AMORTIZING STRUCTURAL INTERVENTIONS IN ONLINE OPINION DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Minimizing opinion polarization and disagreement on social platforms is constrained by incomplete information, as individual innate opinions are private and unobservable. Recent approaches formulate this task as a low rank matrix bandit, employing an initial stage to estimate the latent opinion subspace. We show that the Friedkin-Johnsen objective admits an exact closed form representation in known, opinion free features, eliminating the need for subspace estimation. This structure allows us to delineate the amortization boundary, where offline pretraining replaces online search at a measured price. Under social homophily, a pretrained policy achieves regret governed by the rank of the shared opinion subspace plus its imitation error, with no polynomial dependence on network size and only a logarithmic penalty for enlarging the candidate intervention pool. Guided by this amortizable regime, we introduce the Learned Dynamics Tracker (LDT), a compact filter trained offline to denoise parameter estimates and deployed without updates across drifting streams. In streaming deployments, LDT achieves the lowest dynamic regret across all six horizons with statistically significant and widening margins, reducing regret by 21 percent over the strongest baseline at the longest horizon. Across twenty transfer settings, seven of them on real network topology, LDT leads the prior state of the art in every one.

1 INTRODUCTION

Social platforms facilitate public discussion, but their interaction dynamics often foster ideological clustering and disagreement (Garimella et al., 2018; Cinelli et al., 2021). Because platforms cannot alter personal beliefs directly, their primary lever on division is structural, adjusting connection and exposure strengths. The Friedkin-Johnsen (FJ) model (Friedkin & Johnsen, 1990) formalizes this setting, with fixed innate opinions and network connectivity jointly governing expressed beliefs. Classical studies minimize FJ polarization and disagreement under the assumption that innate opinions are known in advance (Matakos et al., 2017; Musco et al., 2018), whereas practical deployments observe only a noisy aggregate evaluation without access to private beliefs. This information barrier turns structural moderation into an online learning problem. Recent work has progressively reduced the access to private beliefs that intervention methods assume, first restricting the learner to a finite query budget (Cinus et al., 2025) and then removing queries entirely, so that interventions improve from scalar aggregate evaluations alone (Cinus et al., 2026).

Despite this progress, two fundamental challenges remain. First, by treating the objective as an unconstrained quadratic form, existing formulations allocate a dedicated exploration stage to identify the latent opinion subspace on each new graph (Cinus et al., 2026). This recovery phase incurs regret that scales unfavourably with network size, leaving unresolved whether spending interactive rounds on subspace estimation is necessary or an artifact of ignoring network diffusion physics. Second, amortized optimization (Amos, 2023) and in context decision making (Laskin et al., 2023; Lin et al., 2024) pretrain across instances to eliminate online search, yet their guarantees are tied to the pretraining distribution. The structural conditions under which offline pretraining can replace online search remain uncharacterized, leaving the choice between amortization and search without a principled basis, and without a test that can be run before deployment.

We resolve both challenges through the algebraic structure of the FJ equilibrium, proving that an intervention's expected reward is the trace inner product $\langle \Theta^\star, X_a \rangle$ pairing the resolvent $X_a =$

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.

$(I + L_a)^{-1}$ of the intervened Laplacian with the unobservable rank one opinion matrix $\Theta^\star = ss^\top$. Because resolvents depend only on the graph and the candidate interventions, these features are computable prior to interaction, leaving $\Theta^\star$ as the sole latent parameter. The objective is therefore strictly linear in precomputed features, eliminating subspace estimation and reducing the task to a single stage linear bandit solved by our training free algorithm *Lift-TS*. These closed form features also characterize the amortization boundary. Under social homophily, where innate opinions concentrate in a shared rank r subspace, $\Theta^\star$ collapses into a transferable subspace of dimension at most $r(r+1)/2$, which an offline model carries across instances, supplying the principled basis for choosing amortization over online search.

This boundary informs our design. Drifting streams whose episodes share a low rank prior fall inside the amortizable regime, where topological features are already resolved in closed form and tracking nonstationary parameter dynamics is what remains. We therefore introduce the *Learned Dynamics Tracker* (LDT), a causal recurrent filter in the shared parameter space of candidate interventions, maintaining recursive ridge estimates across a window grid with uncertainty summaries that distinguish parameter drift from observation noise and emitting corrections as a zero initialized residual over an analytical reference window. Because the filter operates on spectral parameters decoupled from network size, a single frozen model deploys across networks and candidate pools of differing size, at a per round cost that is independent of network size once the structural features are precomputed, a stage that runs once per network and is reused thereafter.

- We prove that an intervention's expected reward is $\langle ss^\top, X_a \rangle$, pairing the private innate opinions s with the resolvent $X_a = (I + L_a)^{-1}$ of the intervened Laplacian. Because X_a is known in advance, the task reduces to an exact linear bandit on which our training free algorithm Lift-TS attains sublinear regret for arbitrary innate opinions without subspace estimation. The same closed form settles the convexity, nonsubmodularity and intrinsic dimension of the objective.
- We locate the amortization boundary of this objective. Under social homophily that concentrates innate opinions in a shared rank r subspace, a pretrained policy attains $\tilde{\mathcal{O}}(r\sqrt{T})$ regret plus imitation error with no polynomial dependence on network size, paying only $\sqrt{\log K}$ for a larger pool where every structure blind policy pays $\Omega(\min\{K, T\})$.
- Guided by the amortizable regime, we introduce the Learned Dynamics Tracker (LDT). Its design schedule makes the excitation condition a consequence of the algorithm, so on streams whose latent parameter moves within a variation budget $\mathcal{B}$ it attains $\tilde{\mathcal{O}}(\mathcal{B}^{1/3}T^{7/9}) + \mathcal{O}(T^{2/3})$ dynamic regret at fixed rank, and the same policy is certified a second time by its pretraining loss alone.
- Under nonlinear drift, LDT achieves the lowest dynamic regret at all six horizons with widening margins and a 21 percent reduction over the strongest baseline at the longest horizon. Furthermore, it transfers across twenty settings, seven of them on real network topology and the largest at 4039 nodes, leading the prior state of the art in every one.

2 Related work

Friedkin-Johnsen opinion optimization. Interventions in the Friedkin-Johnsen model moderate polarization and disagreement by adjusting innate opinions, edge weights or network ties (Matakos et al., 2017; Musco et al., 2018; Zhu et al., 2021; Sun et al., 2025), while adversarial perturbations maximize both (Gaitonde et al., 2020), and sublinear algorithms estimate equilibrium opinions from graph queries (Neumann et al., 2024; Feng et al., 2026), yet all assume access to explicit opinion values. Under private beliefs, Cinus et al. (2025) permit a finite query budget, whereas Cinus et al. (2026) eliminate queries entirely to learn under scalar bandit feedback alone. We give an exact closed form reward representation, computed from the graph and the candidate set alone, and delineate when offline pretraining replaces online search on fresh graphs.

Structured and nonstationary bandits. Low rank matrix bandits (Lu et al., 2021; Kang et al., 2022) typically allocate exploratory actions to identify an unknown parameter matrix (Jang et al., 2024) before running a reduced space linear bandit (Abbasi-Yadkori et al., 2011). Here the arm matrices themselves are the known quantity and the unknown is a single rank one factor, so Lift-TS runs linear Thompson sampling (Agrawal & Goyal, 2013) directly, attaining Bayes regret guarantees (Russo & Van Roy, 2016) without subspace estimation. Under nonstationary parameter drift (Besbes et al., 2014), sliding window (Garivier & Moulines, 2011; Cheung et al., 2019) and discounting

(Russac et al., 2019) strategies balance bias against variance; our framework exploits invariant resolvent structure to track parameter dynamics without periodic subspace recovery.

Amortized decision making. Amortized optimization exploits shared structure across instances to bypass per instance computation (Amos, 2023). Transformers pretrained on trajectories (Chen et al., 2021; Lee et al., 2023; Laskin et al., 2023; Zisman et al., 2024) learn decision rules offline, supervised pretraining yields in context posterior sampling and LinUCB (Bai et al., 2023; Lin et al., 2024), and such models explore under targeted prompting (Krishnamurthy et al., 2024), yet existing guarantees assume fixed distributions, leaving uncharacterized when amortized inference replaces online search. We characterize that boundary through an amortized regret bound governed by prior rank, with a matching floor for structure blind policies.

3 Problem formulation and structure

3.1 The Friedkin-Johnsen model

Let $G = (V, E)$ be a connected undirected graph of $|V|$ agents with nonnegative weights in adjacency matrix A, all ones vector $\mathbf{1}$, degree matrix $D = \mathrm{diag}(A\mathbf{1})$ and Laplacian $L = D - A \succeq 0$. Under the Friedkin-Johnsen model (Friedkin & Johnsen, 1990), each agent holds an immutable innate opinion $s_i \in \mathbb{R}$, and expressed opinions $z \in \mathbb{R}^{|V|}$ evolve across internal steps τ via $z^{(\tau+1)} = (I + D)^{-1}(Az^{(\tau)} + s)$ from $z^{(0)} = s$, converging geometrically to the unique equilibrium

$$z \;=\; Ms, \qquad M := (I + L)^{-1}, \tag{1}$$

where $M \succ 0$ is the doubly stochastic forest matrix of G (Chebotarev & Agaev, 2002). Mean centering $\mathbf{1}^\top s = 0$ (Musco et al., 2018) passes through $M\mathbf{1} = \mathbf{1}$ to $\mathbf{1}^\top z = 0$, so $\|z\|_2^2$ reads as dispersion about the mean, and scale normalization $\|s\|_2^2 = |V|$ makes networks of differing size comparable, so the index is on one scale across every network we evaluate.

Interventions are scored by the *polarization-disagreement index* (Musco et al., 2018), combining *polarization* $\|z\|_2^2$ (Matakos et al., 2017) with *disagreement* $z^\top Lz = \sum_{\{i,j\}\in E} A_{ij}(z_i - z_j)^2$, which collapses through $z = Ms$ into

$$\|z\|_2^2 + z^\top Lz \;=\; s^\top Ms \;=:\; \mathrm{P}(s, L). \tag{2}$$

Topology L is known to the platform and modifiable by it; innate opinions s are private.

3.2 Interventions and online deployment

Candidate interventions. An intervention, or *arm* a, strengthens connections across k node pairs in E_a, yielding the intervened Laplacian

$$L_a \;=\; L + \sum_{e\in E_a} \delta_e\, b_e b_e^\top, \tag{3}$$

where $\delta_e > 0$ is the weight increment on pair $e = \{i, j\}$ and $b_e = \mathbf{e}_i - \mathbf{e}_j$ its signed incidence vector, an update preserving Laplacian validity (Zhu et al., 2021). The learner selects from a pool $\mathcal{A}$ of $K := |\mathcal{A}|$ arms fixed prior to deployment. The structural feature $X_a := (I + L_a)^{-1}$ is computable from G and a alone without observing s.

Reward. Evaluating P on the intervened graph yields expected reward

$$\mu_a \;=\; s^\top X_a s \;=\; \langle \Theta^\star, X_a\rangle, \qquad \Theta^\star := ss^\top, \tag{4}$$

where $\langle \Theta, X\rangle := \mathrm{tr}(\Theta^\top X)$ is the Frobenius trace inner product. Retaining the explicit resolvent X_a, in place of an unconstrained low rank matrix bandit formulation (Cinus et al., 2026), bounds the reward through the physical model, with $0 \prec X_a \preceq I$, $\|X_a\|_F \le \sqrt{|V|}$ and $0 < \mu_a \le |V|/(1 + \lambda_2(L_a))$ for the algebraic connectivity $\lambda_2(L_a)$. We keep the term *reward* for μ_a although it is minimized, P scoring opinion division: regret is measured against $\min_a \mu_a$ and every algorithm below plays an $\arg\min$.

Protocol. At each round $t \in \{1, \dots, T\}$ the learner selects an arm $a_t \in \mathcal{A}$ using precomputed features $\{X_a\}$ while s, z and all μ_a remain latent, and the environment returns a noisy scalar evaluation $y_t = \mu_{a_t} + \zeta_t$ at equilibrium with independent noise $\zeta_t \sim \mathcal{N}(0, \sigma^2)$.

Drift and streams. Across continuous deployments, innate opinions evolve into round dependent s_t, shifting arm values $\mu_a(t) = s_t^\top X_a s_t$ while X_a stays fixed. Movement of $\Theta_t^\star := s_t s_t^\top$ is bounded by variation budget $\mathcal{B}$ (Besbes et al., 2014), and performance is dynamic regret against the per step optimum,

$$\sum_{t=1}^{T-1} \left\| \Theta_{t+1}^\star - \Theta_t^\star \right\|_F \;\le\; \mathcal{B}, \qquad \mathrm{Regret}_T \;=\; \sum_{t=1}^{T} \left(\mu_{a_t}(t) - \min_{a \in \mathcal{A}} \mu_a(t) \right), \tag{5}$$

which recovers static cumulative regret when $s_t \equiv s$. A *stream* consists of J sequential episodes of horizon T on a fixed network, each drifting independently under a shared innate opinion prior, so the prior is the component of the instance that persists across episodes.

3.3 Structural properties of the objective

The Sherman-Morrison and Woodbury identities (Horn & Johnson, 2012) give the exact objective shift $\Delta\mathrm{P} := \mathrm{P}(s, L_a) - \mathrm{P}(s, L)$ of an intervention.

Proposition 1 (Exact objective shift)**.** With incidence matrix $B = [b_{e_1}, \dots, b_{e_k}]$ collecting the pairs of E_a and weight increments $\mathrm{diag}(\delta) \succ 0$, so that $L_a = L + B\,\mathrm{diag}(\delta)B^\top$,

$$\Delta\mathrm{P} \;=\; -(B^\top z)^\top \left(\mathrm{diag}(\delta)^{-1} + B^\top M B \right)^{-1} (B^\top z) \;\le\; 0, \tag{6}$$

which for a single pair $e = \{i, j\}$ reduces to $\Delta\mathrm{P} = -\frac{\delta_e (z_i - z_j)^2}{1 + \delta_e\, b_e^\top M b_e}$. Given z, itself a global function of topology through M, any k pair intervention depends on the topology only through the $\mathcal{O}(k^2)$ scalars of $B^\top M B$, a count independent of $|V|$.

Symmetric vectorization then renders the reward model strictly linear. Let $\mathrm{vecsym}(\cdot)$ stack the upper triangular entries of a symmetric matrix with $\sqrt{2}$ scaling off the diagonal, preserving Euclidean inner products, so that $\psi_a = \mathrm{vecsym}(X_a)$ transforms equation 4 into

$$\mu_a \;=\; \psi_a^\top \theta^\star, \qquad \theta^\star := \mathrm{vecsym}(ss^\top), \tag{7}$$

an *exact linear bandit* with parameter variation matching the budget in §3.2. Because ψ_a is known prior to interaction, the reward relevant component of $\theta^\star$ is identifiable from played rounds alone, eliminating exploratory subspace estimation and leaving a single learning stage.

Proposition 3 finds continuous weight tuning monotone and convex while discrete pair selection is not submodular and falls outside the greedy guarantees of Nemhauser et al. (1978), and Lemma 1 puts attainable reward vectors on a manifold of Hausdorff dimension at most $|V|$, so the ambient $\mathcal{O}(|V|^2)$ of $\theta^\star$ overstates the degrees of freedom in beliefs and motivates the low rank prior of §4.

4 Bayes regret and the amortization boundary

This section locates the amortization boundary between what offline pretraining certifies under a shared low rank prior and what online interaction supplies, moving from no structural prior through social homophily to drift and closing on the tracker §5.2 deploys. The reward range is $\Delta_{\max} := \max_a \mu_a - \min_a \mu_a \le |V|$. Every certificate below names an algorithm this paper runs, measured in §6 beside the quantity it bounds. Three estimation laws stay apart throughout: the exact posterior benchmarked by Theorem 5, the Gaussian working prior carried by Lift-TS, and the learned filter of §5.2, which operates as a point estimator certified pathwise.

4.1 Bayes regret without a structural prior

Online regret is governed by the feature span dimension $d_\star := \dim \mathrm{span}\{\psi_a : a \in \mathcal{A}\} \le \min(K, |V|(|V|+1)/2)$. The information ratio framework of Russo & Van Roy (2016) yields an immediate certificate.

As established in Theorem 5, exact posterior sampling on $\{\psi_a\}$ attains expected regret at most $\max\{\sigma, \Delta_{\max}/2\}\sqrt{2\, d_\star T \log K}$ under whatever prior generates innate opinions and at any noise level, the benchmark the lifted structure buys any sampler reaching it. The sampler we deploy is certified at the realized opinions and not at their prior average.

Theorem 1 (Lift-TS regret under a misspecified working prior)**.** Fix any innate opinion vector s. Run Lift-TS with ridge design $V_t = \lambda I + \sum_{u<t} \psi_{a_u}\psi_{a_u}^\top$ at any $\lambda > 0$, sample $\tilde\theta_t = \hat\theta_t + c\, V_t^{-1/2}\eta_t$ with $\eta_t \sim \mathcal{N}(0, I_{d_\star})$ independent of the history at any inflation $c > 0$, and play $a_t \in \arg\min_a \psi_a^\top \tilde\theta_t$. Write $\kappa_\lambda := \max\{1, L_\psi^2/\lambda\}$, $\ell_T := \log\big(1 + TL_\psi^2/(d_\star\lambda)\big)$ and $\beta_T(\delta) := \sigma\sqrt{d_\star \ell_T + 2\log(1/\delta)} + \sqrt{\lambda}\, S$, the self normalized radius of the ridge estimate. Let $\mathcal{E}_\delta$ be the event that $\|\hat\theta_t - \theta^\star\|_{V_t} \le \beta_T(\delta)$ at every $t \le T$, of probability at least $1-\delta$, on which Lemma 3 lower bounds the conditional probability of the optimistic event at every round by $p_{\mathrm{opt}}(c,\delta) := \bar\Phi(\beta_T(\delta)/c) > 0$. Then with probability at least $1 - 2\delta$,

$$\mathrm{Regret}_T \;\le\; \frac{C}{p_{\mathrm{opt}}(c,\delta)}\Big(\beta_T(\delta) + c\sqrt{d_\star \log(d_\star/\delta)}\Big)\sqrt{\kappa_\lambda\, d_\star\, T\, \ell_T} \;+\; CSL_\psi d_\star \log T,$$

for an absolute constant C, with $L_\psi := \max_a \|\psi_a\|_2 \le \sqrt{|V|}$ and $S := \|\theta^\star\|_2 \le |V|$, for any design and any law of s. Here $\lambda \ge L_\psi^2$ sets $\kappa_\lambda = 1$ and $c \ge \beta_T(\delta)$ gives $p_{\mathrm{opt}}(c,\delta) \ge \bar\Phi(1) > 1/7$ at the standard normal tail $\bar\Phi$, recovering $\tilde{\mathcal{O}}(\sigma d_\star^{3/2}\sqrt{T})$; every other pair is the same statement at its own κ_λ and $p_{\mathrm{opt}}(c,\delta)$.

Theorem 1 holds $\theta^\star$ fixed and reads the sampling step as a randomization device inside a self normalized ellipsoid, so neither the rank one cone nor the law of s appears, at a leading price of $\tilde{\mathcal{O}}(d_\star)$. The regularizer and inflation enter only through κ_λ, which is explicit, and $p_{\mathrm{opt}}(c,\delta)$, which Lemma 3 evaluates as a measurable tail at any inflation and Appendix E reports at the deployed pair. Because distinct interventions generically yield independent features, $d_\star$ reaches $\Theta(|V|^2)$ on an unrestricted pool, so size independent regret needs the homophily prior.

4.2 Amortized regret under homophily

Social homophily aligns the opinions of connected individuals (Garimella et al., 2018; Cinelli et al., 2021), concentrating innate opinions along dominant spectral directions.

Assumption 1 (Homophily, a low rank innate prior)**.** The innate opinion prior is supported on an r dimensional subspace $\mathcal{S}_r \subseteq \mathbf{1}^\perp$, $r \ll |V|$, canonically the span of the leading nontrivial Laplacian eigenvectors, and the pool $\mathcal{A}$ is finite.

It constrains $\theta^\star$ while leaving ψ_a exact. Let $\varepsilon_{\mathrm{pre}}$ be the per step total variation error between a pretrained policy and exact posterior sampling (Ross & Bagnell, 2010; Lee et al., 2023; Lin et al., 2024), the one quantity through which pretraining enters the bound.

Theorem 2 (Amortized Bayes regret)**.** Under Assumption 1, the problem reduces exactly to a linear bandit of effective dimension $d_{\mathrm{eff}} \le r(r+1)/2$, and an in context policy pretrained for decision making satisfies

$$\mathbb{E}[\mathrm{Regret}_T] \;=\; \mathcal{O}\big(\sqrt{d_{\mathrm{eff}} T \log K}\big) + R_{\mathrm{pre}} \;=\; \tilde{\mathcal{O}}(r\sqrt{T}) + R_{\mathrm{pre}},$$

where $R_{\mathrm{pre}} = \sum_t \mathbb{E}\big[\sum_a |\pi_t(a) - p_t(a)|\Delta_t(a)\big]$ charges the deviation from exact posterior sampling p at the gaps $\Delta_t(a) = \mu_a - \min_b \mu_b$ across which it moves mass, and is at most $\mathcal{O}(T\varepsilon_{\mathrm{pre}})$ on the pretraining distribution and $\mathcal{O}(T^2\varepsilon_{\mathrm{pre}})$ under arbitrary distribution shift.

