\sum_{m<M} \|Q_{m+1} - Q_m\|_F.\end{aligned}$$

Hence, with $R = r + \sqrt{2r}\, V_s$,

$$\sum_{m=1}^M \langle \hat{M}_m, Q_m - X_m\rangle \le \frac{R}{\eta} + \frac{\eta}{2}\sum_{m=1}^M \left\|\hat{M}_m\right\|_F^2. \tag{11}$$

The initial iterate enters only through the r in R.

Step 2: unbiasedness against a predictable direction. X_m is $\mathcal{F}_{m-1}$-measurable, Q_m and M_m are deterministic, and the block-m data are independent of $\mathcal{F}_{m-1}$, so

$$\mathbb{E}[\langle \hat{M}_m - M_m^\circ, Q_m - X_m \rangle \mid \mathcal{F}_{m-1}] = 0.$$

Since $\mathrm{tr} X_m = \mathrm{tr} Q_m = r$, Lemma 23 gives

$$\langle M_m^\circ, Q_m - X_m \rangle = \langle M_m, Q_m - X_m \rangle,$$
$$\mathbb{E} \left\| \hat{M}_m \right\|_F^2 \le G^2.$$

Taking expectations in (11),

$$\sum_{m=1}^{M} \mathbb{E}\langle M_m, Q_m - X_m \rangle \le \Psi_\eta,$$

where

$$\Psi_\eta := \frac{R}{\eta} + \frac{\eta}{2} M G^2.$$

The matrix noise therefore enters only through G^2; no step-size-free noise term is left over.

Step 3: eigengap conversion. By Lemma 21,

$$\langle M_m, Q_m - X_m \rangle \ge \frac{\mu_a}{2} \left\| X_m - Q_m \right\|_F^2,$$

so

$$\sum_m \mathbb{E} \left\| X_m - Q_m \right\|_F^2 \le \frac{2\Psi_\eta}{\mu_a}.$$

Step 4: spectral rounding. Among rank-r projectors P,

$$\|P - X_m\|_F^2 = r - 2\langle P, X_m \rangle + \|X_m\|_F^2$$

is minimized by $\hat{P}_m = \mathrm{Top}_r(X_m)$. As Q_m is also a rank-r projector,

$$\begin{aligned} \left\| \hat{P}_m - Q_m \right\|_F &\le \left\| \hat{P}_m - X_m \right\|_F + \left\| X_m - Q_m \right\|_F \\ &\le 2 \left\| X_m - Q_m \right\|_F, \end{aligned}$$

hence

$$\sum_m \mathbb{E} \left\| \hat{P}_m - Q_m \right\|_F^2 \le \frac{8\Psi_\eta}{\mu_a}.$$

Step 5: fast track under past-measurable projections. In each block the first w rounds predict zero and cost at most $A_{\max}^2/2$ each, in total $A_{\max}^2 M w/2$. A later round whose window contains a coefficient change is charged to that change as in Section E; there are at most $K_T w$ such rounds, each costing at most $2A_{\max}^2$ by clipping. Every other round of block m is covered by Lemma 22, and, because $(I - \hat{P}_m)Q_m = (Q_m - \hat{P}_m)Q_m$,

$$\begin{aligned} e_t^2 &= \left\| (\hat{P}_m - Q_m) Q_m \beta_t \right\|^2 \\ &\le A_{\max}^2 \left\| \hat{P}_m - Q_m \right\|_F^2. \end{aligned}$$

Rounds after MB cost at most $2A_{\max}^2$ each. Hence

$$\begin{aligned} \mathbb{E}[\mathrm{Reg}_T] \le \frac{\sigma^2 r T}{2(w - r - 1)} &+ 2A_{\max}^2 K_T w + \frac{A_{\max}^2 M w}{2} \\ &+ A_{\max}^2 B \sum_{m=1}^{M} \mathbb{E} \left\| \hat{P}_m - Q_m \right\|_F^2 + 2A_{\max}^2 (T - MB). \end{aligned}$$

Step 6: combination. Step 4 bounds the fourth term by $(8A_{\max}^2 B/\mu_a)\Psi_\eta$, which is Equation (9).

Step 7: step size. Choose

$$\eta = \sqrt{\frac{2R}{MG^2}},$$

so that

$$\Psi_\eta = \sqrt{2RMG^2}.$$

Using

$$MB \le T,$$
$$G \le A_{\max}^2 + d\sqrt{C_G/B},$$

we get

$$\frac{8A_{\max}^2 B}{\mu_a}\sqrt{2RMG^2} \le \frac{8\sqrt{2}\,A_{\max}^4}{\mu_a}\sqrt{BTR} + \frac{8\sqrt{2}\,A_{\max}^2 d}{\mu_a}\sqrt{C_G TR}.$$

For fixed $d, r, A_{\max}, \sigma, \mu_a$ and bounded V_s, taking $B \asymp \sqrt{T}$, $w \asymp T^{1/4}$ and $K_T = \Theta(\sqrt{T})$ makes T/w, $K_T w$, Mw and $\sqrt{BT}$ all $O(T^{3/4})$, while the matrix-noise term is $O(d\sqrt{T})$ and $T - MB < B$. □

I Proof of Theorem 10 and Proposition 11: Fixed Unknown Subspace

Roadmap. Lemma 24 bounds the error of the spectral moment of the subspace rounds with high probability, and the Davis–Kahan theorem turns it into a bound on the principal angles of $\hat{U}_j$. Conditionally on the subspace rounds, the regression rounds seen through $\hat{U}_j$ form an r-dimensional instance of the model of Theorem 4 whose noise level varies between segments; Lemma 25 shows that Lemma 17 still holds for it. The proof of Theorem 10 then charges each round by a warm-up term before the estimated subspace is accurate, the per-round charges of Theorem 4 for the single-map run in force, the part of θ_t^* outside $\mathrm{col}(\hat{U}_j)$, and failure events. Proposition 11 applies the van Trees argument of Section G in dimension d to the first r segments and in dimension r to the others.

Notation. Rounds come in pairs $(2i-1, 2i)$, $i \ge 1$; round $2i-1$ is a subspace round and round $2i$ a regression round. Let $\mathcal{O} = \sigma(x_s, \varepsilon_s : s \text{ odd})$. The epoch starts are $T_j = 2^j t_s + 1$, which are odd since t_s is even, O_{T_j} holds the $n_j = 2^{j-1}t_s$ subspace rounds before T_j, and $\hat{M}_j = n_j^{-1}\sum_{s\in O_{T_j}} y_s^2(x_s x_s^\top - I_d)$. Both rounds of a pair lie in the same epoch and use the same prediction, which depends on the regression rounds before the pair. Write $v = A_{\max}^2 + \sigma^2$, $u = \log(8dT^3)$, $\hat{P}_j = \hat{U}_j\hat{U}_j^\top$ and $\epsilon_j = \left\|(I - \hat{P}_j)U\right\|_{\mathrm{op}}$, the sine of the largest principal angle between $\mathrm{col}(\hat{U}_j)$ and $\mathrm{col}(U)$. For isotropic covariates the excess loss is $\frac{1}{2}\left\|\theta - \theta_t^*\right\|^2$, and for $\theta = \hat{U}_j\hat{a}$ the two parts are orthogonal:

$$\tfrac{1}{2}\left\|\hat{U}_j\hat{a} - Ua_t\right\|^2 = \tfrac{1}{2}\left\|\hat{a} - \hat{U}_j^\top U a_t\right\|^2 + \tfrac{1}{2}\left\|(I - \hat{P}_j)Ua_t\right\|^2. \tag{12}$$

Lemma 24 (Moment concentration)**.** *Let $x_s \sim \mathcal{N}(0, I_d)$ and $\varepsilon_s \sim \mathcal{N}(0, \sigma^2)$ be independent over n rounds, $y_s = \theta_s^\top x_s + \varepsilon_s$ with deterministic $\|\theta_s\| \le A_{\max}$, and $\bar{M} = n^{-1}\sum_s 2\theta_s\theta_s^\top$. For every $u \ge 1$, with probability at least $1 - (3n + 2d)e^{-u}$,*

$$\left\|\hat{M} - \bar{M}\right\|_{\mathrm{op}} \le v\Big(\sqrt{\frac{52(d+5)u}{n}} + \frac{8u^2(2d+3u)}{3n} + 5.6(d+8)e^{-u/2}\Big),$$

where $\hat{M} = n^{-1}\sum_s y_s^2(x_s x_s^\top - I_d)$.

Proof. **Mean.** With $v_s = \|\theta_s\|^2 + \sigma^2 \le v$, Gaussian moments give

$$\mathbb{E}[y_s^2 x_s x_s^\top] = v_s I_d + 2\theta_s\theta_s^\top,$$
$$\mathbb{E}[y_s^2] = v_s,$$

so $\mathbb{E}\hat{M} = \bar{M}$.

Truncation. Let

$$B_s = \Big\{y_s^2 \le 2v_s u\Big\} \cap \Big\{\|x_s\|^2 \le d + 2\sqrt{du} + 2u\Big\}.$$

The Gaussian tail and the Laurent–Massart bound give $\Pr(B_s^c) \le 2e^{-u} + e^{-u} = 3e^{-u}$. Put $W_s = y_s^2(x_s x_s^\top - I_d)$ and $\tilde{W}_s = W_s \mathbf{1}_{B_s}$. Since $\|x_s x_s^\top - I_d\|_{\mathrm{op}} = \max\{\|x_s\|^2 - 1, 1\}$ and $2\sqrt{du} \le d + u$,

$$\Big\|\tilde{W}_s\Big\|_{\mathrm{op}} \le R_0 := 2vu(2d + 3u).$$

On $\bigcap_s B_s$, which has probability at least $1 - 3ne^{-u}$, we have $\hat{M} = n^{-1}\sum_s \tilde{W}_s$.

Variance. For a unit vector w, Cauchy–Schwarz gives

$$\mathbb{E}\big[y_s^4 \big\|(x_s x_s^\top - I)w\big\|^2\big] \le \sqrt{\mathbb{E}y_s^8}\,\sqrt{\mathbb{E}\,\|(x_s x_s^\top - I)w\|^4}.$$

Here $\mathbb{E}y_s^8 = 105v_s^4$. Also $\big\|(xx^\top - I)w\big\|^2 = (w^\top x)^2(\|x\|^2 - 2) + 1 \le (w^\top x)^2\,\|x\|^2 + 1$, and with $g = w^\top x$ and $S = \|x\|^2 - g^2 \sim \chi^2_{d-1}$ independent of g,

$$\begin{aligned}\mathbb{E}[g^4\,\|x\|^4] &= \mathbb{E}g^8 + 2\mathbb{E}g^6\,\mathbb{E}S + \mathbb{E}g^4\,\mathbb{E}S^2\\ &= 105 + 30(d-1) + 3(d^2-1)\\ &= 3d^2 + 30d + 72.\end{aligned}$$

Hence $\mathbb{E}\,\big\|(xx^\top - I)w\big\|^4 \le 2(3d^2 + 30d + 72) + 2 \le 6(d+5)^2$ and

$$\mathbb{E}\big[y_s^4 w^\top (x_s x_s^\top - I)^2 w\big] \le \sqrt{630}\,v^2(d+5) \le 26v^2(d+5).$$

Because $\tilde{W}_s^2 \preceq W_s^2$, the centered summands $X_s = \tilde{W}_s - \mathbb{E}\tilde{W}_s$ satisfy $\big\|\sum_s \mathbb{E}X_s^2\big\|_{\mathrm{op}} \le 26nv^2(d+5)$ and $\|X_s\|_{\mathrm{op}} \le 2R_0$.

Matrix Bernstein. The matrix Bernstein inequality (Tropp, 2012, Theorem 1.4), applied to $\pm\sum_s X_s$, gives with probability at least $1 - 2de^{-u}$

$$\Big\|\sum_s X_s\Big\|_{\mathrm{op}} \le \sqrt{52nv^2(d+5)u} + \tfrac{4}{3}R_0 u,$$

because the exponent $\frac{\tau^2/2}{\sigma_B^2 + R\tau/3}$ is at least u at $\tau = \sqrt{2\sigma_B^2 u} + \frac{2}{3}Ru$.