The excess penalty R_{pre} is a measured quantity: evaluating the exact posterior by quadrature places the deployed policy at deviation $\varepsilon_{\mathrm{pre}} = 0.23$ in Appendix F. At fixed rank and horizon the candidate pool then enters solely through $\sqrt{\log K}$. Conversely, Corollary 2 shows that any policy reading the index reward history without the features pays $\Omega(\Delta_0 \min\{K, T\})$ on an instance realizable inside the model, whereas least squares on spectral features localizes the optimal arm from $\mathcal{O}(r^2 \log K/\Delta_{\min}^2)$ observations instead of the $\Omega(K)$ without cross arm transfer.

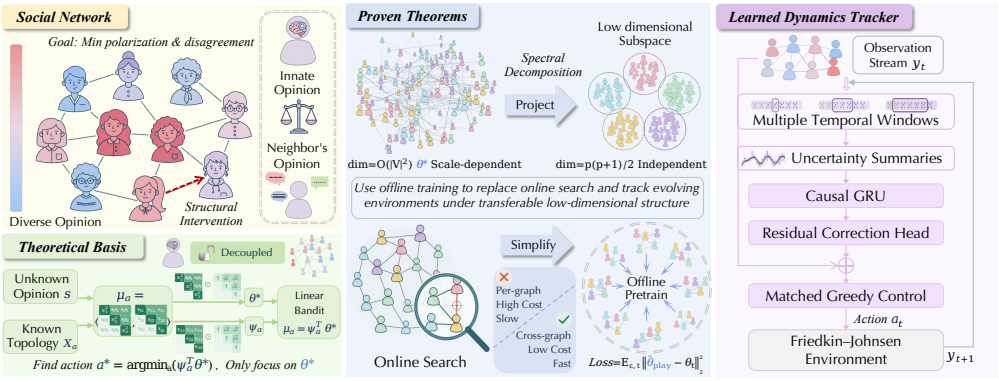

Figure 1: Overview. The closed form representation decouples observable topology from private beliefs, and the Learned Dynamics Tracker filters parameters in a transferable low dimensional subspace.

4.3 Dynamic regret and the deployed tracker

Parameter trajectories $\theta_t^\star = \mathrm{vecsym}(\Theta_t^\star)$ drift under variation budget $\mathcal{B}$ (Besbes et al., 2014); because topology is static, ψ_a stays invariant and the drift is shared across candidates.

Theorem 3 (Nonstationary power law separation)**.** Suppose the innate opinion vector traverses a fixed path at a uniform rate across the horizon, so that the induced parameter trajectory meets the budget in equation 5 with $\mathcal{B} = \mathcal{O}(1)$ (Besbes et al., 2014); call this the budgeted path condition. Suppose for *(i)* that in addition no single intervention is optimal at almost every point of that path. Then *(i)* every commit once policy, playing one fixed arm after sublinear exploration, incurs $\bar{g}(\mathcal{B})\,T - o(T)$ dynamic regret at the time averaged excess $\bar{g}(\mathcal{B}) := \min_a \int_0^1 (\mu_a - \mu^\star) > 0$, while *(ii)* a structure aware sliding window tracker on spectral features under the ρ schedule of Lemma 4 incurs dynamic regret $\tilde{\mathcal{O}}(d_{\mathrm{eff}}^{1/3}\mathcal{B}^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3}$, the second term the forced draws that schedule places.

The bounds differ by a factor of $T^{2/9}$ and Lemma 5 realizes both hypotheses inside the Friedkin-Johnsen model at rank two. Part (ii) carries its own design: draws arriving at density $\rho_t = ct^{-1/3}$ are independent of the history, conditioning the window covariance along whatever trajectory the policy generates. A window of length W ending at t holds $\Theta(cWt^{-1/3})$ of them and the variance divides by that count, which places the balancing memory at $\Theta(T^{7/9})$.

Proposition 2 (Greedy control under parameter error)**.** Let $a_t \in \arg\min_a \varphi_a^\top \hat{\theta}_t$ over the spectral features equation 8, let $a_t^\star$ be optimal at round t, and let $e_t = \hat{\theta}_t - \theta_t$. Then

$$\mathrm{Regret}_T \;\le\; \sum_t (\varphi_{a_t^\star} - \varphi_{a_t})^\top e_t \;=\; T\,\bar{\mathcal{G}}, \qquad \text{and} \qquad \bar{\mathcal{G}} \;\le\; D\sqrt{\bar{\mathcal{L}}},$$

where $\bar{\mathcal{G}}$ averages the estimation error along the direction each round's own decision exposes, $D = \max_{a,b} \|\varphi_a - \varphi_b\|_2$ is the feature diameter and $\bar{\mathcal{L}} = \frac{1}{T}\sum_t \|\hat{\theta}_t - \theta_t\|_2^2$ is the objective equation 11 read along the trajectory the policy realizes. Both inequalities hold path by path, so no excitation, drift or total variation hypothesis is used and the law that produced $\hat{\theta}_t$ is left free.

The certificate is the aligned form on the left, and relaxing its alignment to a norm leaves the pre-training objective as the only data dependent quantity, read on the distribution deployment realizes. On the deployed streams its leading form scales as $T^{0.77}$, $[.73, .81]$ in Table 23, below the rate Theorem 3 certifies at the measured variation budget.

Theorem 4 (Certified dynamic regret for the deployed tracker)**.** Fix $\varepsilon_0 > 0$, run the tracker of §5.2 under the ρ schedule of Lemma 4, and project the played estimate onto the ball of radius $\varepsilon_t = \varepsilon_0 t^{-1/3}$ about the window estimate $\hat{\theta}_t^W$ at the anchor $W = W_{\mathrm{med}}$ the architecture fixes, which is the configuration deployed. Then, under the budgeted path condition of Theorem 3, with probability at least $1 - \delta$,

$$\mathrm{Regret}_T \;\le\; \mathcal{R}_W(T) \;+\; \tfrac{3}{2}\,D\,\varepsilon_0\,T^{2/3},$$

where $\mathcal{R}_W(T)$ is the fixed length bound Corollary 6 carries at that anchor, the forced rounds included. Writing κ for its ratio to the memory that balances the two error sources, the total is $\Lambda(\kappa)\,\tilde{\mathcal{O}}\big(D\,d_{\text{eff}}^{1/3}\mathcal{B}^{1/3}T^{7/9}\big) + \gamma_0 T^{2/3}$ at $\gamma_0 := \frac{3}{2}\big(c\,\Delta_{\max} + D\,\varepsilon_0\big)$ and the inflation $\Lambda(\kappa) = \frac{1}{3}(2\kappa^{-1/2} + \kappa)$ of Corollary 6, for every setting of the filter parameters ω and at every budget including $\mathcal{B} = 0$.

The projection is computable online because $\hat{\theta}_t^W$ is already a coordinate of the filter's state, carrying the rate across the learned component: the schedule of Lemma 4 conditions the design from history independent draws, so the bound holds uniformly over what the filter outputs and over its pretraining distribution, while the objective equation 11 enters Proposition 2's certificate at the square root along the realized rollout. This is the configuration §6 reports throughout, Tables 8 and 9 measuring the design condition and the draws and projections it spends. Appendix B carries the boundary cases: Theorem 6 outside the spectral window, Theorem 7 on the variation budget pooling replaces, Proposition 7 on feature degeneracy, Corollary 5 on the streaming form.

5 Method

Guided by the amortization boundary of §4, we realize structural intervention through two complementary learners matched to the stationary and nonstationary regimes. On static networks where beliefs are invariant, the training free sampler Lift-TS operates directly in the lifted feature space under the sublinear certificate of Theorem 1. Under streaming parameter drift, the Learned Dynamics Tracker (LDT) amortizes invariant spectral geometry to track dynamic beliefs within the low dimensional regime identified in Corollary 5. Figure 1 outlines the design and each learner's role.

5.1 Structure aware sampling

Because expected rewards $\mu_a = \psi_a^\top \theta^\star$ are strictly linear in precomputed resolvent features $\psi_a = \mathrm{vecsym}(X_a)$, structural intervention is an exact linear bandit, on which Lift-TS runs linear Thompson sampling (Agrawal & Goyal, 2013) over a Gaussian working posterior updated conjugately on scalar feedback, playing $a_t = \arg\min_a \psi_a^\top \tilde{\theta}_t$. Because rewards depend on $\theta^\star$ solely through its projection onto $\mathrm{span}\{\psi_a\}$, inference operates strictly in that $d_\star$ dimensional subspace at an update cost of $\mathcal{O}(d_\star^2)$ and sampling cost of $\mathcal{O}(d_\star^3)$ per round, independent of $|V|$. Theorem 1 certifies sublinear regret under arbitrary private beliefs for regularizer and inflation parameters chosen once on held out seeds.

5.2 Learned Dynamics Tracker

To eliminate dependence on network size under streaming drift, LDT compresses candidate interventions into opinion free spectral features,

$$\varphi_a = \mathrm{vecsym}\big(U_p^\top X_a U_p\big) \in \mathbb{R}^{d_p}, \qquad d_p := p(p+1)/2, \tag{8}$$

where $U_p \in \mathbb{R}^{|V|\times p}$ collects the $p \geq r$ leading nontrivial Laplacian eigenvectors. Rewards are linear in φ_a exactly when $\mathcal{S}_r \subseteq \mathrm{span}(U_p)$, which Assumption 1 supplies; otherwise Proposition 5 identifies the residual and Theorem 6 bounds it, with Corollary 4 expressing the bound through the Laplacian eigengap at p, which is also how we set p, at the first sizeable gap above r. Because φ_a and θ_t transform by the same congruence when $U_p \mapsto U_p Q$, every reward is invariant to the sign, ordering and rotation ambiguity of the eigendecomposition; only the filter's input coordinates move. Along a drifting stream, the parameter

$$\theta_t = \mathrm{vecsym}\big(U_p^\top s_t s_t^\top U_p\big) \in \mathbb{R}^{d_p} \tag{9}$$

moves along a shared nonlinear trajectory, varying all candidate rewards while φ_a stays fixed. LDT therefore filters θ_t in this shared parameter space instead of tracking K arm values separately. Because d_p depends only on the spectral cutoff p, one frozen filter deploys across networks and pools of any size without retraining.

Multiscale estimation with uncertainty summaries. Theorem 3 balances variance against drift bias at memory $W = \Theta\big((d_{\text{eff}}^{1/2}T^{7/6}/\mathcal{B})^{2/3}\big)$. Because the variation budget $\mathcal{B}$ is unknown online,

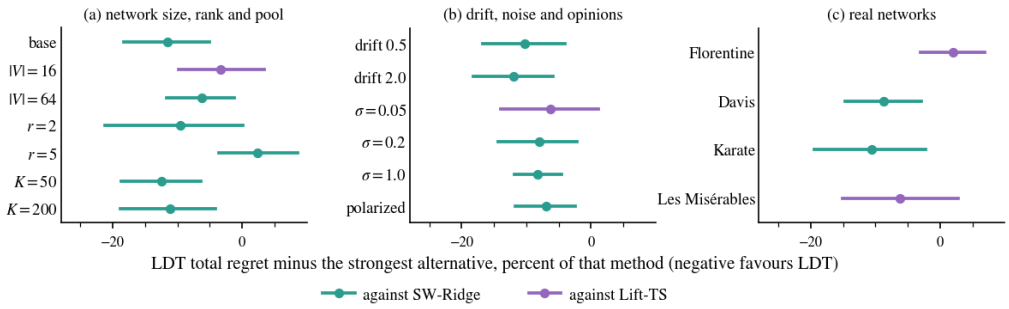

Figure 2: Transfer across seventeen of the twenty settings, with 95% paired bootstrap intervals on LDT's regret against the strongest alternative.

a single fixed window risks statistical inflation or lag, which LDT answers by computing recursive ridge estimates $\hat{\theta}_t^W$ (Abbasi-Yadkori et al., 2011) across a geometric window grid of ratio at most 3, where Corollary 6 holds a member within 1.09 of the bound at that memory over the range it spans. Each window supplies an uncertainty summary $[m_W^{-1/2},\ \hat{\sigma}_W,\ \|\hat{\theta}_t^W - \hat{\theta}_t^{W_{\mathrm{med}}}\|]$ tracking sample count, residual scale and reference divergence, so the grid separates parameter motion from observation noise as evidence while the played estimate stays at the fixed anchor W_{med}. As Appendix D details, these multiwindow sufficient statistics update recursively via rank one operations in $\mathcal{O}(d_p^3)$ time per window, independent of network size $|V|$ and horizon T.

Learned dynamics prediction. A gated recurrent unit (GRU) g_ω (Cho et al., 2014) processes this multiscale diagnostic stream, predicting active parameters as an inductive residual correction to the analytical reference window,

$$\hat{\theta}_{\mathrm{play}}(\omega) \;=\; \hat{\theta}_t^{W_{\mathrm{med}}} + g_\omega(x_{1:t}), \tag{10}$$

where x_t concatenates window estimates with their uncertainty summaries. A residual head set to zero at initialization makes LDT coincide with the reactive tracker certified by Theorem 3, focusing offline optimization on variance reduction, and causal recurrence gives training and deployment the identical state mapping at constant per round latency.

Matched control under the certified schedule. LDT matches the control protocol of reactive baselines, an identical random warmup then greedy exploitation $a_t = \arg\min_a \varphi_a^\top \hat{\theta}_{\mathrm{play}}$, attributing its margin to parameter denoising within the amortized regime of Corollary 5. Around it the policy runs the two mechanisms certified in §4.3: forced exploration samples the G optimal design when $\sum_{s\le t} cs^{-1/3}$ passes an integer, and the played estimate is projected onto the ball of radius $\varepsilon_t = \varepsilon_0 t^{-1/3}$ at $\hat{\theta}_t^{W_{\mathrm{med}}}$. Both depend solely on the round index, so deployment takes no horizon as input. The constants (c, ε_0), the checkpoint and the pretraining collection rule are settled on validation streams disjoint from every reported experiment, and Proposition 2 and Theorem 4 certify it.

The filter is trained via parameter space regression over episodic trajectories,

$$\mathcal{L}_{\mathrm{LDT}}(\omega) \;=\; \mathbb{E}_{\mathrm{episode},\,t}\big\|\,\hat{\theta}_{\mathrm{play}}(\omega) - \theta_t\,\big\|_2^2. \tag{11}$$

Supervising in parameter space decouples pretraining from arm identities, giving a compact frozen tracker that generalizes across pools of any size.

6 Experiments

We evaluate on the online Friedkin-Johnsen benchmark of Cinus et al. (2026) under streaming drift, reporting dynamic regret equation 5 against the per step optimum and the round averaged polarization-disagreement index equation 2, lower better on both.

Experimental setup. Each stream comprises $J = 20$ episodes on an Erdős & Rényi (1959) graph ($|V| = 32$) with $K = 100$ candidate interventions, innate opinions drifting along a nonlinear oscillator at rate 1.0 within a shared rank $r = 3$ subspace, and spectral learners at $p = 6$ ($d_p = 21$). We benchmark against OPD-Min-ESTR of Cinus et al. (2026) from official code, our sampler Lift-TS, two trackers on our structural features at oracle windows, SW-Ridge (Cheung et al., 2019) and

Table 1: Streaming performance at horizon T, reporting dynamic regret cumulated across the J episodes and the mean polarization-disagreement index. Lower is better, best per column in bold.

	$T = 100$		$T = 200$		$T = 300$		$T = 400$		$T = 600$		$T = 800$	
Method	Regret	Index	Regret	Index	Regret	Index	Regret	Index	Regret	Index	Regret	Index
LowPopArt	136.2	0.268	261.0	0.264	382.2	0.263	501.3	0.262	732.9	0.260	962.5	0.259
OPD-Min-ESTR	120.9	0.260	231.7	0.257	341.6	0.256	449.8	0.255	664.3	0.254	879.0	0.254
BOB	121.6	0.260	223.3	0.255	333.8	0.255	439.4	0.254	648.1	0.253	849.0	0.252
D-LinUCB	94.3	0.247	165.5	0.240	231.5	0.238	292.0	0.235	406.2	0.233	511.8	0.231
Vel-Extrap	94.5	0.247	140.3	0.234	182.4	0.229	223.3	0.227	301.6	0.224	377.3	0.222
SW-Ridge	70.4	0.235	113.6	0.228	152.3	0.224	191.7	0.223	268.8	0.221	343.2	0.220
Lift-TS (ours)	81.6	0.240	135.4	0.233	177.2	0.229	213.2	0.226	274.9	0.222	327.9	0.219
LDT (ours)	**66.9**	**0.233**	**100.5**	**0.224**	**131.1**	**0.221**	**160.5**	**0.219**	**215.8**	**0.217**	**270.9**	**0.216**

Vel-Extrap (Jia et al., 2023), and D-LinUCB at oracle discount (Russac et al., 2019), BOB (Cheung et al., 2019) and LowPopArt (Jang et al., 2024). All comparison tuning constants are oracle best and OPD-Min-ESTR receives the exact $t = 0$ subspace, making every margin conservative. Evaluations pair instances, seeds and metric across 20 streams with 95% bootstrap intervals.

Main results. Table 1 places LDT lowest in regret and index at all six horizons, and Table 16 finds all 48 paired differences significant under Holm correction. The margin widens with T, cutting regret by 21 percent over the strongest baseline SW-Ridge at $T = 800$, 270.9 against 343.2, the index moving with it, so the gain lands on the moderation objective. Lift-TS reaches 327.9 there without pretraining, so the closed form and the filter each supply part of it. Figure 2 carries the ordering across seventeen settings, thirteen one factor variations and four real networks supplying measured topology under a controlled opinion process, LDT leading the prior state of the art in every one and the paired interval favouring it in eleven. Appendix E extends this to twenty settings, seven real topologies, six further baselines and the five shifts Theorem 6 prices, LDT ahead throughout.

Complexity and rate scaling. With fixed per step drift, episode variation scales as $\mathcal{B} \propto T^{0.39}$, so Theorem 3 and Corollary 5 certify a $T^{0.91}$ tracking rate for the deployed schedule, and Proposition 2 reaches the same reference from its pretraining loss alone. Fits $\mathrm{Regret}_T \propto T^{\alpha}$ place LDT at $\alpha \approx 0.67$, within that certificate, against 0.95 for OPD-Min-ESTR, 0.76 for SW-Ridge and 0.67 at a larger constant for Lift-TS. Execution costs diverge with size: OPD-Min-ESTR's per round runtime scales as $|V|^{2.15}$ from $|V| = 128$ to 1024 while our structural feature learners stay flat within 1.0% at fixed pool size, letting LDT decide 143× faster, a decision stage ratio Appendix D completes end to end, the one time 669 s pretraining repaid within three streams.

Ablation and tracking mechanism. Adding the three components one at a time lowers regret at both horizons: from SW-Ridge at its oracle window, 191.7 at $T = 400$, the learned dynamics reach 185.1, the multiscale grid 169.4 and the uncertainty summaries 160.5, and the filter cuts parameter estimation error by 10.2%, what equation 11 supervises. Capacity alone does not reproduce this: an arm space in context transformer 31× larger trails by 16 and 42 percent at $T = 200$ and $T = 400$, both intervals excluding zero, locating the gain in the filter's space.

7 Conclusion

We establish an exact closed form for Friedkin-Johnsen opinion moderation, where intervention rewards are linear in resolvent features known before interaction. It removes exploratory subspace estimation and delineates the amortization boundary, pretraining replacing online search at a measured price. The Learned Dynamics Tracker, a causal filter over multiscale ridge estimates, attains the lowest dynamic regret at every horizon, cutting it by 21 percent against the strongest baseline at $T = 800$ and generalizing over twenty settings, seven on real topology under a controlled opinion process. Aligning architecture with the objective's geometry gives a blueprint for structural intervention wherever equilibrium rewards admit closed forms in observable topology.

Reproducibility Statement

Problem formulations, proposed algorithms, and baselines are fully specified in §3–5. Complete theorem statements, formal assumptions, and mathematical proofs are provided in Appendix B. Algorithmic pseudocode, per round complexity analyses, and recursive updates appear in Appendix D. Appendix E details experimental protocols, hyperparameter grids, and model parameter counts. Anonymized source code with reproduction instructions is included in the supplementary material.