Truncation bias. By Cauchy–Schwarz,

$$\begin{aligned}\Big\|\mathbb{E}\tilde{W}_s - \mathbb{E}W_s\Big\|_{\mathrm{op}} &\le \sqrt{\mathbb{E}\big[y_s^4(\|x_s\|^2 + 1)^2\big]}\,\sqrt{\Pr(B_s^c)}\\ &\le 105^{1/4} v\,\big(\mathbb{E}(\|x_s\|^2 + 1)^4\big)^{1/4}\sqrt{3}\,e^{-u/2}\\ &\le 5.6\,v(d+8)e^{-u/2},\end{aligned}$$

using $\mathbb{E}(\chi_d^2 + 1)^4 \le (d+8)^4$, which follows by expanding with $\mathbb{E}(\chi_d^2)^k = d(d+2)\cdots(d+2k-2)$.

Combine. On both events, $\hat{M} - \bar{M} = n^{-1}\sum_s X_s + n^{-1}\sum_s(\mathbb{E}\tilde{W}_s - \mathbb{E}W_s)$, and dividing the two bounds by n gives the claim. □

Lemma 25 (Detection with segment-wise noise levels)**.** *Fix j and condition on $\mathcal{O}$ with $\epsilon_j \le \epsilon \le 1/2$. Let $z_s = \hat{U}_j^\top x_s$ for the regression rounds s. Then the regression rounds form independent observations*

$$y_s = b_s^\top z_s + \xi_s, \qquad z_s \sim \mathcal{N}(0, I_r),$$
$$\xi_s \sim \mathcal{N}(0, \sigma_s^2) \text{ independent of } z_s,$$

with $b_s = \hat{U}_j^\top U a_s$ *and* $\sigma_s^2 = \sigma^2 + \left\|(I - \hat{P}_j)U a_s\right\|^2 \le \sigma^2 + \epsilon^2 A_{\max}^2 \le \bar{\sigma}^2$. *The path* b_s *changes exactly when* a_s *changes,* $\|b_s\| \le A_{\max}$, *each jump has size at least* $\sqrt{1-\epsilon^2}\,\Delta_k$, *and* σ_s *is constant on every segment. If every segment holds at least* $n_0 + 2(\bar{D} + \Delta_c)$ *regression rounds, the learner of Definition 3 with noise level* $\bar{\sigma}$ *in Equation* (3)*, run on these rounds with this single map, satisfies the conclusions of Lemma 17 with* $\mu = L = 1$, σ *replaced by* $\bar{\sigma}$ *and* Δ_0 *by* $\sqrt{1-\epsilon^2}\Delta_0$, *i.e. with delay at most* $\bar{D} + \Delta_c$ *for* $\bar{D} = C_D\big(n_0 + r\bar{\sigma}^2 \log T/((1-\epsilon^2)\Delta_0^2)\big)$, *with conditional probability at least* $1 - 1/T$.

Proof. **Model.** Given $\mathcal{O}$, $\hat{U}_j$ is fixed and the regression rounds are independent of $\mathcal{O}$. For $x \sim \mathcal{N}(0, I_d)$ the Gaussian vectors $\hat{U}_j^\top x$ and $(I - \hat{P}_j)x$ are uncorrelated, hence independent, and $y_s = (\hat{U}_j^\top U a_s)^\top z_s + \xi_s$ with $\xi_s = ((I - \hat{P}_j)U a_s)^\top x_s + \varepsilon_s$. This gives the stated law of ξ_s; σ_s depends on s only through a_s. The jump bound follows from $\sigma_{\min}(\hat{U}_j^\top U) = \sqrt{1-\epsilon_j^2}$, and $\|b_s\| \le \|a_s\|$.

Changes to the proof of Lemma 17. The covariance event and the list of triples are unchanged, and the noise enters only through two tail bounds. First, a triple without a change lies in one segment, so $2\Lambda = \bar{\sigma}^{-2}\xi^\top(P_1 - P_0)\xi$ equals $(\sigma_k^2/\bar{\sigma}^2)$ times a χ_q^2 variable, $q \le r$, and $\Pr(2\Lambda \ge 2h_{\min}) \le \Pr(\chi_q^2 \ge 2h_{\min}) \le e^{-u}$. Second, at a true split let $Q = P_1 - P_0$, m the mean vector and $\bar{\lambda} = \bar{\sigma}^{-2}\|Qm\|^2 = \bar{\sigma}^{-2}\delta^\top(G_L^{-1} + G_R^{-1})^{-1}\delta$. Since Q is an orthogonal projector,

$$\bar{\sigma}^2\, 2\Lambda = \|Qm + Q\xi\|^2 \ge \|Qm\|^2 + 2\langle Qm, \xi\rangle,$$

and $\langle Qm, \xi\rangle$ is centered Gaussian with variance at most $\bar{\sigma}^2\|Qm\|^2$. Hence, with probability at least $1 - e^{-u}$,

$$\begin{aligned} 2\Lambda &\ge \bar{\lambda} - 2\sqrt{2\bar{\lambda}u} \\ &\ge \bar{\lambda}/2 - 4u, \end{aligned}$$

which exceeds $2h$ once $\bar{\lambda} \ge 4h + 8u$, a weaker requirement than the $\lambda \ge 4h + 10u$ used in Lemma 17. The rest of that proof applies verbatim with $\mu = 1$, $\bar{\sigma}$ and $\|\delta\| \ge \sqrt{1-\epsilon^2}\Delta_0$. □

Proof of Theorem 10. **Step 1: subspace event.** Let $\mathcal{G}$ be the event on which the bound of Lemma 24 holds, with $u = \log(8dT^3)$, for every epoch start $T_j \le T$. There are at most $\log_2 T + 1 \le T$ epochs and $3n_j + 2d \le 2dT$, so

$$\Pr(\mathcal{G}^c) \le T \cdot 2dT \cdot \frac{1}{8dT^3} = \frac{1}{4T}.$$

Because $e^{-u/2} = (8dT^3)^{-1/2}$ and $n_j \le T$, the third term of Lemma 24 is at most its first for $T \ge 3$. Hence on $\mathcal{G}$, for every epoch,

$$\left\|\hat{M}_j - \bar{M}_j\right\|_{\mathrm{op}} \le v\sqrt{\frac{208(d+5)u}{n_j}} + \frac{8vu^2(2d+3u)}{3n_j}.$$