AI Use Statement

Generative AI tools were used solely to polish the writing of this paper. The authors take full responsibility for all content, code, and claims.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NeurIPS*, 2011. 2, 8, 19, 46

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. In *AISTATS*, pp. 176–184, 2017. 19, 20

Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD)*, pp. 36–43, 2005. 42

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pp. 127–135, 2013. 2, 7, 20

Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023. 1, 3

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *NeurIPS*, 2023. 3

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *NeurIPS*, 2014. 2, 4, 6, 30, 31, 32, 33

Pavel Chebotarev and Rafig Agaev. Forest matrices around the Laplacian matrix. *Linear Algebra and its Applications*, 356(1–3):253–274, 2002. 3, 13

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *NeurIPS*, 2021. 3

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *AISTATS*, 2019. 2, 8, 9, 46, 47

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pp. 1724–1734, 2014. 8

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *PNAS*, 118(9):e2023301118, 2021. 1, 5

Federico Cinus, Atsushi Miyauchi, Yuko Kuroki, and Francesco Bonchi. Minimizing polarization and disagreement in the Friedkin–Johnsen model with unknown innate opinions. In *IJCAI*, pp. 7365–7373, 2025. 1, 2

Federico Cinus, Yuko Kuroki, Atsushi Miyauchi, and Francesco Bonchi. Online minimization of polarization and disagreement via low-rank matrix bandits. In *ICLR*, 2026. 1, 2, 3, 8, 42, 52

Allison Davis, Burleigh B. Gardner, and Mary R. Gardner. *Deep South: A Social Anthropological Study of Caste and Class*. University of Chicago Press, 1941. 42

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. 25

Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959. 8, 42

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, revised edition, 2015. 17

Weiming Feng, Zelin Li, and Pan Peng. Sublinear-time algorithms for diagonally dominant systems and applications to the Friedkin–Johnsen model. In *COCOON*, pp. 149–162, 2026. 2

Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4):193–206, 1990. 1, 3

Jason Gaitonde, Jon Kleinberg, and Éva Tardos. Adversarial perturbations of opinion dynamics in networks. In *EC*, 2020. 2

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *WWW*, 2018. 1, 5

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *ALT*, pp. 174–188, 2011. 2, 46

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *SciPy*, pp. 11–15, 2008. 42

Eric C. Hall and Rebecca M. Willett. Dynamical models and tracking regret in online convex programming. In *ICML*, pp. 579–587, 2013. 46

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002. 35

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 56

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. 46

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012. 4, 14, 35

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Efficient low-rank matrix estimation, experimental design, and arm-set-dependent low-rank bandits. In *ICML*, pp. 21329–21372, 2024. 2, 9, 48

Su Jia, Qian Xie, Nathan Kallus, and Peter I. Frazier. Smooth non-stationary bandits. In *ICML*, pp. 14930–14944, 2023. 9, 47

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient frameworks for generalized low-rank matrix bandit problems. In *NeurIPS*, 2022. 2

Donald E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. ACM Press, 1993. 42

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? In *NeurIPS*, 2024. 3

Jérôme Kunegis. KONECT: The Koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pp. 1343–1350, 2013. 42

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *ICLR*, 2023. 1, 3

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. 55

Tor Lattimore, Csaba Szepesvári, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *ICML*, pp. 5662–5670, 2020. 24

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *NeurIPS*, 2023. 3, 5, 55, 56

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *ICLR*, 2024. 1, 3, 5, 56

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 38, 45

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *AISTATS*, 2021. 2

Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505, 2017. 1, 2, 3

Julian McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 539–547, 2012. 42

Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. Minimizing polarization and disagreement in social networks. In *WWW*, pp. 369–378, 2018. 1, 2, 3, 42

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. 4

Stefan Neumann, Yinhao Dong, and Pan Peng. Sublinear-time opinion estimation in the Friedkin–Johnsen model. In *WWW*, 2024. 2

John F. Padgett and Christopher K. Ansell. Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology*, 98(6):1259–1319, 1993. 42

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *AISTATS*, pp. 661–668, 2010. 5, 21

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *NeurIPS*, 2019. 3, 9, 47

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *JMLR*, 17(1):2442–2471, 2016. 2, 4, 14, 17, 18

Mohammad Shirzadi, Emilio Cruciani, and Ahad N. Zehmakan. Opinion dynamics: A comprehensive overview. *arXiv preprint arXiv:2511.00401*, 2025. 35

Haoxin Sun, Yubo Sun, Xiaotian Zhou, and Zhongzhi Zhang. Fast computation and optimization for opinion-based quantities of Friedkin–Johnsen model. In *NeurIPS*, 2025. 2

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. 27

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 55

Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977. 42

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *AISTATS*, 2020. 46

Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. In *NeurIPS*, 2021. 2, 3, 42

Ilya Zisman, Vladislav Kurenkov, Alexander Nikulin, Viacheslav Sinii, and Sergey Kolesnikov. Emergence of in-context reinforcement learning from noise distillation. In *ICML*, 2024. 3

APPENDIX CONTENTS

Appendix A **Notation and background facts** (page 13). Notation, the algebraic and metric properties of the forest resolvent, the contraction argument behind the equilibrium, and the scale calibration used throughout.

Appendix B **Full proofs and technical derivations** (page 14). Complete arguments for every result of §3 and §4: nonsubmodularity, intrinsic dimensionality, the regret certificates, zero shot transfer, and the necessity and sufficiency of cross episode amortization.

Appendix C **Limitations** (page 35). Operational scope within the structural intervention literature, and natural extensions to other opinion dynamics and co-evolving topologies.

Appendix D **Algorithms and computational complexity** (page 35). Specifications for Lift-TS, streaming LDT deployment and offline pretraining, with per round cost accounting independent of network size $|V|$ and horizon T.

Appendix E **Experimental details** (page 41). Benchmark environments, homophily prior instantiations, architecture, baseline reproduction and fairness protocols, mechanism verification in parameter space, and transfer across seven real networks.

Appendix F **The learned prior policy** (page 55). The in context transformer that isolates offline amortization without online exploration, measured against the $\sqrt{\log K}$ pool dependence of Corollary 2(i) and the transfer bound of Theorem 6.

Appendix G **Validation streams** (page 58). The disjoint protocol the released checkpoint and the deployment constants were settled on.

A NOTATION AND BACKGROUND FACTS

Table 2 summarizes the mathematical symbols used throughout the paper. Throughout, M denotes the forest resolvent matrix, J denotes the number of streaming episodes, and lowercase Greek vectors $(\psi_a, \varphi_a, \theta^\star, \theta_t)$ denote the symmetric vectorizations of their corresponding matrix counterparts. The equilibrium $z = Ms$ introduced in §3.1 represents the unique stationary state of the Friedkin-Johnsen dynamics, where agents update opinions via $z^{(\tau+1)} = (I + D)^{-1}(Az^{(\tau)} + s)$. This update is a strict contraction in the maximum norm because each row i of $(I + D)^{-1}A$ is nonnegative with sum $d_i/(1 + d_i) \le \max_j d_j/(1 + \max_j d_j) < 1$, ensuring that the spectral radius of $(I + D)^{-1}A$ is strictly less than unity. Consequently, the dynamics converge geometrically from any initial profile $z^{(0)}$ to the unique fixed point solving $(I + L)z = s$.

Properties of the forest matrix. The resolvent $M = (I+L)^{-1}$ satisfies three structural properties that establish our operational bounds.

(i) **Spectral bounds.** M is symmetric positive definite with eigenvalues $1/(1 + \lambda_i(L)) \in (0, 1]$, ensuring that intervention features satisfy $0 \prec X_a \preceq I$ and $\|X_a\|_F \le \sqrt{|V|}$ in §3.2.

(ii) **Stochasticity and centering.** Because $I + L$ is a nonsingular M-matrix, its inverse M is entrywise nonnegative (Chebotarev & Agaev, 2002) and satisfies $M\mathbf{1} = \mathbf{1}$. The equilibrium opinions $z = Ms$ are therefore convex combinations of innate beliefs. Mean centering $\mathbf{1}^\top s = 0$ directly implies $\mathbf{1}^\top z = \mathbf{1}^\top Ms = (M\mathbf{1})^\top s = \mathbf{1}^\top s = 0$, identifying the polarization index as scaled sample variance,

$$\|z\|_2^2 \;=\; \sum_{i \in V} (z_i - \bar{z})^2 \;=\; |V| \cdot \mathrm{Var}(z).$$

(iii) **Intervention scalar bounds.** For candidate pair $e = \{i, j\}$ with incidence vector $b_e = \mathbf{e}_i - \mathbf{e}_j$, expanding the quadratic form yields

$$b_e^\top M b_e \;=\; (\mathbf{e}_i - \mathbf{e}_j)^\top M (\mathbf{e}_i - \mathbf{e}_j) \;=\; M_{ii} + M_{jj} - 2M_{ij}.$$

Because $M_{ii}, M_{jj} \le 1$ and $M_{ij} \ge 0$, this quantity obeys $0 \le b_e^\top M b_e \le 2$, remaining finite even if the underlying graph is disconnected. For multi-pair interventions, the $k \times k$ matrix $B^\top M B$ captures all topological coupling among the k modified edges.

Reward scaling and canonical normalization. Because innate opinions satisfy $s \in \mathbf{1}^\perp$ and $\mathbf{1}$ is an eigenvector of $I + L_a$, the invariant subspace $\mathbf{1}^\perp$ restricts the Rayleigh quotient to the nontrivial spectrum:

$$0 \;<\; \mu_a \;=\; s^\top (I + L_a)^{-1} s \;\le\; \frac{\|s\|_2^2}{1 + \lambda_2(L_a)} \;=\; \frac{|V|}{1 + \lambda_2(L_a)},$$

where $\lambda_2(L_a)$ is the algebraic connectivity of the intervened network under standard scaling $\|s\|_2^2 = |V|$. Because edge additions cannot decrease algebraic connectivity, this bound tightens as interventions strengthen connectivity. The nominal reward range spanning the candidate intervention pool is $\Delta_{\max} := \max_{a \in \mathcal{A}} \mu_a - \min_{a \in \mathcal{A}} \mu_a \le |V|$. Rescaling rewards via

$$\mu_a \;\longmapsto\; \tilde{\mu}_a = \frac{\mu_a - \min_b \mu_b}{\Delta_{\max}} \in [0, 1], \qquad \text{with noise scale} \quad \tilde{\sigma} = \frac{\sigma}{\Delta_{\max}},$$

leaves dynamic regret invariant up to the scale factor $\Delta_{\max}$. This canonical rescaling aligns with the information ratio framework of Russo & Van Roy (2016), where normalized rewards lie in $[0, 1]$ and noise is conditionally sub-Gaussian with parameter at most $\frac{1}{2}$, which is the calibration Theorem 5 runs on when $\sigma \le \Delta_{\max}/2$.

Analytical role of the noise level. The leading constant of Theorem 5 is $\max\{\sigma, \Delta_{\max}/2\}$, continuous in both arguments, so a deployment above $\sigma = \Delta_{\max}/2$ scales it with the noise and leaves the form of the certificate intact. Every algorithm here runs unchanged across the two regimes, and §6 evaluates instances on both sides.

Network scaling and empirical verification. Under population normalization $\|s\|_2^2 = |V|$, the reward range $\Delta_{\max}$ remains $\mathcal{O}(1)$ as the network expands, which is the calibration Table 24 evaluates Theorem 5 on. The streaming generator of §6 adopts unit norm $\|s\|_2 = 1$, causing $\Delta_{\max}$ to contract with network size. Our robustness evaluations further raise the noise tenfold to $\sigma = 1.0$, where performance stays consistent throughout the range evaluated.

B Full proofs and technical derivations

This section provides complete mathematical proofs for all theoretical results stated in §3 and §4. Throughout, $\mathbb{S}^{|V|}$ denotes the Hilbert space of real symmetric $|V| \times |V|$ matrices endowed with the trace inner product $\langle A, B \rangle = \operatorname{tr}(AB)$ and induced Frobenius norm $\|A\|_F = \sqrt{\langle A, A \rangle}$, and $\mathbb{S}^n_{++}$ denotes the cone of symmetric positive definite $n \times n$ matrices. The unperturbed forest resolvent is denoted by $M = (I + L)^{-1} \in \mathbb{S}^{|V|}_{++}$, and $z = Ms$ denotes the unperturbed equilibrium opinion vector. For any candidate edge $e = \{i, j\}$, $b_e = \mathbf{e}_i - \mathbf{e}_j \in \mathbb{R}^{|V|}$ denotes its signed incidence vector, satisfying $b_e^\top z = z_i - z_j$ and $b_e^\top M b_e = M_{ii} + M_{jj} - 2M_{ij}$.

Proof of Proposition 1. Let intervention $a = (E_a, \delta)$ modify $k = |E_a|$ candidate node pairs. Collecting the corresponding incidence vectors as columns of $B = [\, b_e \,]_{e \in E_a} \in \mathbb{R}^{|V| \times k}$ and setting $\Delta_a = \operatorname{diag}(\delta) \in \mathbb{S}^k_{++}$, the intervened Laplacian admits the canonical dyadic perturbation form

$$L_a \;=\; L + \sum_{e \in E_a} \delta_e \, b_e b_e^\top \;=\; L + B \, \Delta_a \, B^\top.$$

The shifted operator satisfies $I + L_a = (I + L) + B \, \Delta_a \, B^\top$. Because $I + L \succ 0$ and $\Delta_a \succ 0$, applying the Woodbury operator identity (Horn & Johnson, 2012) yields the exact resolvent

$$X_a \;=\; (I + L_a)^{-1} \;=\; M - MB \, \Lambda_a^{-1} B^\top M, \qquad \text{where} \quad \Lambda_a \;:=\; \Delta_a^{-1} + B^\top M B \;\in\; \mathbb{S}^k_{++}. \tag{12}$$

Table 2: Summary of notation.

Symbol	Meaning	Symbol	Meaning
$G = (V, E)$	graph, $\lvert V\rvert$ nodes	$\psi_a = \mathrm{vecsym}(X_a)$	full lift feature
L	graph Laplacian	$\theta^\star = \mathrm{vecsym}(ss^\top)$	lifted parameter
$M = (I + L)^{-1}$	resolvent (forest matrix)	$d_\star$	$\dim \mathrm{span}\{\psi_a\}$
s	innate opinions, $\mathbf{1}^\top s = 0$	r	innate rank (homophily)
$z = Ms$	FJ equilibrium	$\mathcal{S}_r = \mathrm{span}(U_r)$	innate subspace
$\mathrm{P}(s, L) = s^\top Ms$	polarization and disagreement	U_r, U_p	leading Laplacian eigenvectors
a, K	arm, number of arms	$p\ (\geq r)$	eigenvectors kept by policies
L_a, $X_a = (I + L_a)^{-1}$	intervened Laplacian, feature	$d_{\mathrm{eff}} \leq r(r+1)/2$	effective dimension
k, δ_e, b_e	pairs per arm, weight, incidence	$\varphi_a = \mathrm{vecsym}(U_p^\top X_a U_p)$	spectral feature
B	incidence matrix of an arm	$d_p = p(p+1)/2$	spectral feature dimension
$\Theta^\star = ss^\top$	rank one parameter (matrix)	θ_t	time varying parameter
$\mu_a = \langle X_a, \Theta^\star \rangle$	mean reward of arm a	W, W_{med}	window, medium window
σ	observation noise	$\mathcal{B}$	variation budget
T, t	horizon, round	J	episodes per stream
Regret_T	cumulative dynamic regret	g_ϕ, g_ω	pretrained policy, LDT filter
Δ	joint margin under degeneracy	$\Delta_{\min}, \Delta_{\max}$	minimum gap, reward range
m, γ	sample size, failure probability	$\varepsilon_{\mathrm{pre}}, R_{\mathrm{pre}}$	imitation error, pretraining regret
$\bar{\theta}_t, \xi_t$	shared drift, residual	$\mathcal{B}_\xi$	residual variation budget
A, D	adjacency, degrees	$\lambda_i(L), \lambda_2$	Laplacian spectrum
$\mathcal{A}$, E_a	arm pool, node pairs of an arm	y_t, ζ_t	observation, noise

The capacitance matrix Λ_a is strictly positive definite because $\Delta_a^{-1} \succ 0$ (all edge weights $\delta_e > 0$) and $B^\top MB \succeq 0$ by virtue of $M \succ 0$. Hence, the inverse Λ_a^{-1} exists and satisfies $\Lambda_a^{-1} \succ 0$.

Contracting equation 12 by the innate opinion vector s and leveraging the symmetry $M = M^\top$ alongside the equilibrium relation $z = Ms$, the objective shift evaluates to

$$\Delta\mathrm{P}(a) \;:=\; s^\top X_a s - s^\top Ms \;=\; -s^\top MB\,\Lambda_a^{-1} B^\top Ms \;=\; -(B^\top z)^\top \Lambda_a^{-1}(B^\top z). \tag{13}$$

Because $\Lambda_a^{-1} \succ 0$, the quadratic form in equation 13 is nonpositive for all $s \in \mathbb{R}^{|V|}$, establishing that $\Delta\mathrm{P}(a) \leq 0$ universally across all graphs and candidate interventions. The shift vanishes if and only if $B^\top z = \mathbf{0}$, corresponding to the degenerate state where every intervened pair connects agents possessing identical equilibrium beliefs, so one disagreeing pair makes the shift negative.

For a rank one intervention ($k = 1$) on edge $e = \{i, j\}$ with weight increment $\delta_e > 0$, the incidence matrix collapses to column vector $B = b_e$. The capacitance matrix reduces to the scalar $\Lambda_a = \delta_e^{-1} + b_e^\top Mb_e$. Evaluating equation 12 via the Sherman-Morrison theorem yields

$$X_a \;=\; M - \frac{\delta_e\, Mb_e b_e^\top M}{1 + \delta_e\, b_e^\top Mb_e}.$$

Contracting with s and substituting $b_e^\top Ms = z_i - z_j$ alongside $b_e^\top Mb_e = M_{ii} + M_{jj} - 2M_{ij}$,

$$\Delta\mathrm{P}(a) \;=\; -\frac{\delta_e\,(b_e^\top z)^2}{1 + \delta_e\, b_e^\top Mb_e} \;=\; -\frac{\delta_e\,(z_i - z_j)^2}{1 + \delta_e\,(M_{ii} + M_{jj} - 2M_{ij})},$$

which establishes the single link expression and concludes the proof. ■

Proposition 3 (Monotonicity, convexity, and nonsubmodularity)**.** The objective functional $\mathrm{P}(\delta) = s^\top (I + L + \sum_e \delta_e b_e b_e^\top)^{-1} s$ is monotone nonincreasing and jointly convex on the nonnegative orthant

$\mathbb{R}_+^{|E_{\text{cand}}|}$ of weight increments carried by the candidate pair set $E_{\text{cand}} = \bigcup_{a\in\mathcal{A}} E_a$. Furthermore, the set function $F(S) := \mathrm{P}(\emptyset) - \mathrm{P}(S)$ induced by unit increments is strictly nonsubmodular, as witnessed by an exact rational counterexample on five nodes.

Proof of Proposition 3. Define the parameterized operator $Y(\delta) := I + L + \sum_e \delta_e b_e b_e^\top$. The mapping $\delta \mapsto Y(\delta)$ is affine, and $Y(\delta) \in \mathbb{S}_{++}^{|V|}$ for all $\delta \geq \mathbf{0}$ because $I + L \succ 0$ and $b_e b_e^\top \succeq 0$. The objective $\mathrm{P}(\delta) = s^\top Y(\delta)^{-1} s$ is infinitely differentiable on this domain.

To establish monotonicity, consider the directional derivative along standard basis vector $\mathbf{e}_e$. Applying the matrix differentiation identity $\nabla_Y Y^{-1} = -Y^{-1}(\nabla Y)Y^{-1}$,

$$\frac{\partial \mathrm{P}}{\partial \delta_e} \;=\; -\,s^\top Y(\delta)^{-1} b_e b_e^\top Y(\delta)^{-1} s \;=\; -\left(b_e^\top Y(\delta)^{-1} s\right)^2 \;\leq\; 0. \tag{14}$$

Because all first partial derivatives are nonpositive throughout $\mathbb{R}_+^{|E_{\text{cand}}|}$, P is monotone nonincreasing, so adding weight to any candidate pair never increases the objective.

For joint convexity, let $H \in \mathbb{S}^{|V|}$ be an arbitrary symmetric matrix and consider the scalar slice $f(\epsilon) = s^\top (Y + \epsilon H)^{-1} s$ for sufficiently small ϵ. Differentiating twice with respect to ϵ,

$$\begin{aligned} f'(0) \;&=\; -\,s^\top Y^{-1} H Y^{-1} s, \\ f''(0) \;&=\; 2\, s^\top Y^{-1} H Y^{-1} H Y^{-1} s \;=\; 2\left\|Y^{-1/2} H Y^{-1} s\right\|_2^2 \;\geq\; 0. \end{aligned}$$

Thus, the operator inversion map $Y \mapsto s^\top Y^{-1} s$ is convex on $\mathbb{S}_{++}^{|V|}$. Because affine precomposition preserves convexity, $\mathrm{P}(\delta)$ is jointly convex in δ.

To analyze submodularity, define the marginal gain operator $g_S(e) := F(S \cup \{e\}) - F(S) = \mathrm{P}(S) - \mathrm{P}(S \cup \{e\})$. Submodularity of F requires $g_S(e) \geq g_{S'}(e)$ whenever $S \subseteq S'$ and $e \notin S'$. Consider the unweighted path graph P_5 on vertices $\{1, 2, 3, 4, 5\}$ with unit edge weights, and set innate opinions to

$$s \;=\; \left(2,\, -1,\, -1,\, \tfrac{3}{2},\, -\tfrac{3}{2}\right)^\top \;\in\; \mathbf{1}^\perp, \qquad S = \emptyset, \qquad S' = \big\{\{2, 4\}\big\}, \qquad e = \{1, 5\},$$

with unit intervention weights $\delta = 1$. Performing matrix inversions in exact rational arithmetic yields

$$\mathrm{P}(\emptyset) \;=\; \frac{393}{110}, \qquad \mathrm{P}(\{e\}) \;=\; \frac{53}{22}, \qquad g_\emptyset(e) \;=\; \mathrm{P}(\emptyset) - \mathrm{P}(\{e\}) \;=\; \frac{64}{55}.$$

Evaluating the perturbed states under S' gives

$$\mathrm{P}(S') \;=\; \frac{155}{44}, \qquad \mathrm{P}(S' \cup \{e\}) \;=\; \frac{1845}{836}, \qquad g_{S'}(e) \;=\; \mathrm{P}(S') - \mathrm{P}(S' \cup \{e\}) \;=\; \frac{25}{19}.$$

Comparing marginal returns reveals a strict increase:

$$g_{S'}(e) - g_\emptyset(e) \;=\; \frac{25}{19} - \frac{64}{55} \;=\; \frac{159}{1045} \;>\; 0.$$

This violates the diminishing returns condition, proving that the set function F is not submodular.