If also $T_j \ge t_e$, then $\bar{M}_j = 2U\bar{A}_j U^\top$ with $\lambda_{\min}(\bar{A}_j) \ge \mu_a$ by Assumption 4, so the gap between the r-th and $(r+1)$-th eigenvalues of $\bar{M}_j$ is at least $2\mu_a$. The Davis–Kahan theorem in the form of Yu et al. (2015, Theorem 2) gives $\epsilon_j \le \left\|\sin\Theta(\hat{U}_j, U)\right\|_F \le 2\sqrt{r}\left\|\hat{M}_j - \bar{M}_j\right\|_{\mathrm{op}}/(2\mu_a)$, hence, with the constants C_U and R_U of Theorem 10,

$$\epsilon_j \le \sqrt{\frac{C_U}{n_j}} + \frac{R_U}{n_j}, \qquad \epsilon_j^2 \le \frac{2C_U}{n_j} + \frac{2R_U^2}{n_j^2}. \tag{13}$$

Step 2: warm-up. Let $m_\epsilon = \max\{4C_U/\epsilon^2,\, 2R_U/\epsilon\}$ and let j_1 be the first epoch with $T_{j_1} \ge t_e$ and $n_{j_1} \ge m_\epsilon$; both conditions persist for $j \ge j_1$, and each term of the first bound in Equation (13) is then at most $\epsilon/2$, so on $\mathcal{G}$ every such epoch has $\epsilon_j \le \epsilon$. If $j_1 \ge 1$, the epoch $j_1 - 1$ violates one condition and $T_{j_1} - 1 = 2(T_{j_1-1} - 1)$, so $T_{j_1} < 2t_e$ or $T_{j_1} = 2n_{j_1} + 1 < 4m_\epsilon + 1$; if $j_1 = 0$, $T_{j_1} = t_s + 1$. Hence $T_{j_1} \le T_\star$; if no epoch $j \ge j_1$ starts before T, every round is covered by this

step. Every prediction satisfies $\|\theta_t\| \le A_{\max}$, so each round before T_{j_1} costs at most $2A_{\max}^2$, in total at most $2A_{\max}^2 T_\star$.

Step 3: per-round charges of the single-map runs. For each j let $\hat{a}_i^{(j)}$ denote the prediction, before regression round $2i$, of the learner of Definition 3 run with the single map $\hat{U}_j$ on all regression rounds; by Definition 9 the prediction in pair i of epoch j is $\hat{U}_j \hat{a}_i^{(j)}$. Index regression rounds by i and let τ_k^{r} be the first regression index of segment k and $n(i) = i - \tau_k^{\mathrm{r}}$ for i in segment k, with $\tau_0^{\mathrm{r}} = 1$. Define the deterministic charges

$$c_i = \begin{cases} 2A_{\max}^2, & n(i) < \bar{D} + \Delta_c + n_0 \ (k \ge 1) \text{ or } n(i) < n_0 \ (k = 0), \\ \dfrac{\bar{\sigma}^2 r}{n(i) - (\bar{D} + \Delta_c)\mathbf{1}_{k \ge 1}}, & \text{otherwise.} \end{cases}$$

Fix $j \ge j_1$ and condition on $\mathcal{O} \in \mathcal{G}$. By Lemma 25, the single-map run satisfies the detection event $\mathcal{E}_{\mathrm{det}}^{(j)}$ of Lemma 17 with conditional probability at least $1 - 1/T$. Repeating Parts 1–3 of the proof of Theorem 4 for this run, with the ghost estimate of Part 3,

$$\mathbb{E}\Big[\tfrac{1}{2}\big\|\hat{a}_i^{(j)} - b_{2i}^{(j)}\big\|^2 \mathbf{1}_{\mathcal{E}_{\mathrm{det}}^{(j)}} \,\Big|\, \mathcal{O}\Big] \le c_i,$$

where $b_s^{(j)} = \hat{U}_j^\top U a_s$: in the first case by clipping; in the second, the reset of segment k lies at most $\bar{D} + \Delta_c$ regression rounds after τ_k^{r}, so the ghost estimate uses $m \ge n(i) - (\bar{D} + \Delta_c)\mathbf{1}_{k \ge 1} \ge n_0 \ge 2r + 2$ fresh rounds of noise level at most $\bar{\sigma}$, and its risk is at most $\bar{\sigma}^2 r/(2(m - r - 1)) \le \bar{\sigma}^2 r/m$. The charge c_i does not depend on j. The odd round of pair i has the coefficient of round $2i$ unless a change occurs at round $2i$, which happens for at most K_T pairs and costs at most $2A_{\max}^2$ each.

Step 4: subspace part. For $j \ge j_1$ on $\mathcal{G}$, every round of epoch j has $\frac{1}{2}\left\|(I - \hat{P}_j)U a_t\right\|^2 \le \frac{1}{2}A_{\max}^2 \epsilon_j^2$, and epoch j has $T_{j+1} - T_j = 2n_j$ rounds, so by Equation (13) it costs at most $2A_{\max}^2 C_U + 2A_{\max}^2 R_U^2/n_j$. At most $\log_2(2T/t_s)$ epochs start before T because $2 \le t_s \le T$, and since n_j doubles from $n_{j_1} \ge 2R_U/\epsilon$, $\sum_{j \ge j_1} 1/n_j \le 2/n_{j_1} \le \epsilon/R_U$. The subspace part is therefore at most $2A_{\max}^2 C_U \log_2(2T/t_s) + 2A_{\max}^2 \epsilon R_U$.

Step 5: failure events. On $\mathcal{G}^c$ or on the complement of $\mathcal{E}_{\mathrm{det}}^{(j(t))}$ a round costs at most $2A_{\max}^2$, so these events add at most $2A_{\max}^2 T\big(\frac{1}{4T} + \frac{1}{T}\big) \le 3A_{\max}^2$.