The algebraic mechanism follows from equation 13, where the marginal efficacy of link $\{1, 5\}$ scales with $(z_1 - z_5)^2$. Strengthening the intermediate link $\{2, 4\}$ alters network diffusion such that $|z_1 - z_5|$ increases from $\frac{8}{5}$ to $\frac{5}{3}$, strictly amplifying the leverage of the subsequent intervention. Because the construction is homogeneous of degree two in s, rescaling s modulates gains by a constant factor, preserving nonsubmodularity under arbitrary scale normalizations. ■

Lemma 1 (Intrinsic dimension of attainable rewards)**.** The nonlinear evaluation map $\Phi : s \mapsto (\mu_a(s))_{a\in\mathcal{A}}$ has differential $d\Phi_s$ with rows $(2X_a s)^\top$, implying that the image set $\Phi(\mathbb{R}^{|V|})$ has Hausdorff dimension at most $|V|$, whereas its linear span can attain $\min(K, |V|(|V|+1)/2)$.

Proof of Lemma 1. Each coordinate $\mu_a(s) = \langle X_a, ss^\top \rangle = s^\top X_a s$ is a quadratic form in s. The Fréchet derivative of μ_a at s in direction $h \in \mathbb{R}^{|V|}$ evaluates to

$$d\mu_a(s)[h] \;=\; 2\, s^\top X_a h \;=\; \langle 2X_a s,\, h \rangle.$$

The Jacobian $J_\Phi(s) \in \mathbb{R}^{K\times|V|}$ has rows $(2X_a s)^\top$. Because $J_\Phi(s)$ is a linear operator from $\mathbb{R}^{|V|}$ to $\mathbb{R}^K$, its rank is bounded by the domain dimension, so $\operatorname{rank}(J_\Phi(s)) \le |V|$ for all $s \in \mathbb{R}^{|V|}$.

The polynomial mapping $\Phi : \mathbb{R}^{|V|} \to \mathbb{R}^K$ is locally Lipschitz continuous on every compact ball in $\mathbb{R}^{|V|}$. By the geometric measure theory theorem on Hausdorff measures under Lipschitz transformations (Evans & Gariepy, 2015), a locally Lipschitz map cannot increase Hausdorff dimension:

$$\dim_{\mathcal{H}}\big(\Phi(\mathbb{R}^{|V|})\big) \;\le\; \dim_{\mathcal{H}}(\mathbb{R}^{|V|}) \;=\; |V|.$$

Conversely, attainable rewards represent the linear image of the quadratic cone $\mathcal{C} := \{ss^\top : s \in \mathbb{R}^{|V|}\}$ under the linear measurement operator $\mathcal{T} : \mathbb{S}^{|V|} \to \mathbb{R}^K$ defined by $\mathcal{T}(\Theta) = (\langle X_a, \Theta\rangle)_{a\in\mathcal{A}}$. Because the linear span of $\mathcal{C}$ is the entire symmetric matrix space $\mathbb{S}^{|V|}$,

$$\operatorname{span}\big(\Phi(\mathbb{R}^{|V|})\big) \;=\; \operatorname{span}\big(\mathcal{T}(\mathcal{C})\big) \;=\; \mathcal{T}\big(\operatorname{span}(\mathcal{C})\big) \;=\; \mathcal{T}\big(\mathbb{S}^{|V|}\big) \;=\; \operatorname{range}(\mathcal{T}).$$

The dimension of $\operatorname{range}(\mathcal{T})$ equals the rank of $\mathcal{T}$, which is precisely $\dim \operatorname{span}\{X_a : a \in \mathcal{A}\}$. This subspace lies in $\mathbb{S}^{|V|}$, of dimension $|V|(|V|+1)/2$, and is spanned by the K feature operators, so its dimension is at most $\min(K, |V|(|V|+1)/2)$ and attains that value exactly when the K operators are linearly independent, which is the generic case and is the quantity $d_\star$ certified by Theorem 5. The span therefore grows to $\min(K, |V|(|V|+1)/2)$ while the image itself stays within a set of Hausdorff dimension $|V|$, confirming the claimed dimensional gap. ■

Theorem 5 (Structure aware Bayes regret)**.** Exact posterior sampling on the lifted structural features $\{\psi_a\}_{a\in\mathcal{A}}$ satisfies

$$\mathbb{E}[\mathrm{Regret}_T] \;\le\; \max\{\sigma, \Delta_{\max}/2\}\,\sqrt{2\,d_\star T \log K} \;=\; \mathcal{O}\big(\max\{\sigma, \Delta_{\max}\}\sqrt{d_\star T \log K}\big),$$

with no low rank requirement on s and no condition relating σ to $\Delta_{\max}$: the expectation runs over whatever prior generates innate opinions.

Proof of Theorem 5. Under symmetric vectorization $\psi_a = \operatorname{vecsym}(X_a) \in \mathbb{R}^d$ and parameter $\theta^\star = \operatorname{vecsym}(ss^\top) \in \mathbb{R}^d$ with ambient dimension $d = |V|(|V|+1)/2$, the trace inner product identity yields $\mu_a = \langle X_a, ss^\top\rangle = \psi_a^\top\theta^\star$ exactly. The interaction model therefore constitutes an exact linear bandit, with features fixed by the graph and the parameter by the innate opinions.

Let $\mathcal{L} := \operatorname{span}\{\psi_a : a \in \mathcal{A}\} \subset \mathbb{R}^d$ denote the structural feature span with intrinsic dimension $d_\star = \dim\mathcal{L} \le \min(K, d)$, and let $\Pi_\mathcal{L}$ denote the orthogonal projection operator onto $\mathcal{L}$. Because $\psi_a \in \mathcal{L}$ for all $a \in \mathcal{A}$,

$$\mu_a \;=\; \psi_a^\top\theta^\star \;=\; \psi_a^\top(\Pi_\mathcal{L}\theta^\star + \Pi_\mathcal{L}^\perp\theta^\star) \;=\; \psi_a^\top\Pi_\mathcal{L}\theta^\star.$$

Let $V_\mathcal{L} \in \mathbb{R}^{d\times d_\star}$ be an orthonormal basis matrix for $\mathcal{L}$, so that $\psi_a = V_\mathcal{L}\tilde\psi_a$ with $\tilde\psi_a \in \mathbb{R}^{d_\star}$. The system is algebraically isomorphic to a well specified linear bandit of ambient dimension $d_\star$ governed by parameter $\tilde\theta^\star = V_\mathcal{L}^\top\theta^\star$, with the orthogonal component $\Pi_\mathcal{L}^\perp\theta^\star$ unidentifiable and uncoupled from rewards, so the effective dimension of the problem is the intrinsic $d_\star$.

Under canonical scale calibration, we normalize payoffs to the unit interval via $\tilde\mu_a = (\mu_a - \min_b \mu_b)/\Delta_{\max} \in [0,1]$ and define normalized noise $\tilde\zeta_t = \zeta_t/\Delta_{\max}$. Under the standing condition $\sigma \le \Delta_{\max}/2$, the noise scale satisfies $\tilde\sigma \le \frac{1}{2}$. The information ratio framework of Russo & Van Roy (2016) upper bounds the Bayesian regret of Thompson sampling by

$$\mathbb{E}\big[\mathrm{Regret}_T\big] \;\le\; \Delta_{\max}\sqrt{\bar\Gamma\, H(A^\star)\, T}, \tag{15}$$

where $\bar\Gamma$ is the worst case information ratio and $H(A^\star)$ is the Shannon entropy of the optimal action. For a linear bandit in Euclidean space $\mathbb{R}^{d_\star}$ with sub-Gaussian observation noise variance $\tilde\sigma^2$, Proposition 5 of Russo & Van Roy (2016) establishes that $\bar\Gamma \le d_\star/2$. Specifically, for any two candidate parameters predicting action means $\tilde\mu_1, \tilde\mu_2$, the Kullback-Leibler divergence between observation distributions satisfies

$$D_{\mathrm{KL}}\big(\mathcal{N}(\tilde\mu_1, \tilde\sigma^2)\,\|\,\mathcal{N}(\tilde\mu_2, \tilde\sigma^2)\big) \;=\; \frac{(\tilde\mu_1 - \tilde\mu_2)^2}{2\tilde\sigma^2} \;\ge\; 2\,(\tilde\mu_1 - \tilde\mu_2)^2,$$

which supplies the mutual information lower bound required for $\bar{\Gamma} \le d_\star/2$. Bounding entropy by $H(A^\star) \le \log K$ and substituting into equation 15 yields

$$\mathbb{E}\big[\mathrm{Regret}_T\big] \;\le\; \Delta_{\max}\sqrt{\tfrac{1}{2}\,d_\star T\log K} \;=\; \tfrac{\Delta_{\max}}{2}\sqrt{2\,d_\star T\log K}.$$

Above that noise level the same chain runs on a different calibration. Normalizing instead by 2σ, so that $\tilde{\mu}_a = (\mu_a - \min_b \mu_b)/(2\sigma)$ and $\tilde{\sigma} = \frac{1}{2}$ exactly, leaves the noise hypothesis of Proposition 5 of Russo & Van Roy (2016) satisfied and $\bar{\Gamma} \le d_\star/2$ unchanged, that proposition constraining the observation noise. Reading equation 15 in those units and returning to the original scale gives $\mathbb{E}[\mathrm{Regret}_T] \le 2\sigma\sqrt{\tfrac{1}{2}d_\star T\log K} = \sigma\sqrt{2\,d_\star T\log K}$. Taking whichever calibration applies gives the stated $\max\{\sigma, \Delta_{\max}/2\}\sqrt{2\,d_\star T\log K}$ at every noise level, with the first branch in force throughout the instances evaluated in Appendix E. The linear decomposition $\mu_a = \psi_a^\top\theta^\star$ holds identically without restricting s, and the information ratio argument uses only that the sampler draws from the exact posterior induced by whatever prior generates s. The certificate is therefore a statement about the lifted structure $\{\psi_a\}$ together with exact posterior sampling on it, and it is valid for arbitrary innate opinion priors in that sense. ■

Theorem 1 is stated about a sampler written in the ridge estimate and design matrix of the played features, while Algorithm 1 carries a conjugate Gaussian posterior. The next lemma identifies the two exactly, so the constants the theorem asks for can be read off the implemented ones.

Lemma 2 (The implemented sampler is the analysed one)**.** Run Algorithm 1 at prior scale τ, noise scale σ and sampling scale c, and write (μ_t, Σ_t) for the pair it carries after t observations. Put $\lambda := \sigma^2/\tau^2$, $V_t := \lambda I + \sum_{u\le t}\psi_{a_u}\psi_{a_u}^\top$ and $b_t := \sum_{u\le t}\psi_{a_u}y_u$. Then at every round

$$\Sigma_t = \sigma^2 V_t^{-1}, \qquad \mu_t = V_t^{-1}b_t = \hat{\theta}_t, \tag{16}$$

so the carried mean is the ridge estimate at regularizer λ, and the played draw $\mu_t + c\,C_t\epsilon$ has the law of $\hat{\theta}_t + c_{\mathrm{eff}}V_t^{-1/2}\epsilon$ at $c_{\mathrm{eff}} = c\sigma$. Algorithm 1 is therefore the sampler of Theorem 1 run at ridge $\lambda = \sigma^2/\tau^2$ and inflation $c_{\mathrm{eff}} = c\sigma$.

Proof. Both identities follow by induction. At $t = 0$, $\Sigma_0^{-1} = \tau^{-2}I = \sigma^{-2}\lambda I = \sigma^{-2}V_0$ and $\mu_0 = \mathbf{0} = V_0^{-1}b_0$. Assume equation 16 at $t-1$ and write $\psi := \psi_{a_t}$, $y := y_t$. The covariance step of Algorithm 1 is $\Sigma_t = \Sigma_{t-1} - \Sigma_{t-1}\psi\psi^\top\Sigma_{t-1}/(\sigma^2 + \psi^\top\Sigma_{t-1}\psi)$, which is the Sherman-Morrison form of $(\Sigma_{t-1}^{-1} + \sigma^{-2}\psi\psi^\top)^{-1}$, so

$$\Sigma_t^{-1} = \Sigma_{t-1}^{-1} + \sigma^{-2}\psi\psi^\top = \sigma^{-2}\big(V_{t-1} + \psi\psi^\top\big) = \sigma^{-2}V_t.$$

For the mean, that same step gives $\Sigma_t\psi = \sigma^2\Sigma_{t-1}\psi/(\sigma^2 + \psi^\top\Sigma_{t-1}\psi)$, so the update $\mu_t = \mu_{t-1} + (y - \psi^\top\mu_{t-1})\Sigma_{t-1}\psi/(\sigma^2 + \psi^\top\Sigma_{t-1}\psi)$ reads $\mu_t = \mu_{t-1} + \sigma^{-2}\Sigma_t\psi(y - \psi^\top\mu_{t-1})$. Multiplying by Σ_t^{-1} and substituting $\Sigma_t^{-1} = \Sigma_{t-1}^{-1} + \sigma^{-2}\psi\psi^\top$ cancels the two terms in $\psi\psi^\top\mu_{t-1}$ and leaves $\Sigma_t^{-1}\mu_t = \Sigma_{t-1}^{-1}\mu_{t-1} + \sigma^{-2}\psi y$; multiplying by σ^2 turns this into $V_t\mu_t = V_{t-1}\mu_{t-1} + \psi y = b_{t-1} + \psi y = b_t$, which is the second identity. Finally C_t is a Cholesky factor of Σ_t, so $C_t\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma_t) = \mathcal{N}(\mathbf{0}, \sigma^2V_t^{-1})$ and $c\,C_t\epsilon$ has the law of $c\sigma V_t^{-1/2}\epsilon$. ■

Lemma 3 (The optimism probability of a Gaussian draw)**.** Let $a^\star \in \arg\min_a \mu_a$, note $\psi_{a^\star} \neq 0$ because $X_{a^\star} \succ 0$, write $w_t := \|\psi_{a^\star}\|_{V_t^{-1}} > 0$, and let

$$G_t \;:=\; \frac{\psi_{a^\star}^\top(\hat{\theta}_t - \theta^\star)}{w_t}$$

be the prediction error at that arm in units of its own confidence width, a quantity measurable with respect to $\mathcal{F}_{t-1}$. Then the draw of Theorem 1 is optimistic with conditional probability

$$\Pr\Big[\min_a \psi_a^\top\tilde{\theta}_t \le \min_a \mu_a \;\Big|\; \mathcal{F}_{t-1}\Big] \;\ge\; \bar{\Phi}\big(G_t/c\big)$$

at every inflation $c > 0$, with equality when the minimum on the left is attained at $a^\star$. Two readings follow. On the event $\mathcal{E}_\delta$ of Theorem 1, $G_t \le \beta_T(\delta)$ at every round, which is the per history constant $p_{\mathrm{opt}}(c,\delta) = \bar{\Phi}(\beta_T(\delta)/c)$ that theorem carries. Off that event the reading is a frequency: for any level u and any $q \le \Pr[G_t \le u]$ the event occurs at rate at least $q\,\bar{\Phi}(u/c)$, which at $u = 0$ is $q/2$ at every inflation, a floor the deployed inflation inherits with no horizon dependence.

Proof. Since $\min_a \mu_a = \mu_{a^\star} = \psi_{a^\star}^\top \theta^\star$ and $\min_a \psi_a^\top \tilde{\theta}_t \le \psi_{a^\star}^\top \tilde{\theta}_t$, the optimistic event contains $\{\psi_{a^\star}^\top \tilde{\theta}_t \le \psi_{a^\star}^\top \theta^\star\}$, and the two coincide when $a^\star$ attains the left minimum. At the unit vector $u_t := V_t^{-1/2} \psi_{a^\star} / w_t$, which is $\mathcal{F}_{t-1}$ measurable, that event reads $w_t\big(G_t + c\, u_t^\top \eta_t\big) \le 0$, and $u_t^\top \eta_t$ is standard normal given $\mathcal{F}_{t-1}$ because η_t is isotropic and drawn fresh, so its conditional probability is $\bar{\Phi}(G_t/c)$. On the self normalized event Cauchy-Schwarz gives $G_t \le \|\hat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta)$ and $\bar{\Phi}$ decreases, which is the first reading. For the second, take expectations and discard the nonnegative contribution of $\{G_t > u\}$, leaving $\mathbb{E}[\bar{\Phi}(G_t/c)] \ge \bar{\Phi}(u/c) \Pr[G_t \le u]$. ■

Proof of Theorem 1. Lift-TS maintains a Gaussian working posterior on the lift: by Lemma 2 it samples $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t,\, c^2 V_t^{-1})$ from the ridge estimate and design matrix of the played features and plays $a_t \in \arg\min_a \psi_a^\top \tilde{\theta}_t$, and we write c for the inflation in those coordinates throughout. The true parameter $\theta^\star = \mathrm{vecsym}(ss^\top)$ is positive semidefinite and rank one, so it lies on a measure zero cone that carries no Gaussian mass, and the working posterior is never the Bayes posterior of the generative model. Frequentist analyses of linear Thompson sampling do not require that it be. They fix $\theta^\star$ arbitrarily and treat the Gaussian as a randomization device, so the analysis rests on four algorithmic hypotheses, and we discharge each on our instance.

(H1) Initialization and design. The design is initialized at $V_0 = \lambda I$ for an arbitrary $\lambda > 0$ and updated only by played features, so $V_t \succeq \lambda I$ at every round and $\det V_T \le (\lambda + T L_\psi^2 / d_\star)^{d_\star}$. The elliptical potential argument of Abbasi-Yadkori et al. (2011) therefore runs on the realized design whatever arms greedy optimism selects, and no minimum eigenvalue or excitation condition on $\{\psi_{a_u}\}$ is needed: regularization, not exploration, is what keeps V_t invertible. The ridge reaches the rate through that potential alone. Writing $x_u = \|\psi_{a_u}\|^2_{V_{u-1}^{-1}} \le L_\psi^2/\lambda =: R$, the elementary inequality $\min\{1, x\} \ge x / \max\{1, R\}$ on $[0, R]$ turns Lemma 11 of Abbasi-Yadkori et al. (2011) into

$$\sum_{u \le T} \|\psi_{a_u}\|^2_{V_{u-1}^{-1}} \;\le\; 2 \max\{1, L_\psi^2/\lambda\}\; d_\star \log\big(1 + T L_\psi^2/(d_\star \lambda)\big),$$

so the leading constant carries the factor $\sqrt{\kappa_\lambda}$ the statement displays, which $\lambda \ge L_\psi^2$ sets to one and which every smaller regularizer carries at its own value. By Lemma 2 the deployed configuration is this hypothesis at $\lambda = \sigma^2/\tau^2$.

(H2) Covariance inflation and optimism. The sampled parameter is $\tilde{\theta}_t = \hat{\theta}_t + c V_t^{-1/2} \eta_t$ with $\eta_t \sim \mathcal{N}(0, I_{d_\star})$ drawn fresh at each round independently of $\mathcal{F}_{t-1}$. The perturbation is therefore a multivariate normal in the V_t geometry, which satisfies both properties the analysis of Abeille & Lazaric (2017) requires of a sampling distribution: concentration, $\Pr[\|\eta_t\|_2 \le \sqrt{d_\star \log(d_\star/\delta)}] \ge 1 - \delta$, and anti concentration, $\Pr[u^\top \eta_t \le -x] = \bar{\Phi}(x)$ for every unit u, because η_t is isotropic and independent of the arm set. Both hold at every inflation, since η_t is a standard normal whatever scale multiplies it. The scale enters at one place only, the probability p_{opt} that a draw is optimistic, and it enters explicitly. Lemma 3 evaluates it: conditionally on the history the probability is at least $\bar{\Phi}(G_t/c)$ at the pointwise error G_t the lemma names, and on $\mathcal{E}_\delta$ that error is at most $\beta_T(\delta)$ at every round and for every history the event admits. Hence $p_{\mathrm{opt}}(c, \delta) = \bar{\Phi}(\beta_T(\delta)/c)$ there, nondecreasing in the inflation, and $c \ge \beta_T(\delta)$ secures $p_{\mathrm{opt}}(c, \delta) \ge \bar{\Phi}(1) > 1/7$. The constant is a property of $\mathcal{E}_\delta$, whose complement carries probability at most δ and enters the conclusion additively. Appendix E measures the frequency the lemma reads off a quantile of G_t at the deployed $c_{\mathrm{eff}} = c\sigma$ of Lemma 2.

(H3) Bounded features and parameter, (H4) sub-Gaussian noise. These are properties of the instance and we verify them next.

Bounded features. Since $L_a \succeq 0$ we have $I + L_a \succeq I$, hence $0 \prec X_a = (I + L_a)^{-1} \preceq I$ and all $|V|$ eigenvalues of X_a lie in $(0, 1]$, giving $\|X_a\|_F \le \sqrt{|V|}$. Because vecsym is a linear isometry from $\mathbb{S}^{|V|}$ under the Frobenius inner product onto $\mathbb{R}^d$ under the Euclidean one, $\|\psi_a\|_2 = \|X_a\|_F \le \sqrt{|V|} =: L_\psi$.