Step 6: summation. By Equation (12) and Steps 2–5,

$$\mathbb{E}[\mathrm{Reg}_T] \le 2A_{\max}^2 T_\star + 2\sum_i c_i + 2A_{\max}^2 K_T + 2A_{\max}^2 C_U \log_2 \frac{2T}{t_s} + 2A_{\max}^2 \epsilon R_U + 3A_{\max}^2.$$

Within segment $k \ge 1$ the first case of c_i covers at most $\bar{D} + \Delta_c + n_0$ indices and the second case sums to at most $\bar{\sigma}^2 r \sum_{m=n_0}^{J_k^{\mathrm{r}}} 1/m \le \bar{\sigma}^2 r \log(e J_k^{\mathrm{r}}/n_0)$, where J_k^{r} is the number of regression rounds of the segment; segment 0 is the same without $\bar{D} + \Delta_c$. By concavity of $\log$ and $\sum_k J_k^{\mathrm{r}} \le T/2$,

$$\sum_i c_i \le 2A_{\max}^2 K_T(\bar{D} + \Delta_c) + 2A_{\max}^2 (K_T + 1) n_0 + (K_T + 1)\bar{\sigma}^2 r \log \frac{eT}{(K_T + 1)n_0},$$

and substituting gives the bound of Theorem 10.

Step 7: order. For fixed $(r, \sigma, A_{\max}, \Delta_0, \mu_a, \epsilon)$, $K_T = \Theta(\sqrt{T})$, $t_e = O(\sqrt{T})$ and $t_s = O(\sqrt{T})$, the first line of the bound is $O\big((\bar{\sigma}^2 \rho_D + A_{\max}^2) r \sqrt{T} \log T\big)$ as in Theorem 4 with $\kappa = 1$. With $u = \log(8dT^3)$, $C_U = O(rv^2 du/\mu_a^2)$ and $R_U = O(\sqrt{r}\, v(d + u)u^2/\mu_a)$, so $C_U \log_2(2T/t_s)$, C_U/ϵ^2, R_U/ϵ and ϵR_U are all $O\big((rv^2/\mu_a^2 + \sqrt{r}\, v/\mu_a)\, d \log^2(dT)\big)$ whenever $\log(dT) \le d$, and at most $O(\log^3(dT))$ otherwise; every one of them is a cost of learning U and none grows with K_T. □

Proof of Proposition 11. **Family.** Let J be even, $K+1 \geq 2r$ and $T = (K+1)J$, with segment k on rounds $kJ+1, \ldots, (k+1)J$. Draw $v_0, \ldots, v_{r-1}$ independently from the product prior on $[-w, w]^d$ with the density q_w of Section G in each coordinate, $w = A_{\max}/(4\sqrt{rd})$, and put $\theta_k = \frac{A_{\max}}{2} e_{k+1} + v_k$ for $k < r$. Then $\|v_k\| \leq A_{\max}/(4\sqrt{r})$, the matrix $V = [v_0, \ldots, v_{r-1}]$ has $\|V\|_F \leq A_{\max}/4$, and $\Theta = [\theta_0, \ldots, \theta_{r-1}]$ has $\sigma_{\min}(\Theta) \geq A_{\max}/2 - A_{\max}/4 = A_{\max}/4$. Let U be an orthonormal basis of $\mathrm{col}(\Theta)$, chosen as a fixed measurable function of Θ, and $a_k^\circ = U^\top \theta_k$. For segments $k \geq r$ draw g_k independently from the product prior on $[-w', w']^r$ with $w' = A_{\max}/(16r)$, so $\|g_k\| \leq A_{\max}/(16\sqrt{r})$, and set $a_k = (-1)^{\lfloor k/r \rfloor}(a^\circ_{k \bmod r} + g_k)$; for $k < r$ let $a_k = a_k^\circ$. All priors are independent of the design and the noise.

Membership. $\|a_k\| \leq A_{\max}/2 + A_{\max}/4 + A_{\max}/16 < A_{\max}$. For $r \geq 2$, consecutive coefficients differ by $\pm(a^\circ_{j+1} - a^\circ_j + g - g')$ inside a cycle and by $\pm(a^\circ_{r-1} + a^\circ_0 + g + g')$ between cycles; since $\|\theta_{j+1} \mp \theta_j\| \geq A_{\max}/\sqrt{2} - A_{\max}/(2\sqrt{r}) \geq A_{\max}/(2\sqrt{2})$ and $\|g\| + \|g'\| \leq A_{\max}/(8\sqrt{2})$, every jump is at least $A_{\max}/5$. For $r = 1$ the jumps are $\|2a_0^\circ \pm (g + g')\| \geq 2(A_{\max}/4) - A_{\max}/8 \geq A_{\max}/5$. For excitation, every complete cycle $\{a_k\}_{k=mr}^{mr+r-1}$ satisfies $\sum_k a_k a_k^\top \succeq (A_{\max}/4 - A_{\max}/16)^2 I_r = \frac{9A_{\max}^2}{256} I_r$, because $[a_0^\circ, \ldots, a_{r-1}^\circ] = U^\top \Theta$ has smallest singular value at least $A_{\max}/4$ and the perturbations have Frobenius norm at most $A_{\max}/16$. Each segment has $J/2$ odd rounds. Let $t \geq 2rJ + 1$ and let $S = \lfloor (t-1)/J \rfloor \geq 2r$ be the number of complete segments before t. They contain $C = \lfloor S/r \rfloor \geq S/(2r) \geq (t-1)/(4rJ)$ complete cycles, each contributing at least $\frac{J}{2} \cdot \frac{9A_{\max}^2}{256} I_r$ to $\sum_{s \in O_t} a_s a_s^\top$, and $|O_t| \leq t/2$; hence

$$\lambda_{\min}\Big(\frac{1}{|O_t|} \sum_{s \in O_t} a_s a_s^\top\Big) \geq \frac{t-1}{t} \cdot \frac{9A_{\max}^2}{1024r} \geq \frac{A_{\max}^2}{250r},$$

i.e. Assumption 4 holds with $\mu_a = A_{\max}^2/(250r)$ and $t_e = 2rJ + 1$.