Bounded parameter. $\|\theta^\star\|_2 = \|ss^\top\|_F = \|s\|_2^2 = |V| =: S$ under the normalization fixed in §3, and only $\Pi_{\mathcal{L}} \theta^\star$, whose norm is no larger, is identified or used.

Noise. Observations are $\mu_{a_t} + \zeta_t$ with $\zeta_t \sim \mathcal{N}(0, \sigma^2)$ drawn independently of the history, hence conditionally σ sub-Gaussian, which is exactly what (H4) requires.

With (H1)–(H4) in hand, restricting to the identified subspace $\mathcal{L}$ exactly as in the proof of Theorem 5 leaves a well specified $d_\star$ dimensional linear bandit with the constants above, and Theorem 1 of Agrawal & Goyal (2013) in the form of Abeille & Lazaric (2017) applies verbatim on $\mathcal{E}_\delta$, giving with probability at least $1 - 2\delta$

$$\mathrm{Regret}_T \;\le\; \frac{C}{p_{\mathrm{opt}}(c,\delta)}\Big(\beta_T(\delta) + c\sqrt{d_\star \log(d_\star/\delta)}\Big)\sqrt{\kappa_\lambda\, d_\star\, T\, \ell_T} \;+\; C\, S\, L_\psi\, d_\star \log T,$$

with C an absolute constant, the first factor the radius of the sampled ellipsoid, the square root the elliptical potential of (H1) and $1/p_{\mathrm{opt}}(c,\delta)$ the price of anti concentration from (H2). Every quantity on the right is an instance constant discharged above, and no step of the argument evaluates the law of $\theta^\star$: the statement is made for each fixed s separately, which is what makes a working prior that assigns the rank one cone zero mass admissible. Network size enters through $L_\psi \le \sqrt{|V|}$ and $S \le |V|$ alone, and at $\kappa_\lambda = 1$ and $p_{\mathrm{opt}}(c,\delta) \ge \bar{\Phi}(1)$ the noise driven part of the leading term is $\tilde{\mathcal{O}}(d_\star^{3/2}\sqrt{T})$ in the normalized units of §4 under $\sigma \le \Delta_{\max}/2$, a factor $\tilde{\mathcal{O}}(d_\star)$ above the $\Delta_{\max}\sqrt{\frac{1}{2}d_\star T \log K}$ of Theorem 5, which exact posterior sampling attains and which no conjugate Gaussian update realizes on this problem. ■

Proof of Theorem 2. Under Assumption 1, innate opinions satisfy $s \in \mathcal{S}_r$, where $\mathcal{S}_r = \mathrm{span}(U_r) \subset \mathbf{1}^\perp$. Consequently, the rank one matrix $\Theta^\star = ss^\top$ is supported strictly on the low rank tensor subspace

$$\mathcal{W}_r \;:=\; \mathrm{Sym}(\mathcal{S}_r) \;=\; \Big\{ \sum_{i,j\le r} \alpha_{ij} u_i u_j^\top : \alpha_{ij} = \alpha_{ji} \Big\} \;\subset\; \mathbb{S}^{|V|},$$

which has intrinsic dimension $\dim \mathcal{W}_r = r(r+1)/2$. Let $\Pi_{\mathcal{W}_r} : \mathbb{S}^{|V|} \to \mathcal{W}_r$ denote the orthogonal projection operator onto $\mathcal{W}_r$ under the Frobenius inner product, and write $d_{\mathrm{eff}} := \dim \mathrm{span}\{\Pi_{\mathcal{W}_r} X_a : a \in \mathcal{A}\} \le r(r+1)/2$ for the dimension the pool actually spans there, with equality when it spans $\mathcal{W}_r$. Because $\Theta^\star \in \mathcal{W}_r$, the orthogonal complement vanishes, so $\Pi^\perp_{\mathcal{W}_r}\Theta^\star = \mathbf{0}$. Self-adjointness of orthogonal projections implies

$$\mu_a \;=\; \langle X_a, \Theta^\star\rangle \;=\; \langle \Pi_{\mathcal{W}_r} X_a + \Pi^\perp_{\mathcal{W}_r} X_a, \Theta^\star\rangle \;=\; \langle \Pi_{\mathcal{W}_r} X_a, \Theta^\star\rangle.$$

The interaction model is therefore isomorphic to a linear bandit on the low dimensional feature space $\mathcal{W}_r$. Applying Theorem 5 on this projected space yields the intrinsic Bayes regret bound $\mathcal{O}(\sqrt{d_{\mathrm{eff}} T \log K}) = \tilde{\mathcal{O}}(r\sqrt{T})$ in the normalized regret units of §4.

Now consider an in context policy pretrained via imitation learning to approximate exact posterior sampling within per step total variation error $\varepsilon_{\mathrm{pre}}$. Let $p_t(\cdot|h_t)$ and $\pi_t(\cdot|h_t)$ denote the action distributions of the exact posterior sampler and the pretrained policy at round t, respectively, satisfying $\|p_t - \pi_t\|_{\mathrm{TV}} \le \varepsilon_{\mathrm{pre}}$. Let $\Delta_t(a) = \mu_a - \mu^\star \le \Delta_{\max}$ denote the instantaneous suboptimality gap, where $\mu^\star = \min_{b\in\mathcal{A}} \mu_b$. The expected regret increment at round t decomposes as

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_t}[\Delta_t(a)] \;&=\; \mathbb{E}_{a\sim p_t}[\Delta_t(a)] + \sum_{a\in\mathcal{A}} \big(\pi_t(a) - p_t(a)\big)\Delta_t(a) \\
&\le\; \mathbb{E}_{a\sim p_t}[\Delta_t(a)] + \underbrace{\sum_{a\in\mathcal{A}} \big|\pi_t(a) - p_t(a)\big|\,\Delta_t(a)}_{\text{gap weighted}} \\
&\le\; \mathbb{E}_{a\sim p_t}[\Delta_t(a)] + 2\Delta_{\max}\,\|p_t - \pi_t\|_{\mathrm{TV}}.
\end{aligned}$$

The gap weighted middle quantity is recorded as Proposition 4 below, because the last inequality, which replaces every gap by the largest one, is the only place $\Delta_{\max}$ and hence total variation enters, and it is loose exactly when the pool contains near ties. When pretraining data is collected interactively on-policy (Algorithm 3), the expectation is evaluated with respect to the state occupancy measure induced by the learner itself. Summing across horizon T yields a linear imitation error penalty:

$$R_{\mathrm{pre}} \;\le\; \Delta_{\max} \sum_{t=1}^{T} \|p_t - \pi_t\|_{\mathrm{TV}} \;=\; \mathcal{O}(T\varepsilon_{\mathrm{pre}}).$$

Under arbitrary distribution shift, where the evaluation trajectory is autonomous and open loop, standard error propagation bounds for sequential decision processes (Ross & Bagnell, 2010) apply, and the distribution of visited histories h_t drifts from the pretraining distribution by at most $t\,\varepsilon_{\text{pre}}$ in total variation distance. Bounding the regret increment at round t by $\Delta_{\max} t\,\varepsilon_{\text{pre}}$ and summing across rounds gives

$$R_{\text{pre}} \;\le\; \Delta_{\max} \sum_{t=1}^{T} t\,\varepsilon_{\text{pre}} \;=\; \mathcal{O}(T^2 \varepsilon_{\text{pre}}),$$

which completes the proof. ■

Proposition 4 (Gap weighted amortization)**.** Let π be any policy and let p be exact posterior sampling evaluated on the same histories. With $\Delta_t(a) = \mu_a - \min_b \mu_b$,

$$\mathbb{E}[\text{Regret}_T(\pi)] \;\le\; \mathbb{E}[\text{Regret}_T(p)] \;+\; \sum_{t=1}^{T} \mathbb{E}\Big[\sum_{a \in \mathcal{A}} \big|\pi_t(a) - p_t(a)\big|\, \Delta_t(a) \Big],$$

and the second term is the R_{pre} of Theorem 2, at most $2\Delta_{\max} \sum_t \mathbb{E}\|\pi_t - p_t\|_{\text{TV}} = \mathcal{O}(T\varepsilon_{\text{pre}})$.

Proof. Immediate from the displayed decomposition in the proof of Theorem 2, summed over t and over the on policy history distribution; no hypothesis beyond measurability of the two action kernels is used. The final inequality substitutes $\Delta_t(a) \le \Delta_{\max}$ and $\sum_a |\pi_t(a) - p_t(a)| = 2\|\pi_t - p_t\|_{\text{TV}}$. ■

The two forms differ by what they charge for. Total variation charges a policy for every unit of probability mass it moves; the gap weighted form charges it only for mass moved onto arms that are actually worse. Appendix F measures both on held out instances and finds the distinction decisive: on an intervention pool that is nearly tied, total variation resolves the deployed in context policy only to the scale of the horizon, while the gap weighted form that Theorem 2 states lands within 3% of its realized regret on the same rounds.

The certificate for LDT differs in kind, because LDT is a filter that plays greedily, so it is stated in the currency the filter controls. Proposition 2 needs neither: it bounds the instantaneous regret of any estimator by the aligned error $(\varphi_{a_t^\star} - \varphi_{a_t})^\top e_t$, and relaxing that alignment reaches $D\|\hat{\theta}_t - \theta_t\|_2$, the quantity $\bar{\mathcal{L}}$ that equation 11 shapes offline and Table 23 evaluates along the deployed trajectory.

Proof of Proposition 2. Let $a^\star \in \arg\min_a \varphi_a^\top \theta_t$ and $e = \hat{\theta}_t - \theta_t$. Greedy selection gives $\varphi_{a_t}^\top \hat{\theta}_t \le \varphi_{a^\star}^\top \hat{\theta}_t$, so

$$\begin{aligned} \mu_{a_t} - \mu_{a^\star} \;&=\; (\varphi_{a_t} - \varphi_{a^\star})^\top \theta_t \;=\; \underbrace{(\varphi_{a_t} - \varphi_{a^\star})^\top \hat{\theta}_t}_{\le 0} + (\varphi_{a^\star} - \varphi_{a_t})^\top e \\ &\le\; \underbrace{\max_{a,b \in \mathcal{A}} (\varphi_a - \varphi_b)^\top e}_{\text{pairwise form}} \;\le\; \|\varphi_{a^\star} - \varphi_{a_t}\|_2\, \|e\|_2 \;\le\; D\|e\|_2, \end{aligned}$$

by Cauchy-Schwarz. Summing over t gives the first bound, and Cauchy-Schwarz in t gives $\sum_t \|e_t\|_2 \le \sqrt{T \sum_t \|e_t\|_2^2} = T\sqrt{\bar{\mathcal{L}}}$. The underbraced pairwise form is the last quantity in the chain that still reads e through the arm geometry, and Appendix E evaluates the whole chain on the deployed streams. ■

The constant is the diameter, because an $\arg\min$ is invariant to a shift common to every arm, so only the spread of the arm features converts estimation error into regret. The two differ sharply here: every $X_a = (I + L_a)^{-1}$ is a rank $|E_a|$ update of one common resolvent by Proposition 1, so the arm features cluster tightly around a shared center, and Appendix E measures $D = 0.285$ against $2\Phi_{\max} = 1.411$. Holding for every realization of the stream and for every estimator, it is the right certificate for the shipped configuration.

Corollary 1 (What the correction moves inside the certificate)**.** Run two policies on one stream, each greedy on its own estimate: the deployed one, whose played estimate is the filter output under the safeguard of Theorem 4, and the zero correction one, which plays on the anchor estimate $\hat{\theta}_t^{W_{\text{med}}}$

alone. Write $\bar{\mathcal{L}}_\omega$ and $\bar{\mathcal{L}}_0$ for the objective equation 11 read along each policy's own rollout. Their leading certificates in Proposition 2 are $D\,T\sqrt{\bar{\mathcal{L}}_\omega}$ and $D\,T\sqrt{\bar{\mathcal{L}}_0}$, so they stand in the ratio $\sqrt{\bar{\mathcal{L}}_\omega/\bar{\mathcal{L}}_0}$ at every ω and on every realization.

Proof. Proposition 2 applies to each policy along its own rollout at the same D and T, so the two leading terms are $D\,T\sqrt{\bar{\mathcal{L}}_\omega}$ and $D\,T\sqrt{\bar{\mathcal{L}}_0}$ and the common factor $D\,T$ cancels realization by realization. ■

The quantity pretraining minimizes therefore enters the certificate at the square root, on the trajectory the policy itself generates, while Theorem 4 holds both members at the same order and the same γ_0: the safeguard fixes the rate and leaves the ratio to the objective. Appendix E runs the two policies under common random numbers and reads both sides.

Corollary 2 (Arm count and sample complexity)**.** Let Assumption 1 hold.

(i) For fixed rank r and horizon T, amortized policy regret depends on pool size only through $\sqrt{\log K}$, whereas every structure blind policy (Definition 1), one reading the index reward history alone and equivariant under arm relabelling, incurs $\Omega(\Delta_0 \min\{K, T\})$ on an instance realizable inside the Friedkin-Johnsen model with gap $\Delta_0 > 0$.

(ii) Under the ρ schedule of Lemma 4, which supplies the design condition by construction, least squares on spectral features identifies the optimal arm from $m = \mathcal{O}\big(r^2 \log(K/\gamma)/\Delta_{\min}^2\big)$ observations with probability $1-\gamma$ at minimum gap $\Delta_{\min}$, against $\Omega(K)$ without cross arm transfer.

Proof of Corollary 2. Part (i). By Theorem 2, the regret of the structure aware policy scales with candidate pool cardinality K solely through the entropy factor $\sqrt{\log K}$. For the converse we lower bound the whole structure blind class.

Definition 1 (Structure blind policy)**.** A policy is *structure blind* if its action distribution at round t is a measurable function of the index reward history $h_t = (a_1, y_1, \ldots, a_{t-1}, y_{t-1})$ alone and is equivariant under relabelling of arm indices: $\pi_t(\tau(a) \mid \tau \cdot h_t) = \pi_t(a \mid h_t)$ for every permutation τ of $\mathcal{A}$. Equivalently, it never reads φ_a or ψ_a.

A hard instance inside the Friedkin-Johnsen model. Let $|V| = n$ be even and let $G = C_n$ be the unweighted cycle, so L is circulant and vertex transitive. Let s be the alternating vector $s_i = (-1)^i$, which satisfies $\mathbf{1}^\top s = 0$, $\|s\|_2^2 = n$, and $Ls = \lambda s$ with $\lambda = 4$; Assumption 1 holds with $r = 1$ and $\mathcal{S}_1 = \mathrm{span}(s)$, hence $d_{\mathrm{eff}} = 1$. Fix an odd offset m with $1 \le m < n/2$ and let arm a_k, $k = 0, \ldots, K-1$ with $K \le n$, add weight δ_k to the single pair $e_k = \{k,\, k + m \bmod n\}$. By Proposition 1,

$$\mu_{a_k} \;=\; s^\top M s - \frac{\delta_k\, g_k^2}{1 + \delta_k R_k}, \qquad g_k = b_{e_k}^\top M s, \quad R_k = b_{e_k}^\top M b_{e_k}.$$

Because $M = (I + L)^{-1}$ is circulant, $R_k = 2(M_{00} - M_{0m})$ does not depend on k. Because s is an eigenvector of L, $Ms = s/(1+\lambda)$, so $g_k = (s_k - s_{k+m})/(1+\lambda) = \pm 2/5$ and $g_k^2 = 4/25$ does not depend on k either. Setting $\delta_k = \delta_0$ for every k except one index $k^\star$ carrying $\delta_0 + \eta$ therefore makes the $K-1$ ordinary arms *exactly* tied, with the distinguished arm strictly better by

$$\Delta \;=\; \frac{g^2\,\eta}{(1+\delta_0 R)\,(1 + (\delta_0 + \eta) R)} \;=\; \Theta(\eta),$$

since $x \mapsto x/(1+xR)$ is strictly increasing. Place a uniform prior on $k^\star \in \{0, \ldots, K-1\}$.

The bound. All K arms are exchangeable under relabelling, and a structure blind policy is equivariant under relabelling, so at any round its conditional law of $k^\star$ given the history is uniform on the arms it has not yet pulled. After t rounds at most t distinct arms have been pulled, so $\Pr[N > t] \ge 1 - t/K$ where N is the first round at which $a_{k^\star}$ is played. Every round before N plays an ordinary arm and hence incurs instantaneous regret exactly Δ, so

$$\mathbb{E}\big[\mathrm{Regret}_T\big] \;\ge\; \Delta\,\mathbb{E}\big[\min\{N, T\}\big] \;\ge\; \Delta \sum_{t=0}^{\min\{K,T\}-1} \Big(1 - \frac{t}{K}\Big) \;\ge\; \frac{\Delta}{2}\min\{K, T\},$$

which is the claimed $\Omega(\Delta_0 \min\{K, T\})$ with $\Delta_0 = \Delta$. Being a Bayes bound under an explicit prior, it is also a minimax bound over the family. A structure aware learner instead reads φ_{a_k} before the first action and pays only the $\tilde{\mathcal{O}}(\sqrt{d_{\text{eff}} T \log K})$ of Theorem 2. Explore-Commit is one member of the class, and the arm pool study of Appendix F instantiates it.

Part (ii). Write expected rewards as $\mu_a = \varphi_a^\top \theta^\star$ where $\varphi_a \in \mathbb{R}^{d_{\text{eff}}}$, with bounded features $\max_a \|\varphi_a\|_2 \le L_\varphi$. Under the stated well posed design, the empirical covariance after m pulls satisfies $\hat{\Sigma}_m = \frac{1}{m}\sum_{i=1}^m \varphi_{a_i}\varphi_{a_i}^\top \succeq \lambda_0 I$ with $\lambda_0 > 0$. The least squares estimate $\hat{\theta}_m$ satisfies

$$\varphi_a^\top(\hat{\theta}_m - \theta^\star) \;=\; \frac{1}{m}\sum_{i=1}^m \varphi_a^\top \hat{\Sigma}_m^{-1}\varphi_{a_i}\,\zeta_i.$$

Conditioned on the sequence of selected actions, this error is zero mean and sub-Gaussian with variance proxy

$$\sigma_a^2 \;=\; \frac{\sigma^2}{m^2}\sum_{i=1}^m \big(\varphi_a^\top \hat{\Sigma}_m^{-1}\varphi_{a_i}\big)^2 \;=\; \frac{\sigma^2}{m}\,\varphi_a^\top \hat{\Sigma}_m^{-1}\varphi_a \;\le\; \frac{\sigma^2 L_\varphi^2}{m\lambda_0}.$$

Applying standard sub-Gaussian tail bounds alongside a union bound over all K candidate arms,

$$\Pr\Big[\max_{a\in\mathcal{A}} \big|\varphi_a^\top(\hat{\theta}_m - \theta^\star)\big| \ge \frac{\Delta_{\min}}{2}\Big] \;\le\; 2K\,\exp\Big(-\frac{m\lambda_0\Delta_{\min}^2}{8\sigma^2 L_\varphi^2}\Big) \;\le\; \gamma.$$

Solving for sample size m demonstrates that deviation is bounded by $\Delta_{\min}/2$ with probability at least $1-\gamma$ provided

$$m \;\ge\; \frac{8\sigma^2 L_\varphi^2 \log(2K/\gamma)}{\lambda_0 \Delta_{\min}^2}. \tag{17}$$

Furthermore, the trace of the empirical covariance satisfies

$$d_{\text{eff}}\lambda_0 \;\le\; \operatorname{tr}(\hat{\Sigma}_m) \;=\; \frac{1}{m}\sum_{i=1}^m \|\varphi_{a_i}\|_2^2 \;\le\; L_\varphi^2 \quad\Longrightarrow\quad \frac{L_\varphi^2}{\lambda_0} \;\ge\; d_{\text{eff}},$$

so d_{eff} is the smallest value this ratio can take, and a design that spreads the spectral features isotropically over $\mathcal{W}_r$ attains it up to an absolute constant, giving $L_\varphi^2/\lambda_0 = \Theta(d_{\text{eff}})$. Substituting into equation 17 yields sample complexity $m = \mathcal{O}(d_{\text{eff}}\log(K/\gamma)/\Delta_{\min}^2) = \mathcal{O}(r^2\log(K/\gamma)/\Delta_{\min}^2)$. When each arm is estimated independently, confidence intervals shrink at rate $\sqrt{1/m_a}$, requiring $\sum_a m_a = \Omega(K\log(1/\gamma)/\Delta_{\min}^2) = \Omega(K)$ observations, completing the proof. ■

Theorem 6 (Zero shot transfer under model error)**.** Write

$$\varepsilon_{\text{mis}} \;:=\; \sup_{t\le T}\,\max_{a\in\mathcal{A}}\; \big|\mu_{a,t} - \varphi_a^\top\theta_t\big| \tag{18}$$

for the largest departure, over the deployment, of the expected rewards from the linear encoding the policy reads.

(i) The deployment is an ε_{mis} misspecified linear bandit of dimension $d_{\text{eff}} \le r(r+1)/2$, so its regret is bounded by the guarantee that holds when the encoding is exact, Theorem 2 on a fixed instance and Theorem 4 along a drifting stream, plus an excess penalty $C\,T\,\varepsilon_{\text{mis}}$ with $C = \mathcal{O}(\sqrt{d_{\text{eff}}}\log T) = \mathcal{O}(r\log T)$. The penalty reads equation 18 alone, so it is the same whichever mechanism opens that departure.

(ii) When the encoding is built on the deployment graph and the test network's innate opinions satisfy $s' \notin \mathcal{S}_r$, the departure obeys $\varepsilon_{\text{mis}} \le 2\sqrt{|V|}\,\|q\|_2 + \|q\|_2^2$ with $q = \Pi_{\mathcal{S}_r}^{\perp} s'$, so the excess penalty scales linearly in $\|q\|_2$ as $\|q\|_2 \to 0$ and vanishes identically when $s' \in \mathcal{S}_r$.

Proof of Theorem 6. Part *(ii)* first. Decompose the test innate opinion profile into orthogonal components $s' = s_\parallel + q$, where $s_\parallel = \Pi_{\mathcal{S}_r} s' \in \mathcal{S}_r$ and $q = \Pi_{\mathcal{S}_r}^{\perp} s' \perp \mathcal{S}_r$. Expanding the expected reward functional $\mu_a(s') = s'^\top X_a s'$ around $s_\parallel$,

$$\mu_a(s') - \mu_a(s_\parallel) \;=\; (s_\parallel + q)^\top X_a (s_\parallel + q) - s_\parallel^\top X_a s_\parallel \;=\; 2\,s_\parallel^\top X_a q + q^\top X_a q. \tag{19}$$

Because $I + L_a \succeq I$ for every intervention, the operator norm satisfies $\|X_a\|_2 = \lambda_{\max}((I + L_a)^{-1}) \le 1$. Applying the Cauchy-Schwarz inequality to equation 19 yields

$$\big|\mu_a(s') - \mu_a(s_\parallel)\big| \;\le\; 2\,\|s_\parallel\|_2\,\|X_a\|_2\,\|q\|_2 + \|X_a\|_2\,\|q\|_2^2 \;\le\; 2\,\|s_\parallel\|_2\,\|q\|_2 + \|q\|_2^2.$$

Under the population normalization of Appendix A, $\|s'\|_2^2 \le |V|$, so orthogonality implies $\|s_\parallel\|_2 \le \|s'\|_2 \le \sqrt{|V|}$. The perturbation is therefore uniformly bounded across all candidate interventions,

$$\max_{a \in \mathcal{A}} \big|\mu_a(s') - \mu_a(s_\parallel)\big| \;\le\; 2\sqrt{|V|}\,\|q\|_2 + \|q\|_2^2 \;=:\; \varepsilon_{\mathrm{mis}}. \tag{20}$$

Since $s_\parallel \in \mathcal{S}_r$ makes $\mu_a(s_\parallel) = \varphi_a^\top \theta$ exactly by Proposition 5(i), the left side of equation 20 is the departure equation 18 on this deployment, which proves *(ii)*.

Part *(i)*. Fix any deployment whose departure equation 18 is at most $\varepsilon_{\mathrm{mis}}$. Expected payoffs then deviate from the linear model $\mathcal{W}_r$ the policy reads by at most $\varepsilon_{\mathrm{mis}}$ in ℓ_∞, uniformly over arms and rounds, which is the definition of an $\varepsilon_{\mathrm{mis}}$ misspecified linear bandit of dimension $d_{\mathrm{eff}} = r(r+1)/2$. By the regret analysis of misspecified linear bandits (Lattimore et al., 2020), an ℓ_∞ departure of $\varepsilon_{\mathrm{mis}}$ inflates each round's instantaneous regret by $\mathcal{O}(\sqrt{d_{\mathrm{eff}}}\log T)\,\varepsilon_{\mathrm{mis}}$ over its value under an exact encoding, so on a fixed instance, where Theorem 2 supplies that value,

$$\mathrm{Regret}_T \;\le\; \tilde{\mathcal{O}}\big(r\sqrt{T}\big) + R_{\mathrm{pre}} + C\,T\,\varepsilon_{\mathrm{mis}}, \qquad \text{where} \quad C \;=\; \mathcal{O}\big(\sqrt{d_{\mathrm{eff}}}\log T\big) \;=\; \mathcal{O}(r\log T),$$

and along a drifting stream the same per round inflation sits on top of Theorem 4. This step consults the scalar $\varepsilon_{\mathrm{mis}}$ and nothing else about the deployment, so it holds equally when the departure comes from innate opinions leaving the window, as in *(ii)*, and when it comes from an encoding built against a graph the deployment has since moved away from, with $\varepsilon_{\mathrm{mis}}$ then read off equation 18 directly. Substituting the bound of *(ii)* makes the penalty linear in $\|q\|_2$ to leading order near $\|q\|_2 = 0$, and $q = \mathbf{0}$ gives $\varepsilon_{\mathrm{mis}} = 0$, recovering exact transfer. ■

Theorem 6 is stated for the innate subspace; the spectral window U_p that the deployed tracker reads is the same statement with $\mathcal{S}_r$ replaced by $\mathrm{span}(U_p)$, and the next proposition records when that window is exact and what it does under the ambiguity of an eigendecomposition.

Proposition 5 (Exactness and basis invariance of the spectral window)**.** Let $U_p \in \mathbb{R}^{|V|\times p}$ have orthonormal columns and set $\varphi_a = \mathrm{vecsym}(U_p^\top X_a U_p)$ and $\theta_t = \mathrm{vecsym}(U_p^\top s_t s_t^\top U_p)$.

(i) If $s_t \in \mathrm{span}(U_p)$ then $\mu_{a,t} = \varphi_a^\top \theta_t$ exactly, for every arm and every round. Otherwise, writing $s_t = s_\parallel + q$ with $s_\parallel = U_p U_p^\top s_t$, the residual is exactly $\mu_{a,t} - \varphi_a^\top \theta_t = 2 s_\parallel^\top X_a q + q^\top X_a q$, the quantity Theorem 6(ii) bounds.

(ii) For every orthogonal $Q \in \mathrm{O}(p)$, replacing U_p by $U_p Q$ leaves $\varphi_a^\top \theta_t$ unchanged for every a and every t.

Proof. (i) Write $w = U_p^\top s_t$, so $s_\parallel = U_p w$ and $U_p^\top s_t s_t^\top U_p = w w^\top$. Since vecsym is a linear isometry for the Frobenius inner product,

$$\varphi_a^\top \theta_t = \big\langle U_p^\top X_a U_p,\; w w^\top \big\rangle = w^\top U_p^\top X_a U_p w = s_\parallel^\top X_a s_\parallel = \mu_a(s_\parallel).$$

If $q = \mathbf{0}$ this is $\mu_{a,t}$; in general $\mu_{a,t} - \mu_a(s_\parallel) = 2 s_\parallel^\top X_a q + q^\top X_a q$ by equation 19.

(ii) With $A = U_p^\top X_a U_p$ and $B = U_p^\top s_t s_t^\top U_p$, the same isometry gives

$$\big\langle \mathrm{vecsym}(Q^\top A Q),\, \mathrm{vecsym}(Q^\top B Q) \big\rangle = \mathrm{tr}(Q^\top A Q Q^\top B Q) = \mathrm{tr}(Q^\top A B Q) = \mathrm{tr}(AB),$$

using $Q Q^\top = I$ and the cyclic property of the trace, and the last expression is $\langle \varphi_a, \theta_t \rangle$. ■

Part (ii) settles the sign, ordering and degeneracy ambiguities of `eigh` at the level of the decision problem: rewards, gaps and the optimal arm are exactly invariant, so the bandit instance is well posed whichever basis is returned. The coordinate vector the pretrained filter reads moves with the basis, which is what the canonicalization rule of Appendix E pins and Table 11 measures.

Corollary 3 (Degenerate eigenspaces admit no transferable canonicalization)**.** Let f map a graph to an orthonormal eigenbasis of its Laplacian and be equivariant under vertex relabelling, $f(P_\sigma L P_\sigma^\top) = P_\sigma f(L)$ for every permutation matrix P_σ. If $\sigma \in \mathrm{Aut}(G)$ and $\mathcal{E}$ is an eigenspace of L on which the automorphism group fixes no nonzero vector, then no such f exists.

Proof. For $\sigma \in \mathrm{Aut}(G)$ we have $P_\sigma L P_\sigma^\top = L$, so equivariance forces $f(L) = P_\sigma f(L)$, that is every column of $f(L)$ is fixed by P_σ for every $\sigma \in \mathrm{Aut}(G)$. The columns spanning $\mathcal{E}$ would then be nonzero vectors of $\mathcal{E}$ fixed by the whole group, contradicting the hypothesis. ■

On the cycle C_n every nontrivial Laplacian eigenvalue $2 - 2\cos(2\pi k/n)$ is exactly doubly degenerate and the rotation subgroup acts on each such plane by a nontrivial rotation, which fixes no nonzero vector, so Corollary 3 applies. A rule that reads the arbitrary vertex indices instead of only relabelling invariant statistics can fix a basis, but it then returns different features on two relabelled copies of the same network, which is precisely the transfer setting it would be needed for. Appendix E reports the canonicalization rule we ship, the fraction of degenerate clusters it separates on each graph family, and the regret cost of the residual ambiguity.

Corollary 4 (Eigengap form of the transfer penalty)**.** Let the deployment window be $\mathrm{span}(U_p)$, the leading p dimensional nontrivial invariant subspace of L, with eigengap $\gamma_p := \lambda_{p+1}(L) - \lambda_p(L) > 0$, and let the test network's innate subspace $\mathcal{S}_r$, $r \le p$, be the leading r dimensional invariant subspace of a perturbed Laplacian $L + E$. If $\|E\|_2 < \gamma_p/2$ then every unit $s' \in \mathcal{S}_r$ satisfies $\|q\|_2 = \|\Pi^\perp_{\mathrm{span}(U_p)} s'\|_2 \le 2\|E\|_2/\gamma_p$, so the penalty of Theorem 6(ii) obeys

$$C\,T\,\varepsilon_{\mathrm{mis}} \;\le\; C\,T\Big(\tfrac{4\sqrt{|V|}\,\|E\|_2}{\gamma_p} + \tfrac{4\|E\|_2^2}{\gamma_p^2}\Big), \qquad C = \mathcal{O}(r \log T).$$

Proof. By Weyl's inequality $\lambda_r(L+E) \le \lambda_r(L) + \|E\|_2 \le \lambda_p(L) + \|E\|_2$, so the spectrum of $L + E$ inside $\mathcal{S}_r$ is separated from $\{\lambda_j(L)\}_{j>p}$ by at least $\gamma_p - \|E\|_2 > \gamma_p/2$. The $\sin\Theta$ theorem of Davis & Kahan (1970) gives $\|\sin\Theta(\mathcal{S}_r, \mathrm{span}(U_p))\|_2 \le \|E\|_2/(\gamma_p - \|E\|_2) \le 2\|E\|_2/\gamma_p$, and for a unit vector $\|\Pi^\perp s'\|_2$ is at most that principal sine. Substituting into $\varepsilon_{\mathrm{mis}} = 2\sqrt{|V|}\|q\|_2 + \|q\|_2^2$ gives the display. ■

Proof of Theorem 3. Part (i). Let the parameter trajectory $\theta_t^\star$ traverse a continuous rectifiable path in $\mathbb{R}^{d_{\mathrm{eff}}}$ parametrized by normalized time $u = t/T$, with total variation bounded by $\mathcal{B}$. The instantaneous regret incurred by playing a fixed intervention a at round t is $\Delta_t(a) = \mu_{a,t} - \mu_t^\star$, where $\mu_t^\star = \min_{b \in \mathcal{A}} \mu_{b,t}$. The time averaged regret of committing to a single action a over horizon T satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\big(\mu_{a,t} - \mu_t^\star\big) \;\xrightarrow[T\to\infty]{}\; \bar{g}_a(\mathcal{B}) \;:=\; \int_0^1 \big(\mu_a(u) - \mu^\star(u)\big)\,\mathrm{d}u.$$

By hypothesis no single intervention $a \in \mathcal{A}$ satisfies $\mu_a(u) = \mu^\star(u)$ for almost every $u \in [0,1]$. Since the pool $\mathcal{A}$ is finite and each map $u \mapsto \mu_a(u)$ is continuous, the minimal mean excess across all candidate interventions is strictly positive,

$$\bar{g}(\mathcal{B}) \;:=\; \min_{a \in \mathcal{A}} \bar{g}_a(\mathcal{B}) \;>\; 0.$$

Any commit once policy executes a sublinear exploration phase of length $t_0 = o(T)$, followed by an irrevocable commitment to a single arm $\hat{a} \in \mathcal{A}$ for all rounds $t \in [t_0 + 1, T]$. Its dynamic regret is lower bounded by

$$\mathrm{Regret}_T \;\ge\; \sum_{t=t_0+1}^{T}\big(\mu_{\hat{a},t} - \mu_t^\star\big) \;\ge\; T\,\bar{g}(\mathcal{B}) - o(T) \;=\; \Omega(T).$$

Part (ii). Consider a sliding window regularized least squares estimator operating over a backward window of length W. At round t, the estimator solves

$$\hat{\theta}_t^W \;=\; \arg\min_{\theta} \sum_{i=t-W+1}^{t} (\varphi_{a_i}^\top \theta - y_i)^2 + \lambda\|\theta\|_2^2.$$

Under a well posed design across sliding windows, the estimation error decomposes into statistical variance and nonstationary drift bias. What the variance divides by is the number of rounds inside the window on which the block covariance is conditioned, and not the length of the window: Lemma 4 places $m(t,W) \geq c\min\{W,\, t-n_0\}\, t^{-1/3} - 1$ of them in a window of length W ending at round t, so

$$\big\|\hat{\theta}_t^W - \theta_t^\star\big\|_2 \;\leq\; \underbrace{\mathcal{O}\Big(\sigma\sqrt{\frac{d_{\text{eff}}\log(KT)}{\lambda^\star\, m(t,W)}}\Big)}_{\text{statistical estimation error}} \;+\; \underbrace{\mathcal{O}\Big(\sum_{i=t-W+1}^{t-1}\|\theta_{i+1}^\star - \theta_i^\star\|_2\Big)}_{\text{temporal drift bias}}.$$

Because the trajectory moves at uniform speed under variation budget $\mathcal{B}$, the per round drift is bounded by $\|\theta_{i+1}^\star - \theta_i^\star\|_2 \leq \mathcal{B}/T$, implying that drift bias across a window of length W is bounded by $\mathcal{O}(\mathcal{B}W/T)$. At $W \leq t - n_0$ that count reads $m(t,W) = \Theta(cWt^{-1/3})$, and because expected rewards are linear in θ, the instantaneous regret at round t satisfies

$$\mathbb{E}[\Delta_t] \;\leq\; \tilde{\mathcal{O}}\Big(\sqrt{\frac{d_{\text{eff}}}{c\lambda^\star W}}\, t^{1/6} + \frac{\mathcal{B}W}{T}\Big). \tag{21}$$

The rounds $t \leq W + n_0$, over which the window is not yet full, together contribute $\mathcal{O}\big(\sqrt{d_{\text{eff}}/(c\lambda^\star)}\, W^{2/3}\big)$, below the first term of equation 21 summed over the horizon at every $W \leq T$. Since $\sum_{t\leq T} t^{1/6} = \Theta(T^{7/6})$, balancing the two contributions of equation 21 summed over the horizon yields optimal sliding window length

$$W^\star \;=\; \Theta\Big(\Big(\frac{d_{\text{eff}}^{1/2}\, T^{7/6}}{\sqrt{c\lambda^\star}\,\mathcal{B}}\Big)^{2/3}\Big), \tag{22}$$

which is $\Theta(T^{7/9})$ at a fixed budget. Substituting $W^\star$ yields total dynamic regret

$$\text{Regret}_T \;\leq\; \tilde{\mathcal{O}}\Big(\Big(\frac{d_{\text{eff}}}{c\lambda^\star}\Big)^{1/3}\mathcal{B}^{1/3}T^{7/9}\Big) \;+\; \tfrac{3}{2}c\,\Delta_{\max}T^{2/3},$$

the second term the cost Lemma 4 charges for the draws that condition the window covariance, carried explicitly at every budget. The density is the algorithm's to set and it sets the exponent: a schedule $\rho_t = ct^{-\alpha}$ places the windowed term at order $T^{(2+\alpha)/3}$ against a scheduled term at order $T^{1-\alpha}$, and the deployed $\alpha = \frac{1}{3}$ puts the first at $T^{7/9}$ and the second at $T^{2/3}$. Comparing the static lower bound $\Omega(T)$ with the dynamic tracking bound $\tilde{\mathcal{O}}(T^{7/9})$ confirms the $T^{2/9}$ power law separation. ■

Corollary 5 (Amortized streaming dynamic regret)**.** Consider a stream of J sequential episodes on a fixed network sharing the rank r prior, each drifting with variation budget $\mathcal{B}$ over horizon T. An amortized tracker running the ρ schedule of Lemma 4 achieves total dynamic regret bounded by $J\cdot\big(\tilde{\mathcal{O}}(\mathcal{B}^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3}\big) + J\cdot R_{\text{pre}}$, whereas an estimator that restarts at every episode and is structure blind within it (Definition 1) incurs $\Omega(J\min\{K,T\})$ by Corollary 2(i), a converse about that class. Subspace structure is shared across episodes while drift is not, so an offline model carries the structure across them and spends its capacity on the dynamics.

Proof of Corollary 5. Consider a streaming horizon comprising J sequential episodes of length T. Let each episode $j \in \{1,\dots,J\}$ traverse a dynamic parameter trajectory $\theta_{j,t}^\star$ within the shared homophily subspace $\mathcal{S}_r$, each bounded by variation budget $\mathcal{B}$.

Applying Theorem 3 conditionally on each episode j, the structure aware dynamic tracker achieves expected dynamic regret bounded by $\tilde{\mathcal{O}}(\mathcal{B}^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3} + R_{\text{pre}}$, where R_{pre} is the amortized imitation penalty from Theorem 2. By linearity of expectation across the J episodes, the cumulative stream regret satisfies

$$\mathbb{E}\big[\text{Regret}_{1:J}\big] \;=\; \sum_{j=1}^{J}\mathbb{E}\big[\text{Regret}^{(j)}\big] \;\leq\; J\cdot\Big(\tilde{\mathcal{O}}\big(\mathcal{B}^{1/3}T^{7/9}\big) + \tfrac{3}{2}c\,\Delta_{\max}T^{2/3}\Big) + J\cdot R_{\text{pre}}.$$

Conversely, consider an estimator that restarts learning at the beginning of each episode without carrying parameter estimates forward across episode boundaries. On each episode j, such an estimator

re-identifies the active parameter subspace from scratch; if it is moreover structure blind within the episode in the sense of Definition 1, that costs $C_{\mathrm{est}} = \Omega(\min\{K, T\})$ by Corollary 2(i). Because information is erased at each episode boundary, these exploration penalties compound additively across episodes:

$$\mathbb{E}\big[\mathrm{Regret}_{1:J}^{\mathrm{restart}}\big] \;\geq\; \sum_{j=1}^{J} C_{\mathrm{est}} \;=\; \Omega\big(J\, C_{\mathrm{est}}\big).$$

Restarting the search on every episode therefore carries a penalty that grows linearly with J, whereas the structure aware tracker amortizes structural inference across the whole stream, concluding the proof. ■

Theorem 3(ii) and Corollary 2(ii) are stated under a well posed design. The next lemma discharges that hypothesis by making the design part of the algorithm, so the condition holds by construction.

Lemma 4 (An anytime ρ schedule makes the design condition a consequence)**.** Let $\pi^\star$ be a G optimal design computed over the whole pool $\{\varphi_a\}_{a\in\mathcal{A}} \subset \mathbb{R}^{d_p}$, exactly the ambient space Algorithm 2 computes it in, let $\Pi_{\mathrm{eff}} \in \mathbb{R}^{d_{\mathrm{eff}}\times d_p}$ select the coordinates of the effective block that carries θ_t under Assumption 1, and write $\lambda^\star := \lambda_{\min}\big(\Pi_{\mathrm{eff}}\big(\sum_a \pi^\star(a)\varphi_a\varphi_a^\top\big)\Pi_{\mathrm{eff}}^\top\big) > 0$, read on that block alone, and $\Phi_{\max} := \max_a \|\varphi_a\|_2$. That sign is a coverage condition on the pool read on the block: the matrix inside equals $\sum_a \pi^\star(a)(\Pi_{\mathrm{eff}}\varphi_a)(\Pi_{\mathrm{eff}}\varphi_a)^\top$, so $\lambda^\star > 0$ exactly when $\{\Pi_{\mathrm{eff}}\varphi_a : \pi^\star(a) > 0\}$ spans $\mathbb{R}^{d_{\mathrm{eff}}}$, and a G optimal design supports such a set whenever the pool does, its criterion being finite only at designs nonsingular on the span of the pool. A pool whose projection reaches a proper subspace is read at the dimension of that span, which is what d_{eff} denotes throughout. Fix $c > 0$ and let the policy draw $a_t \sim \pi^\star$ independently of the history at each round where the running sum $S(t) := \sum_{n_0<s\leq t} cs^{-1/3}$, accumulated from the first round after a warmup of length n_0 exactly as Algorithm 2 accumulates it, passes an integer, a schedule of density $\rho_t = ct^{-1/3}$ that reads the round index and takes no horizon as input. Every window of length W ending at round t then meets that schedule in $m \geq c\min\{W,\, t-n_0\}\, t^{-1/3} - 1$ rounds, and

$$\Pr\Big[\lambda_{\min}\big(\Pi_{\mathrm{eff}}\hat{\Sigma}_W\Pi_{\mathrm{eff}}^\top\big) \;\geq\; \tfrac{1}{2}m\lambda^\star\Big] \;\geq\; 1 - d_{\mathrm{eff}}\exp\Big(-\frac{m\lambda^\star}{8\Phi_{\max}^2}\Big),$$

while the scheduled rounds cost at most $\frac{3}{2}c\,\Delta_{\max}T^{2/3}$ in regret over a horizon T. At $W = W^\star = \Theta(T^{7/9})$ the warmup is a fixed offset against a growing length, $\min\{W^\star, t-n_0\} \geq W^\star - n_0 = \Theta(T^{7/9})$ at every $t \geq W^\star$, so every window of that length carries $m = \Theta(T^{4/9})$ design draws and the failure probability is $d_{\mathrm{eff}}e^{-\Theta(T^{4/9})}$, so Theorem 3(ii) and Corollary 2(ii) hold unconditionally at the same rate, at one constant c held fixed across horizons.