Lower bound. Fix an algorithm. In round n of segment $k < r$, θ_k is independent of all earlier segments, whose parameters are independent of v_k, and the Fisher information of the $n-1$ earlier rounds of the segment is $(n-1)I_d/\sigma^2$; the prior has information $(10/w^2)I_d = (160rd/A_{\max}^2)I_d$. The van Trees inequality of Section G, in dimension d with $\Sigma_z = I_d$, bounds the Bayes risk of the round by $d\sigma^2/(2(n - 1 + c_d))$, and summing over $n \leq J$ gives $\frac{d\sigma^2}{2} \log(1 + J/c_d)$ per segment. In round n of a segment $k \geq r$, reveal $(v_0, \ldots, v_{r-1})$, hence U and all a°, to the learner, which can only lower its Bayes risk; the remaining unknown g_k is independent of all other data, and $\frac{1}{2}\|\theta - U a_k\|^2 \geq \frac{1}{2}\|U^\top \theta - a_k\|^2$, so the r-dimensional argument of Section G with prior information $(10/w'^2)I_r = (2560r^2/A_{\max}^2)I_r$ gives $\frac{r\sigma^2}{2} \log(1 + J/c_r)$ per segment. The Bayes risk lower-bounds the supremum over the family, and adding the r segments of the first kind and the $K + 1 - r$ of the second gives Equation (7). □

J Extended Related Work

Differences from the closest analyses. *SVUCRL* (Dam, 2025) studies non-stationary reinforcement learning in a communicating MDP whose transition drift is low-rank plus sparse; its non-stationarity terms scale with the drift rank and give $\sqrt{T}$-type dynamic regret upper bounds. Its mechanism is low-rank tracking with robust PCA, confidence widening and forecasting, without a detect-and-reset component. For passive regression, Theorems 4 and 8 form an upper/lower pair of the same order on one jump class.

Change-point dynamic pricing (Zhao et al., 2026) studies an active pricing problem with sparsity as the exploited structure and regret measured in expectation against a clairvoyant seller. Our analysis concerns passive tracking on a known r-dimensional subspace, with a finite-sample expected-regret bound in which false alarms, detection delay and post-reset estimation enter one decomposition, together with the learner-class separation into $\Omega(T)$ monotone-step, $O(T^{3/4})$ fixed-window and $O(\sqrt{T \log T})$ adaptive rates across one observation model.

GLR-CUCB (Zhou et al., 2020) is a piecewise-stationary combinatorial semi-bandit method analyzed in expected α-approximation regret; its $\sqrt{NKT \log T}$-type rate is the branch in which the number of stationary segments N is known. Theorem 4 needs no knowledge of the growth rate of K_T, and Theorem 8 is compared with it, via Corollary 6, on the same well-spaced fixed-gap class.

ISD-linUCB (Lazzaretto et al., 2026) studies linear contextual bandits under slowly varying drift, learning an invariant subspace from offline historical data and exploring online only in the residual subspace. Our setting has abrupt coefficient jumps, a detection cost inside the regret bound, and a lower bound over the jump class.

Multiple change points in mean tracking. Gafni et al. (2026) study expected cumulative squared-error dynamic regret for tracking a mean with multiple change points by detection and restart, with nearly matching upper and lower bounds; an appendix of their paper outlines an extension to vector-valued observations, with the full argument omitted. In both cases the tracked parameter is observed directly up to additive noise. For isotropic designs, $\Sigma_z = I_r$ with known U, a coordinatewise reduction to their setting is available: the statistic $V_{t,i} = \sqrt{\pi/2}\, y_t \operatorname{sgn}(z_{t,i})$ satisfies $\mathbb{E}[V_{t,i}] = a_{t,i}$, it is independent and identically distributed within a segment, and $V_{t,i} - a_{t,i}$ is sub-Gaussian with variance proxy of order $A_{\max}^2 + \sigma^2$. Their scalar tracker run on each of the r coordinates, with the estimate projected onto the $A_{\max}$-ball, therefore has regret of order $r(A_{\max}^2 + \sigma^2)(K_T + 1)\log T$ on this subclass, the order of Theorem 4 at $K_T = \Theta(\sqrt{T})$. For a general design covariance with $\mu I \preceq \Sigma_z \preceq LI$ the same statistic has mean $(\Sigma_z a_t)_i / \sqrt{(\Sigma_z)_{ii}}$, which mixes the coordinates through the unknown covariance, and the results here address this regression structure directly. They give the guarantee of the GLR least-squares learner for unknown anisotropic covariances with explicit dependence on r, κ and σ, its per-jump transient uniform in the jump size, the lower bound $\Omega(\sqrt{T}\log T)$ with per-segment cost $r\sigma^2 \log \frac{T}{K_T+1}$ for the regression class, the polynomial separation from every fixed window and from monotone steps, and the additive cost of an unknown subspace together with its lower bound. Here an r-vector is observed only through scalar responses $y_t = z_t^\top a_t + \varepsilon_t$ to random Gaussian designs, and three steps of our analysis have no counterpart under direct observation. First, the GLR statistic is a design-dependent quadratic form with noncentrality $\sigma^{-2}\delta^\top (G_L^{-1} + G_R^{-1})^{-1}\delta$ built from two random Gram matrices, so the uniform detection event of Lemma 17 needs a lower bound on this form over every tested triple. Second, after a reset the estimate is clipped least squares, whose risk is the inverse-Wishart term $\sigma^2 r/(\mu(n - r - 1))$, and its expectation is taken over restart times selected by the same responses, as handled in Section F. Third, during a detection delay the least-squares fit mixes two coefficients through the design covariance, which is where κ enters the bias bound $9\kappa\Delta_k$ of Theorem 5. Consequently the upper and lower bounds carry r and κ, and the ratio in Corollary 6 is $O(\kappa^3 + LA_{\max}^2/\sigma^2)$ uniformly in the jump size. Hüselitz et al. (2026) study online jump detection in segmented regression with localization and detection-delay objectives and constant per-sample computation; our objective is cumulative regret.