Proof. Write $\hat{\Sigma}_W = \sum_{i\in\mathcal{W}} \varphi_{a_i}\varphi_{a_i}^\top$ and let $\mathcal{F} \subseteq \mathcal{W}$ index the scheduled rounds it contains. The schedule places $\lfloor S(t)\rfloor$ draws in the first t rounds, so, writing $t_- := \max\{t-W,\, n_0\}$ for the earliest round of the window the schedule reaches, $|\mathcal{F}| = \lfloor S(t)\rfloor - \lfloor S(t_-)\rfloor > S(t) - S(t_-) - 1$, and $S(t) - S(t_-) = \sum_{s=t_-+1}^{t} cs^{-1/3} \geq c(t-t_-)t^{-1/3} = c\min\{W, t-n_0\}\, t^{-1/3}$ because the summand decreases in s. Hence $m = |\mathcal{F}| \geq c\min\{W, t-n_0\}\, t^{-1/3} - 1$ for every window, a count fixed before the run because the schedule is deterministic. Every summand is positive semidefinite, so $\Pi_{\mathrm{eff}}\hat{\Sigma}_W\Pi_{\mathrm{eff}}^\top \succeq \sum_{i\in\mathcal{F}} Y_i$ with $Y_i = (\Pi_{\mathrm{eff}}\varphi_{a_i})(\Pi_{\mathrm{eff}}\varphi_{a_i})^\top$, and the Y_i are independent because the forced draws use fresh randomness and a fixed distribution on an index set chosen in advance. They satisfy $0 \preceq Y_i \preceq \Phi_{\max}^2 I_{d_{\mathrm{eff}}}$, since selecting coordinates contracts the norm, and $\lambda_{\min}\big(\sum_{i\in\mathcal{F}}\mathbb{E}Y_i\big) = m\lambda^\star$, the design entering through that single number, so nothing beyond $\lambda^\star > 0$ is asked of $\pi^\star$ on the block. The matrix Chernoff bound (Theorem 5.1.1 of Tropp, 2012) gives, for $\epsilon \in (0,1)$,

$$\Pr\Big[\lambda_{\min}\Big(\sum_{i\in\mathcal{F}} Y_i\Big) \leq (1-\epsilon)\, m\lambda^\star\Big] \;\leq\; d_{\mathrm{eff}}\,\exp\Big(-\frac{\epsilon^2 m\lambda^\star}{2\Phi_{\max}^2}\Big),$$

and $\epsilon = \frac{1}{2}$ yields the display. Each scheduled round is charged at most $\Delta_{\max}$ and there are $\lfloor S(T)\rfloor \leq \sum_{n_0<s\leq T} cs^{-1/3} \leq \sum_{s=1}^{T} cs^{-1/3} \leq \frac{3}{2}cT^{2/3}$ of them, so scheduled exploration contributes $\mathcal{O}(T^{2/3})$, matching the tracking term, and $m \geq c(W^\star - n_0)T^{-1/3} - 1 \to \infty$ at $W^\star = \Theta(T^{7/9})$ drives the failure probability to zero. ■

This is the schedule §5.2 deploys, at the constant c fixed once on the validation streams of Appendix G, and Appendix D reports the draws it places and the minimum eigenvalue on the effective block they produce along the streams the tables are run on.

The schedule is also what lets the rate reach the deployed filter. By equation 10 that filter emits its correction on one member of its own window grid, $\hat{\theta}_t^{\mathrm{play}} = \hat{\theta}_t^{W} + g_\omega(x_{1:t})$ at $W = W_{\mathrm{med}}$ fixed by the architecture, so the distance between the two is read off the state at no cost and can be held to a schedule the algorithm fixes in advance of the stream.

Proof of Theorem 4. Write $a_t^\star \in \arg\min_a \varphi_a^\top \theta_t$, and let $a_t \in \arg\min_a \varphi_a^\top \hat{\theta}_t^{\mathrm{play}}$ be the arm played on a greedy round. Optimality of a_t for the played estimate gives $\varphi_{a_t}^\top \hat{\theta}_t^{\mathrm{play}} \le \varphi_{a_t^\star}^\top \hat{\theta}_t^{\mathrm{play}}$, so the instantaneous regret obeys

$$\mu_{a_t,t} - \mu_{a_t^\star,t} = (\varphi_{a_t} - \varphi_{a_t^\star})^\top \theta_t \le (\varphi_{a_t} - \varphi_{a_t^\star})^\top \big(\theta_t - \hat{\theta}_t^{\mathrm{play}}\big) \le D\big\|\hat{\theta}_t^{\mathrm{play}} - \theta_t\big\|_2,$$

by Cauchy-Schwarz and the definition of the feature diameter D, exactly as in Proposition 2. The projection places $\hat{\theta}_t^{\mathrm{play}}$ within ε_t of $\hat{\theta}_t^{W}$ whatever g_ω returns, so the triangle inequality splits the estimation error into a part carried by the window and a part carried by the schedule,

$$\big\|\hat{\theta}_t^{\mathrm{play}} - \theta_t\big\|_2 \;\le\; \big\|\hat{\theta}_t^{W} - \theta_t\big\|_2 + \varepsilon_t.$$

Summing over the greedy rounds, the first term contributes $\sum_t D\|\hat{\theta}_t^W - \theta_t\|_2$, which is the quantity equation 23 bounds by $\mathcal{R}_W(T)$ at that length once the window covariance is conditioned. Lemma 4 conditions it, and does so along the realized trajectory, because the forced draws use fresh randomness at rounds fixed in advance and their law therefore does not depend on the filter; the forced rounds themselves are charged at most $\Delta_{\max}$ each and are already inside $\mathcal{R}_W(T)$ by that lemma. The second term contributes

$$D\sum_{t=1}^{T} \varepsilon_0 t^{-1/3} \;\le\; D\varepsilon_0\Big(1 + \int_1^T u^{-1/3}\,\mathrm{d}u\Big) \;=\; D\varepsilon_0\Big(\tfrac{3}{2}T^{2/3} - \tfrac{1}{2}\Big) \;\le\; \tfrac{3}{2}D\varepsilon_0 T^{2/3}.$$

Adding the two gives the display. No step reads g_ω beyond the radius the projection enforces, so the bound is uniform in ω. Summing equation 21 at the anchor $W = W_{\mathrm{med}}$ gives $\mathcal{R}_W(T) = \tilde{\mathcal{O}}\big(D\sqrt{d_{\mathrm{eff}}/(c\lambda^\star W_{\mathrm{med}})}\,T^{7/6} + D\,\mathcal{B}W_{\mathrm{med}}\big) + \frac{3}{2}c\,\Delta_{\max}T^{2/3}$, which is equation 23 at that length and which Corollary 6 reads as $\Lambda(\kappa)$ times its value at the balancing length equation 22, $\kappa = W_{\mathrm{med}}/W^\star$. The safeguard term makes the total $\Lambda(\kappa)\,\tilde{\mathcal{O}}(D\,d_{\mathrm{eff}}^{1/3}\mathcal{B}^{1/3}T^{7/9}) + \gamma_0 T^{2/3}$ at $\gamma_0 = \frac{3}{2}(c\,\Delta_{\max} + D\,\varepsilon_0)$, a constant with no window length or budget. Both forms are read at the length the architecture fixes, so neither asks a deployment to select a member against an unobserved $(\mathcal{B}, T)$. ■

The window that balances equation 21 depends on the budget $\mathcal{B}$ and the horizon T, neither of which a deployment observes, while the architecture fixes its grid once. The next corollary prices that separation exactly: it gives the rate a fixed length carries at every $(\mathcal{B}, T)$ and the factor by which a grid member departs from the balancing length.

Corollary 6 (What a fixed window length carries, and the span a grid covers)**.** Fix any window length W and run the tracker of Theorem 3(ii) at that length under the schedule of Lemma 4 and the safeguard projection of Theorem 4. Then, under the budgeted path condition of Theorem 3, at every budget $\mathcal{B} \ge 0$ and horizon T,

$$\mathrm{Regret}_T \;\le\; \tilde{\mathcal{O}}\Big(\sqrt{\tfrac{d_{\mathrm{eff}}}{c\lambda^\star W}}\;T^{7/6} \;+\; \mathcal{B}\,W\Big) \;+\; \gamma_0\,T^{2/3}, \qquad \gamma_0 \;:=\; \tfrac{3}{2}\big(c\,\Delta_{\max} + D\,\varepsilon_0\big), \tag{23}$$

the second term collecting the forced draws Lemma 4 schedules and the radius the projection allows, at the constant γ_0 of Theorem 4, which carries the two schedule constants and no window length. No quantity of the instance enters the choice of W. Writing $W = \kappa W^\star$ for the balancing length equation 22, the windowed term of equation 23 exceeds its value at $W^\star$ by exactly the factor

$$\Lambda(\kappa) \;=\; \tfrac{1}{3}\big(2\kappa^{-1/2} + \kappa\big), \qquad \Lambda(1) = 1, \tag{24}$$

and the whole of equation 23 exceeds its value at $W^\star$ by at most $\Lambda(\kappa)$, so the $T^{7/9}$ order of Theorem 3(ii) is retained at every κ bounded away from 0 and ∞, and at $\mathcal{B} = 0$, where the

balancing length leaves every finite one behind, the member retaining the whole history carries $\tilde{\mathcal{O}}(\sqrt{d_{\text{eff}}/(c\lambda^\star)}\,T^{2/3})+\gamma_0 T^{2/3}$, the whole of it at the order the schedule itself places. Consequently a geometric grid of ratio at most ϱ spanning $[W_{\min}, W_{\max}]$ holds, for every $W^\star$ in that span, a member at $\Lambda(\kappa) \le \max\{\Lambda(\varrho^{1/2}), \Lambda(\varrho^{-1/2})\}$, and a single anchor W meets every $W^\star \in [W/\varrho, \varrho W]$ at $\Lambda(\kappa) \le \max\{\Lambda(\varrho), \Lambda(\varrho^{-1})\}$. In the units of equation 23 the span $[W_{\min}, W_{\max}]$ is the budget range

$$\frac{\mathcal{B}}{T^{7/6}} \;\in\; \Big[\frac{\sqrt{d_{\text{eff}}/(c\lambda^\star)}}{2W_{\max}^{3/2}},\; \frac{\sqrt{d_{\text{eff}}/(c\lambda^\star)}}{2W_{\min}^{3/2}}\Big], \tag{25}$$

and once $\mathcal{B} \le \frac{1}{2}\sqrt{d_{\text{eff}}/(c\lambda^\star)}\,T^{-1/3}$ the balancing length reaches the horizon, where the member retaining the whole history minimizes equation 23 over every admissible length.

Proof. Summing equation 21 over the greedy rounds gives the windowed term of equation 23, its two contributions being $\sqrt{d_{\text{eff}}/(c\lambda^\star W)}\,T^{7/6}$ and $\mathcal{B}W$. Lemma 4 charges every forced draw at most $\Delta_{\max}$ and schedules at most $\frac{3}{2}cT^{2/3}$ of them, while the projection holds the played estimate within $\varepsilon_t = \varepsilon_0 t^{-1/3}$ of the window estimate, which through the Cauchy-Schwarz step of Proposition 2 adds at most $D\varepsilon_t$ to a round and $D\varepsilon_0 \sum_{t\le T} t^{-1/3} \le \frac{3}{2}D\varepsilon_0 T^{2/3}$ over the horizon. The two together are $\gamma_0 T^{2/3}$ and complete equation 23. Write the windowed term as $f(W) = AW^{-1/2} + CW$ with $A = \sqrt{d_{\text{eff}}/(c\lambda^\star)}\,T^{7/6}$ and $C = \mathcal{B}$ in the $\tilde{\mathcal{O}}$ units. Then $f'(W) = -\frac{1}{2}AW^{-3/2} + C$ vanishes at $W^\star = (A/2C)^{2/3}$, which is equation 22, and there $A(W^\star)^{-1/2} = 2^{1/3}A^{2/3}C^{1/3}$ and $CW^\star = 2^{-2/3}A^{2/3}C^{1/3}$, so $f(W^\star) = 3\cdot 2^{-2/3}A^{2/3}C^{1/3}$. At $W = \kappa W^\star$ the first term scales by $\kappa^{-1/2}$ and the second by κ, hence

$$\frac{f(\kappa W^\star)}{f(W^\star)} \;=\; \frac{2^{1/3}\kappa^{-1/2} + 2^{-2/3}\kappa}{3\cdot 2^{-2/3}} \;=\; \tfrac{1}{3}\big(2\kappa^{-1/2} + \kappa\big) \;=\; \Lambda(\kappa),$$

an identity, not a bound. Since $\Lambda'(\kappa) = \frac{1}{3}(1-\kappa^{-3/2})$, Λ decreases on $(0,1)$, increases on $(1,\infty)$ and attains its minimum $\Lambda(1) = 1$, so $\Lambda \ge 1$ on $(0,\infty)$ and it is finite at every κ there. The term $\gamma_0 T^{2/3}$ carries no W, so $W^\star$ minimizes the whole of equation 23 as well, and writing $G = \gamma_0 T^{2/3} \ge 0$,

$$\frac{\Lambda(\kappa)f(W^\star) + G}{f(W^\star) + G} \;\le\; \Lambda(\kappa),$$

since $\Lambda(\kappa) \ge 1$ makes $G \le \Lambda(\kappa)G$; the inflation of the complete bound is therefore at most that of its windowed term, and the T dependence of equation 23 at $W = \kappa W^\star$ is that of $W^\star$ for any fixed κ. For the grid claim, let $W_i < W_{i+1} \le \varrho W_i$ be consecutive members bracketing $W^\star$. If $W^\star$ lies at or below their geometric mean then $W_i/W^\star \in [\varrho^{-1/2}, 1]$, and $\Lambda(W_i/W^\star) \le \Lambda(\varrho^{-1/2})$ by monotonicity on $(0,1)$; otherwise $W_{i+1}/W^\star \in [1, \varrho^{1/2}]$ and $\Lambda(W_{i+1}/W^\star) \le \Lambda(\varrho^{1/2})$ by monotonicity on $(1,\infty)$. The anchor claim is the same monotonicity read at $\kappa \in [\varrho^{-1}, \varrho]$. For the span, $W^\star = (A/2C)^{2/3}$ gives $W^\star \ge W_{\min}$ exactly when $\mathcal{B}/T^{7/6} \le \sqrt{d_{\text{eff}}/(c\lambda^\star)}/(2W_{\min}^{3/2})$ and $W^\star \le W_{\max}$ exactly when $\mathcal{B}/T^{7/6} \ge \sqrt{d_{\text{eff}}/(c\lambda^\star)}/(2W_{\max}^{3/2})$, which is equation 25. Finally $W^\star \ge T$ reads $\sqrt{d_{\text{eff}}/(c\lambda^\star)}\,T^{7/6}/(2\mathcal{B}) \ge T^{3/2}$, that is $\mathcal{B} \le \frac{1}{2}\sqrt{d_{\text{eff}}/(c\lambda^\star)}\,T^{-1/3}$, and f decreases on $(0, W^\star)$, so on that range f is minimized over the admissible lengths $W \in (0, T]$ at $W = T$, the length the unbounded member holds. ∎

Theorem 3 and Theorem 7 below are stated on drifting instances with a prescribed variation budget and a moving optimal arm. The next lemma exhibits such instances inside the Friedkin-Johnsen model itself, with the parameter exactly rank one and the innate opinion inside the homophily subspace at every round, so both standing structural conditions hold along the path.

Lemma 5 (A drifting Friedkin-Johnsen instance at a prescribed budget)**.** Let $G = C_n$ be the unweighted cycle on an odd number $n \ge 5$ of vertices, whose smallest nontrivial Laplacian eigenvalue $\lambda_2 = 2 - 2\cos\beta$, $\beta := 2\pi/n$, carries the two dimensional eigenspace $\mathcal{S}_2 = \operatorname{span}(u_1, u_2)$ with $u_1(k) = \sqrt{2/n}\cos(k\beta)$ and $u_2(k) = \sqrt{2/n}\sin(k\beta)$. Take Assumption 1 at $r = 2$ with $\mathcal{S}_2$, write $s_\alpha = \sqrt{n}\,U_2 c(\alpha)$ for $c(\alpha) = (\cos\alpha, \sin\alpha)^\top$, and let $\mathcal{A} = \{a_0, \dots, a_{n-1}\}$, where a_k adds weight $\delta > 0$ to the single pair $e_{m_k} = \{i + m_k,\, j + m_k\}$ obtained by rotating a fixed pair $e_0 = \{i, j\}$ by $m_k \equiv \frac{n+1}{2}k$ vertices modulo n, an ordering that lists every rotation once because n is odd. Then, with $w := U_2^\top b_{e_0} \ne \mathbf{0}$, $R := b_{e_0}^\top M b_{e_0}$, $\alpha_0 := \arg(w)$ and $\gamma := \pi/n$:

(i) every candidate reward is one rigid shift of a single nonconstant profile,

$$\mu_{a_k}(\alpha) = f(\alpha - k\gamma), \qquad f(\alpha) = \frac{n}{1+\lambda_2} - \kappa\cos^2(\alpha - \alpha_0), \qquad \kappa = \frac{n\,\delta\,\|w\|_2^2}{(1+\lambda_2)^2(1+\delta R)},$$

so the minimizing arm advances by one index each time α advances by γ and traverses all of $\mathcal{A}$ over a half turn;

(ii) $\left\|\frac{d}{d\alpha}\mathrm{vecsym}(s_\alpha s_\alpha^\top)\right\|_2 = \sqrt{2}\,n$ exactly, so a sweep of angle A at uniform rate meets equation 5 with $\mathcal{B} = \sqrt{2}\,nA$, and $A > \gamma$ moves the optimal arm;

(iii) for any margin $0 < \epsilon < \kappa\sin^2\gamma$ and any batch count m_B, holding α at $\alpha_0 + \gamma/2 \pm \eta$ with $\sin 2\eta = \epsilon/(\kappa\sin\gamma)$ and drawing the sign independently on each batch leaves a_0 and a_1 the two best arms, separated by exactly ϵ in the direction the sign selects and ahead of every other arm by $\kappa\sin 2\gamma\sin(\gamma - 2\eta) > 0$, at a total variation cost of at most $2\sqrt{2}\,n\,m_B\eta$.

Proof. Let P be the permutation matrix of the rotation $k \mapsto k+1$, so $P\mathbf{e}_k = \mathbf{e}_{k+1}$ and $b_{e_m} = P^m b_{e_0}$. The cycle is vertex transitive, $PLP^\top = L$, hence $L_{a_m} = P^m L_{a_0} P^{m\top}$ and $X_{a_m} = P^m X_{a_0} P^{m\top}$. From $u_1(k+1) = \cos\beta\,u_1(k) - \sin\beta\,u_2(k)$ and $u_2(k+1) = \sin\beta\,u_1(k) + \cos\beta\,u_2(k)$ we get $P^{m\top}U_2 = U_2 R_{m\beta}$ for the plane rotation R_ϱ acting as $R_\varrho c(\alpha) = c(\alpha - \varrho)$, so

$$\mu_{a_k}(\alpha) = n\,c(\alpha)^\top R_{m_k\beta}^\top G_0 R_{m_k\beta}\,c(\alpha) = n\,c(\alpha - m_k\beta)^\top G_0\,c(\alpha - m_k\beta), \qquad G_0 := U_2^\top X_{a_0} U_2.$$

The phase ordering is what makes this pool uniform: $m_k\beta = k\pi + k\gamma$ agrees with $k\gamma$ modulo π, the period of the profile below. Sherman-Morrison gives $X_{a_0} = M - \delta M b_{e_0} b_{e_0}^\top M/(1+\delta R)$, and $\mathcal{S}_2$ is an eigenspace of L, so $MU_2 = (1+\lambda_2)^{-1}U_2$ and

$$G_0 \;=\; \frac{1}{1+\lambda_2} I_2 \;-\; \frac{\delta}{(1+\lambda_2)^2(1+\delta R)}\,ww^\top.$$

Here $w = (u_1(i) - u_1(j),\, u_2(i) - u_2(j))^\top$ vanishes only if $i \equiv j \bmod n$, which no pair satisfies, so G_0 is a strict rank one perturbation of a multiple of I_2 and f is nonconstant. Substituting $(w^\top c(\alpha))^2 = \|w\|_2^2\cos^2(\alpha - \alpha_0)$ gives the displayed f, whose minimizer over k is the shift placing $\alpha - k\gamma$ nearest α_0 modulo π, and the n phases $k\gamma$ divide the half turn evenly, proving *(i)*. Note also $\mathbf{1}^\top s_\alpha = 0$ and $\|s_\alpha\|_2^2 = n$, so the normalization of §3.1 holds at every α and $\Theta_\alpha^\star = s_\alpha s_\alpha^\top$ is exactly rank one, so the entire path lies inside the admissible parameter set.