Adaptive-regret and discounting methods. After projection onto the known U, the problem is r-dimensional online least squares with a piecewise-constant comparator. The interval guarantees of follow-the-leading-history (Hazan & Seshadhri, 2009) and of the strongly adaptive reduction of Daniely et al. (2015) need bounded losses or gradients, so Gaussian responses require truncation, and they do not isolate false alarms, detection delay and post-reset estimation. Jacobsen & Cutkosky (2024) analyze discounted online linear regression with learned discount factors; their protocol lets the prediction use the current feature, and their bounds depend on path variation rather than on K_T and Δ_0. Theorem 4 instead isolates false alarms, detection delay and post-reset estimation for Gaussian random designs and is paired with a lower bound on the same jump class in Corollary 6; on the well-spaced class Theorem 5 is one logarithm below the guarantee obtained by truncating the responses and summing FLH interval bounds over segments, at fixed κ, as derived in Section F.1; this is a comparison with that route, not a limit for adaptive-regret methods. Zhang et al. (2025) show that exponential weights with fixed-share updates attain dynamic regret $O(dT^{1/3}P_T^{2/3}\log T)$ for mixable losses, including bounded squared loss, where P_T is the path length of the comparator; their comparator class is measured by path length and the losses are bounded, whereas our class is measured by the number and size of jumps, the responses are Gaussian and unbounded, and the bounds are matched by a lower bound on the same jump class. Empirically, FLH with clipped ridge experts, fixed-discount forgetting least squares and learned-discount forecasters with fixed-share mixing are among the baselines of Tables 1 and 5 and Section A.2.

What the analysis adds. Five statements are specific to random-design regression with abrupt jumps and do not follow from the cited results by substitution. The first is the uniform detection event of Lemma 17 for random-design GLR statistics, whose noncentrality is a random quadratic form. The second is the post-reset risk of least squares after a response-selected restart, handled in Part 3 of the proof of Theorem 4 by a ghost estimator whose window start is a stopping time. The third is the

mixed-window bias bound $9\kappa\Delta_k$ during a delay, which makes the transient of Theorem 5 uniform in the jump size. The fourth is the van Trees bound on the separated well-spaced class, matched to the first three up to $O(\kappa^3 + LA_{\max}^2/\sigma^2)$ in Corollary 6. The fifth is the every-window lower bound for fixed-window least squares, which gives the polynomial separation of Corollary 7 at every jump rate T^γ, $\gamma < 1$.

Shared linear representations. Theorem 10 uses the spectral moment $y^2(xx^\top - I)$ of Tripuraneni et al. (2021), whose tasks become the coefficient segments of one stream; the moment is estimated on half of the rounds so that the other half stays independent of the estimated subspace, which lets the detection analysis of Lemma 17 run unchanged on the projected rounds, and Proposition 11 shows that its rd cost is necessary. Lifelong and sequential multi-task representation learning (Alquier et al., 2017; Cao et al., 2022) also carries a representation across tasks that arrive one after another, but the learner is told when a new task starts and the guarantees are stated per task or on average over tasks; in Theorem 10 the task boundaries are unknown change points inside one stream, their detection is paid for inside the cumulative dynamic regret, and Proposition 11 lower-bounds the representation cost under this protocol. Theorem 12 tracks the trace-centered moment $y^2(xx^\top - \|x\|^2 I/d)$ online, following Tripuraneni et al. (2021), with Fantope OGD against a moving comparator; the estimation error enters a dynamic-regret bound and block excitation plays the role of task diversity.

K Registered Confirmatory Tests

The studies of Sections A.2 and A.3 were added as amendments to the preregistration that already held the protocols of the theorem-matched experiments; each amendment was registered before the runs it governs. A4 and A5 were registered before any test run of E5 and E6; A6–A11 revise the baselines without changing the decision rules, and A12, A14 and A16 add descriptive analyses, each registered before it was computed; A18 is the study of Section A.4 and A19 that of Section A.5. The registered wording of F1 covers every non-oracle alternative, whereas the analysis excludes the untuned frozen-subspace control. The GLR arms of E5 use the capped scan of Section A.2, unlike the uncapped implementation of Table 1.

Amendment	Registered content	Outcome
A4, E5	D1: material advantage at a condition iff the relative reduction against the strongest competitor is $\geq 10\%$ and the paired interval lies below zero; D2: advantage at $T = 50$k and 100k in at least two path families, one of them at $\sigma = 0.5$ with Gaussian covariates, for each of GLR-1, GLR with tuned threshold and discard-all reset, and GLR-S	D2 met for GLR-S on P1 and P2; not met for GLR-1 or the tuned-threshold discard-all rule
A5, E6	F1: substantial advantage at a condition iff the relative reduction against the strongest non-oracle alternative, among the full-dimensional baselines and an exponentially weighted spectral-moment tracker feeding the same coefficient rule, is $\geq 15\%$ with the interval below zero; F2: F1 at $d \in \{50, 100\}$ and $T \in \{40, 160\}$k in two settings, one with weak excitation or smooth rotation	F2 not met
A6/A7	two-parameter baselines re-selected jointly over both grid parameters; grids widened	applied; E5/E6 rows of the affected baselines replaced
A8, E5	second widening of baseline grids at selected endpoints	applied; D2 outcome unchanged
A9, E5	learned-discount baseline with coefficient projection added	strongest competitor in 0 of 43 conditions; D2 unchanged
A10, E6	widening of the full-dimensional baseline grids; correction before any A10 test result was examined: a diverged tuning run counts as $+\infty$ in the selection	applied; F2 outcome unchanged
A11, E5	FLH weights renormalized after each update, as in standard pruned FLH, retuned and retested in all 43 conditions	applied; D2 outcome unchanged

Amendment	Registered content, continued	Outcome
A12	descriptive analyses and short reruns: discard-all rule at GLR-S's multiplier; FLH with tuned ridge on the streams of Table 1; regret decomposition and cadence runs of GLR-1; spectral-moment tracker of E6; standard errors of Table 3	reported in Sections A and A.1 to A.3; no decision rule changed
A14	jump-size sweep of GLR-1 over $\Delta \in \{0.8, \ldots, 0.05\}$ and $T \in \{5, 10, 20\}$k with two checks: A14-1, at each T the mean excess per change at $\Delta = 0.05$ is at most that at $\Delta = 0.8$; A14-2, at each T the largest mean excess per change over the grid is at most 16 times the smallest; regret–time–memory runs on the streams of Table 1	A14-1 met at every T; A14-2 not met at any T, since the smallest mean excess is negative; reported in Section A.1
A16	descriptive studies: round-type decomposition of GLR-S and the discard-all rule at equal multiplier in 28 conditions; learned-discount forecasters with fixed-share mixing in 8 conditions; forgetting least squares in information form on 150 E6 tuning cells with $d = 100$ and $\gamma \in \{0.5, 0.7\}$; regret, time and peak memory on streams with 0, 2 or 10 changes and ranks 4, 8, 16	reported in Sections A.1 to A.3; no decision rule changed
A18	GLR-U of Definition 9 on a fixed unknown subspace, $d \in \{20, 50, 100\}$, $\sigma \in \{0.1, 0.5\}$, $T \in \{20, 50, 100\}$k, against tuned full-dimensional sliding-window ridge and forgetting least squares; A18-1, at $d \in \{50, 100\}$, $T \in \{50, 100\}$k the paired interval of GLR-U minus the baseline with the lower tuning regret lies below zero; A18-2, for each (d, σ) the log–log slope of GLR-U over T is below 0.75	A18-1 not met in any of the 8 cells; A18-2 met in all 6 pairs; reported in Section A.4 and Table 15
A18b	the full-dimensional baseline grids of A18 extended towards shorter windows, faster forgetting and larger ridges before any A18 test run was examined	applied; the selection stays at the short-window or fast-forgetting end in 9 of the 36 cells of condition and baseline
A19	GLR-U with a change every $J \in \{150, 300\}$ rounds, $d \in \{50, 100, 200\}$, $T = 100$k, against the tuned full-dimensional baselines; A19-1 at $\sigma = 0.1$ and A19-2 at $\sigma = 0.5$: in each of the six (J, d) cells the paired interval of GLR-U minus the baseline with the lower tuning regret lies below zero	A19-1 met in 5 of 6 cells, all but $J = 300$, $d = 50$, which are reported in Section A.5; A19-2 not met in any cell; all cells in Table 16

Table 15: Registered study A18 on a fixed unknown subspace, all conditions: means over 20 paired test streams. The comparator is selected by the lower mean tuning regret between full-dimensional sliding-window ridge (W) and forgetting least squares (RLS). Δ is GLR-U minus comparator; wins count streams with lower GLR-U regret. All registered conditions are listed.

$T/10^3$	d	σ	comparator	GLR-U	comparator	Δ [95% CI]	wins
20	20	0.1	W	3 743	1 409	2 334 [2 141, 2 528]	0/20
20	20	0.5	RLS	6 906	2 598	4 307 [4 207, 4 407]	0/20
20	50	0.1	RLS	6 565	2 694	3 871 [3 657, 4 085]	0/20
20	50	0.5	W	8 269	4 554	3 715 [3 633, 3 796]	0/20
20	100	0.1	W	9 006	4 849	4 157 [4 000, 4 314]	0/20
20	100	0.5	W	9 325	6 047	3 278 [3 176, 3 381]	0/20
50	20	0.1	RLS	4 917	2 322	2 595 [2 331, 2 859]	0/20
50	20	0.5	RLS	12 172	4 773	7 400 [7 222, 7 577]	0/20
50	50	0.1	RLS	9 071	4 389	4 682 [4 345, 5 020]	0/20
50	50	0.5	W	15 078	8 506	6 572 [6 365, 6 779]	0/20
50	100	0.1	RLS	13 851	8 260	5 590 [5 098, 6 083]	0/20
50	100	0.5	W	18 063	12 687	5 376 [5 170, 5 582]	0/20
100	20	0.1	RLS	6 228	3 354	2 873 [2 717, 3 030]	0/20
100	20	0.5	RLS	17 847	7 796	10 051 [9 865, 10 237]	0/20
100	50	0.1	RLS	10 878	6 531	4 346 [3 977, 4 716]	0/20
100	50	0.5	RLS	21 803	14 006	7 797 [7 544, 8 049]	0/20
100	100	0.1	RLS	17 593	11 990	5 603 [4 987, 6 220]	0/20
100	100	0.5	W	27 210	21 525	5 685 [5 385, 5 984]	0/20

Table 16: Registered study A19 with frequent changes, all conditions at $T = 100$k: means over 20 paired test streams. The comparator is selected by the lower mean tuning regret between full-dimensional sliding-window ridge (W) and forgetting least squares (RLS). Δ is GLR-U minus comparator; wins count streams with lower GLR-U regret. All registered conditions are listed.

J	d	σ	comparator	GLR-U	comparator	Δ [95% CI]	wins
150	50	0.1	W	15 914	23 529	−7 615 [−7 810, −7 420]	20/20
150	50	0.5	W	41 774	28 975	12 799 [12 637, 12 960]	0/20
150	100	0.1	W	21 532	33 554	−12 023 [−12 357, −11 689]	20/20
150	100	0.5	W	44 155	37 915	6 240 [6 044, 6 436]	0/20
150	200	0.1	W	30 497	41 214	−10 716 [−11 024, −10 408]	20/20
150	200	0.5	W	46 635	43 023	3 612 [3 460, 3 765]	0/20
300	50	0.1	RLS	12 664	12 817	−153 [−596, 290]	12/20
300	50	0.5	W	32 194	21 689	10 504 [10 299, 10 710]	0/20
300	100	0.1	W	18 295	23 619	−5 324 [−5 657, −4 991]	20/20
300	100	0.5	W	35 812	29 732	6 079 [5 791, 6 367]	0/20
300	200	0.1	W	28 165	33 468	−5 303 [−5 960, −4 645]	20/20
300	200	0.5	W	40 731	37 042	3 688 [3 448, 3 929]	0/20