For *(ii)*, $\dot{s}_\alpha = \sqrt{n}\,U_2 c'(\alpha)$ satisfies $\|\dot{s}_\alpha\|_2^2 = n$ and $\dot{s}_\alpha^\top s_\alpha = 0$, so $\|\dot{s}_\alpha s_\alpha^\top + s_\alpha\dot{s}_\alpha^\top\|_F^2 = 2\|s_\alpha\|_2^2\|\dot{s}_\alpha\|_2^2 = 2n^2$, and vecsym is an isometry for the Frobenius norm. The bound transfers verbatim to the spectral coordinate $\theta_\alpha = \mathrm{vecsym}(U_p^\top s_\alpha s_\alpha^\top U_p)$ for any $p \geq 2$, since $s_\alpha \in \mathrm{span}(U_p)$ makes the two representations isometric.

For *(iii)*, writing $x = \alpha - \alpha_0$ and using $\cos^2 x - \cos^2(x-\gamma) = -\sin(2x-\gamma)\sin\gamma$,

$$f(\alpha) - f(\alpha - \gamma) \;=\; \kappa\sin\gamma\,\sin(2x - \gamma),$$

which vanishes at $x = \gamma/2$, where a_0 and a_1 are tied, and equals $\pm\kappa\sin\gamma\sin 2\eta = \pm\epsilon$ at $x = \gamma/2 \pm \eta$, the margin ceiling $\kappa\sin^2\gamma$ being exactly what keeps $2\eta < \gamma$. Arm a_k enters through $\cos^2$ of the distance from x to $k\gamma$ modulo π, a distance at most $\pi/2$ on which $\cos^2$ is strictly decreasing, so the pool is ordered by phase distance and a_0, a_1 lead at both held angles. The identity $\cos^2 u - \cos^2 v = -\sin(u+v)\sin(u-v)$ at $u = \gamma/2 + \eta$ and $v = 3\gamma/2 - \eta$ puts the nearest of the remaining $n-2$ arms above the leading pair by $\kappa\sin 2\gamma\sin(\gamma - 2\eta)$, so all of them stay strictly worse throughout and the instance is the two point construction of Besbes et al. (2014) embedded in $\mathcal{A}$. Each batch boundary moves α by at most 2η, which by *(ii)* costs at most $2\sqrt{2}\,n\eta$ of variation, and there are m_B of them. ∎

The next two results characterize dynamic regret reduction, establishing the necessity of cross episode amortization and delineating the information theoretic boundary between offline amortized policies and interactive search, one a variation reduction and one a regret separation.

Proposition 6 (Amortized variation reduction)**.** Let the nonstationary parameter trajectory admit the additive decomposition $\theta_t = \bar{\theta}_t + \xi_t$, where $\bar{\theta}_t$ denotes the structured trajectory predictable from shared cross episode dynamics and ξ_t represents the stochastic residual, with total variation budgets denoted by $\mathcal{B} = \mathrm{TV}(\theta_{1:T}) = \sum_{t=1}^{T-1}\|\theta_{t+1} - \theta_t\|_2$ and $\mathcal{B}_\xi = \mathrm{TV}(\xi_{1:T}) = \sum_{t=1}^{T-1}\|\xi_{t+1} - \xi_t\|_2$,

satisfying $\mathcal{B}_\xi \le \mathcal{B}$. A reactive sliding window tracker operating on θ_t under the ρ schedule of Lemma 4 incurs dynamic regret $\tilde{\mathcal{O}}(\mathcal{B}^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3}$, as certified by Theorem 3. Conversely, an amortized tracker that subtracts the predicted drift $\hat{\bar{\theta}}_t$ and tracks solely the residual ξ_t achieves cumulative dynamic regret bounded by

$$\mathbb{E}\big[\mathrm{Regret}_T\big] \;\le\; \tilde{\mathcal{O}}\big(\mathcal{B}_\xi^{1/3}T^{7/9}\big) \;+\; \tfrac{3}{2}c\,\Delta_{\max}T^{2/3} \;+\; \mathcal{O}\big(L_\varphi E_{\mathrm{pred}}\big),$$

where $E_{\mathrm{pred}} = \sum_{t=1}^{T}\|\bar{\theta}_t - \hat{\bar{\theta}}_t\|_2$ denotes cumulative trajectory prediction error and $L_\varphi = \max_{a\in\mathcal{A}}\|\varphi_a\|_2$ is the uniform structural feature bound. Whenever the shared dynamics are learnable such that $\mathcal{B}_\xi \ll \mathcal{B}$ and $L_\varphi E_{\mathrm{pred}} \le \delta\,\mathcal{B}_\xi^{1/3}T^{7/9}$ for a small constant $\delta > 0$, both algorithms attain the rate $\tilde{\mathcal{O}}(T^{7/9})$ of Theorem 3(ii), and the amortized policy carries the strictly smaller leading constant, yielding a regret reduction of order $(\mathcal{B}^{1/3} - \mathcal{B}_\xi^{1/3})T^{7/9}$ that expands monotonically with horizon T, the schedule term entering both members alike and cancelling in that difference.

Theorem 7 (Variation budget separation in the shared drift regime)**.** Say a streaming environment is in the *shared drift regime* when the shared drift trajectory $\bar{\theta}_{1:T}$ with total variation $\mathcal{B}$ is drawn from the minimax least favourable prior of Besbes et al. (2014), realized inside the Friedkin-Johnsen model by Lemma 5*(iii)*, and remains invariant across all J consecutive episodes. Within each episode $j \in \{1,\dots,J\}$, the residual trajectory $\xi_{1:T}^{(j)}$ has total variation at most $\mathcal{B}_\xi \le \mathcal{B}$, drawn independently across episodes, and satisfies the order preservation condition

$$\sup_{1\le t\le T}\max_{a\in\mathcal{A}}\big|\varphi_a^\top \xi_t^{(j)}\big| \;\le\; \frac{\epsilon}{4},$$

where $\epsilon = \Theta((\sigma^2\mathcal{B}/T)^{1/3})$ is the per batch suboptimality margin the minimax prior induces.

(i) **Information theoretic lower bound.** Every per episode policy, whose decisions within episode j are measurable solely with respect to intra-episode history, incurs expected dynamic regret lower bounded by

$$\mathbb{E}\big[\mathrm{Regret}^{(j)}\big] \;=\; \Omega\big(\mathcal{B}^{1/3}T^{2/3}\big).$$

(ii) **Amortized upper bound.** An estimator leveraging cross episode data pooling across an initial burn-in phase of $j_0 = \mathcal{O}(K\log(\mathcal{B}/\mathcal{B}_\xi))$ episodes under round-robin exploration, where j_0 is independent of horizon T, and subsequently tracking residual variations, incurs per episode dynamic regret bounded by $\tilde{\mathcal{O}}(\mathcal{B}_\xi^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3}$, as established in Proposition 6.

The two bounds share the budget exponent $1/3$, whereas the budget entering their leading terms differs by the factor $(\mathcal{B}/\mathcal{B}_\xi)^{1/3} \ge 1$, which is strictly greater than unity whenever $\mathcal{B}_\xi < \mathcal{B}$. In this regime the global variation budget $\mathcal{B}$ is therefore the right constant for every per episode policy, and cross episode pooling replaces it with the intrinsic residual variation $\mathcal{B}_\xi$.

Alignment of the two routes. Part (ii) establishes the statistical benchmark via cross episode parameter pooling across an initial online burn-in phase. In contrast, LDT as formulated in Algorithm 2 resets its recurrent state and memory buffer at each episode boundary, having acquired the invariant shared dynamical prior entirely offline. This eliminates online burn-in latency altogether. When $\bar{\theta}_{1:T}$ is governed by dynamics drawn from pretraining distribution $\mathcal{D}$, the frozen recurrent filter g_ω predicts $\bar{\theta}_t$ from running intra-episode observations. Proposition 6 then applies to LDT on the hypothesis it states, that the total tracking error satisfy $L_\varphi E_{\mathrm{pred}} \le \delta\,\mathcal{B}_\xi^{1/3}T^{7/9}$, yielding per episode dynamic regret bounded by $\tilde{\mathcal{O}}(\mathcal{B}_\xi^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3} + R_{\mathrm{pre}}$, where R_{pre} reflects the pretraining approximation error characterized in Theorem 2. Both routes deliver the same operational benefit, replacing the total drift variation $\mathcal{B}$ with the residual variation $\mathcal{B}_\xi$, reached by offline pretraining in LDT and by online multi episode sample pooling in part (ii).

Proof of Proposition 6. Let the true parameter trajectory decompose as $\theta_t = \bar{\theta}_t + \xi_t$. Let $\hat{\bar{\theta}}_t$ denote the causal prediction of the shared component and let $e_t := \bar{\theta}_t - \hat{\bar{\theta}}_t$ denote its instantaneous prediction error. The amortized tracker forms the composite estimate $\hat{\theta}_t = \hat{\bar{\theta}}_t + \hat{\xi}_t$, where $\hat{\xi}_t$ is a sliding window regularized least squares estimate of residual process ξ_t.

Because expected reward is exactly linear in θ, the instantaneous suboptimality gap of the chosen action $a_t = \arg\min_{a\in\mathcal{A}} \varphi_a^\top \hat{\theta}_t$ decomposes into estimation error on the residual and prediction discrepancy on the shared component:

$$\begin{aligned}\mu_{a_t,t} - \mu_t^\star &= \varphi_{a_t}^\top \theta_t - \min_{a\in\mathcal{A}} \varphi_a^\top \theta_t \\ &\le 2\max_{a\in\mathcal{A}} \left|\varphi_a^\top(\hat{\theta}_t - \theta_t)\right| \\ &\le 2\max_{a\in\mathcal{A}} \left|\varphi_a^\top(\hat{\xi}_t - \xi_t)\right| + 2\max_{a\in\mathcal{A}} \left|\varphi_a^\top e_t\right|.\end{aligned}$$

We bound the two cumulative contributions separately. Rounds given to the forced design are counted apart: Lemma 4 places at most $\frac{3}{2}cT^{2/3}$ of them and each costs at most $\Delta_{\max}$.

For the first term, the residual process $\xi_{1:T}$ has total variation bounded by $\mathcal{B}_\xi$. Because the sliding window tracker observes this residual through the identical linear spectral features φ_a, the windowed estimation analysis of Theorem 3 applies verbatim with budget $\mathcal{B}$ replaced by $\mathcal{B}_\xi$. Balancing the window length against residual drift yields

$$\sum_{t=1}^{T} \mathbb{E}\Big[\max_{a\in\mathcal{A}} \left|\varphi_a^\top(\hat{\xi}_t - \xi_t)\right|\Big] = \tilde{\mathcal{O}}\big(\mathcal{B}_\xi^{1/3} T^{7/9}\big).$$

For the second term, applying the Cauchy-Schwarz inequality with uniform feature bound $\|\varphi_a\|_2 \le L_\varphi$,

$$\sum_{t=1}^{T} 2\max_{a\in\mathcal{A}} \left|\varphi_a^\top e_t\right| \le 2L_\varphi \sum_{t=1}^{T} \|e_t\|_2 = 2L_\varphi E_{\text{pred}}. \tag{26}$$

Summing the residual tracking bound, the forced draws and the predictive error bound yields cumulative regret $\tilde{\mathcal{O}}(\mathcal{B}_\xi^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3} + \mathcal{O}(L_\varphi E_{\text{pred}})$.

Setting $\bar{\theta}_t \equiv \mathbf{0}$ removes amortization, which enforces $E_{\text{pred}} = 0$ and $\mathcal{B}_\xi = \mathcal{B}$, recovering the reactive bound $\tilde{\mathcal{O}}(\mathcal{B}^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3}$ of Theorem 3. Both learners therefore scale at the identical rate $T^{7/9}$ with respect to horizon, differing strictly in their leading constants ($\mathcal{B}_\xi^{1/3}$ versus $\mathcal{B}^{1/3}$), the schedule contributing the same $\frac{3}{2}c\,\Delta_{\max}T^{2/3}$ to each. Under learnable shared dynamics where $L_\varphi E_{\text{pred}} \le \delta\,\mathcal{B}_\xi^{1/3}T^{7/9}$, the additive prediction overhead in equation 26 contributes at most $\delta\,\mathcal{B}_\xi^{1/3}T^{7/9}$, so for δ small enough that $(1+\delta)\mathcal{B}_\xi^{1/3} < \mathcal{B}^{1/3}$ the amortized learner keeps the strictly smaller constant and the regret gap of order $(\mathcal{B}^{1/3} - \mathcal{B}_\xi^{1/3})T^{7/9}$ expands with horizon T. ■

Proof of Theorem 7. Part (i). Let $\mathcal{Q}$ denote the least favourable prior over parameter trajectories that establishes the minimax lower bound of Besbes et al. (2014). Under this construction, a parameter trajectory of total variation $\mathcal{B}$ is drawn across horizon T such that the Bayes dynamic regret of any policy is lower bounded by $\Omega(\mathcal{B}^{1/3}T^{2/3})$. Lemma 5*(iii)* carries that prior inside the model. Partition the horizon into $N = T/\Delta_T$ blocks of length $\Delta_T = \Theta(\sigma^2/\epsilon^2)$, hold the innate angle at $\alpha_0 + \gamma/2 \pm \eta$ with $\sin 2\eta = \epsilon/(\kappa \sin\gamma)$, and draw the sign independently on each block, which is the lemma at $m_B = N$. Within a block the two leading interventions are separated by exactly ϵ in the direction the sign selects while every other one trails by $\kappa \sin 2\gamma \sin(\gamma - 2\eta) > 0$, so each block poses the two point decision the minimax prior poses, and $\epsilon < \kappa \sin^2\gamma$ gives $2\eta < \gamma$, which is what holds that pair on top through the excursion. The variation this spends is $2\sqrt{2}\,n\,N\eta$, at most $\frac{\pi}{\sqrt{2}}\,n\,N\epsilon/(\kappa\sin\gamma)$ since $2\eta \le \frac{\pi}{2}\sin 2\eta$ below $\pi/2$, and $N\epsilon = T\epsilon^3/\Theta(\sigma^2) = \Theta(\mathcal{B})$ at $\epsilon = \Theta((\sigma^2\mathcal{B}/T)^{1/3})$, so the realized budget is $\Theta(\mathcal{B})$ with a constant fixed by the cycle. Let the shared component $\bar{\theta}_{1:T} \sim \mathcal{Q}$ be drawn once and held invariant across all J streaming episodes. Within each episode j, the active parameter trajectory is $\theta_t^{(j)} = \bar{\theta}_t + \xi_t^{(j)}$, whose total variation satisfies

$$\mathrm{TV}\big(\theta_{1:T}^{(j)}\big) \in \big[\mathcal{B} - \mathcal{B}_\xi,\ \mathcal{B} + \mathcal{B}_\xi\big],$$

so that $\mathrm{TV}(\theta_{1:T}^{(j)}) = \Theta(\mathcal{B})$, since $\mathcal{B}_\xi \le \mathcal{B}$ with $\mathcal{B}_\xi \ll \mathcal{B}$ in the asymptotic regime.

Fix an arbitrary per episode policy π. By definition, actions executed by π at round t of episode j are measurable with respect to intra-episode history $\mathcal{H}_t^{(j)}$ alone, preventing information transfer across episode boundaries. Write Π^+ for the class of policies that additionally observe the complete residual trajectory $\xi_{1:T}^{(j)}$ at the beginning of episode j, and let

$$\mathcal{R}^+ \;:=\; \inf_{\pi^+\in\Pi^+} \mathbb{E}\big[\mathrm{Regret}^{(j)}(\pi^+)\big].$$

Any per episode π belongs to Π^+, since it may discard the side information, so $\mathbb{E}[\mathrm{Regret}^{(j)}(\pi)] \geq \mathcal{R}^+$ and it suffices to lower bound $\mathcal{R}^+$. The argument is therefore a statement about the whole residual informed class, which is what makes it uniform over π.

Conditioned on side information $\xi_{1:T}^{(j)}$, the mean reward of arm a at round t evaluates to

$$\mu_{a,t}^{(j)} \;=\; \varphi_a^\top \bar{\theta}_t \;+\; \varphi_a^\top \xi_t^{(j)}.$$

Because $\varphi_a^\top \xi_t^{(j)}$ is known to every member of Π^+, subtracting this offset is a bijection between Π^+ and the policies of the single episode Bayes problem under unknown prior $\bar{\theta} \sim \mathcal{Q}$, preserving expected regret round by round; the residual being known is what makes the reduction exact. By the order preservation condition $\sup_t \max_a |\varphi_a^\top \xi_t^{(j)}| \leq \epsilon/4$, the residual perturbation cannot invert the suboptimality gaps between candidate interventions within any temporal block:

$$(\varphi_a - \varphi_{a^\star})^\top \theta_t^{(j)} \;\geq\; (\varphi_a - \varphi_{a^\star})^\top \bar{\theta}_t \;-\; 2\max_{b\in\mathcal{A}} \big|\varphi_b^\top \xi_t^{(j)}\big| \;\geq\; \epsilon - \frac{\epsilon}{2} \;=\; \frac{\epsilon}{2}.$$

The Bayes value of the underlying decision problem is therefore preserved within an absolute constant factor, and the minimax bound of Besbes et al. (2014) for that problem gives $\mathcal{R}^+ = \Omega(\mathcal{B}^{1/3}T^{2/3})$, which lower bounds the regret of every per episode π. Summing across all J independent episodes establishes the cumulative lower bound $\Omega(J\mathcal{B}^{1/3}T^{2/3})$.

Part (ii). Under prior $\mathcal{Q}$, the shared trajectory is piecewise constant across $N = T/\Delta_T$ temporal blocks of length $\Delta_T = \Theta(\sigma^2/\epsilon^2)$, where $\epsilon = \Theta((\sigma^2\mathcal{B}/T)^{1/3})$ denotes the per batch suboptimality gap. A single episode provides only Δ_T observations per block, which is insufficient to statistically resolve gap ϵ, establishing the lower bound in Part (i).

However, because the block structure of $\bar{\theta}$ is invariant across episodes while observation noise and residuals are independent, pooling observations across the first j_0 episodes supplies $j_0\Delta_T$ total samples per block. Under round-robin exploration during this burn-in phase, each candidate intervention receives at least $m_a = \lfloor j_0\Delta_T/K \rfloor$ pulls.

Applying sub-Gaussian concentration, the empirical mean estimates separate the optimal intervention in any given block with error probability bounded by

$$\Pr\big[\text{misidentification in block } k\big] \;\leq\; \exp\Big(-\Omega\Big(\frac{j_0\Delta_T\epsilon^2}{K\sigma^2}\Big)\Big) \;=\; \exp\big(-\Omega(j_0/K)\big), \tag{27}$$

which is independent of horizon T because $\Delta_T\epsilon^2 = \Theta(\sigma^2)$ by construction of $\mathcal{Q}$. Order preservation guarantees that residuals do not invert the sign of this empirical gap.

A misidentified block incurs at most $\epsilon\Delta_T$ error, whereas a correctly identified block incurs zero error on the shared component. Summing equation 27 across all $N = T/\Delta_T$ blocks yields

$$\mathbb{E}\big[E_{\mathrm{pred}}\big] \;\leq\; \sum_{k=1}^{N} \epsilon\Delta_T \exp\big(-\Omega(j_0/K)\big) \;=\; \mathcal{O}\big(T\epsilon\, e^{-\Omega(j_0/K)}\big) \;=\; \mathcal{O}\big(\mathcal{B}^{1/3}T^{2/3}e^{-\Omega(j_0/K)}\big).$$

Setting $j_0 = CK\log(\mathcal{B}/\mathcal{B}_\xi)$ for sufficiently large absolute constant $C > 0$ drives the expected prediction error down to $L_\varphi\,\mathbb{E}[E_{\mathrm{pred}}] \leq \delta\,\mathcal{B}_\xi^{1/3}T^{7/9}$ for any target constant $\delta > 0$. Because j_0 is independent of horizon T, this error reduction operates entirely on the leading constant.

Invoking Proposition 6 for all subsequent episodes $j > j_0$ yields per episode regret bounded by $\tilde{\mathcal{O}}(\mathcal{B}_\xi^{1/3}T^{7/9}) + \frac{3}{2}c\,\Delta_{\max}T^{2/3}$. The initial burn-in phase contributes cumulative regret $\mathcal{O}(j_0\Delta_{\max}T)$, whose contribution to average stream regret vanishes as $J \to \infty$. Comparing the per episode bounds of Part (i) and Part (ii) reveals that both share the budget exponent $1/3$, whereas the budget entering their leading terms differs by the factor $(\mathcal{B}/\mathcal{B}_\xi)^{1/3} > 1$, which is the separation asserted by the theorem. ■

The subsequent results formalize the operational boundary between purely offline amortized decision making and active online exploration. They establish an exact information theoretic dichotomy, where feature degeneracy is precisely the condition under which offline observable features alone stop determining the optimal intervention and interactive rollouts become indispensable.

Definition 2 (Policy classes and feature degeneracy)**.** Let $x(I)$ denote the feature representation of instance I computable prior to running opinion dynamics. An *amortized policy* maps observable features $x(I)$ to an action selection without conducting interactive online rollouts. A *search policy* adaptively allocates interactive rollouts to evaluate candidate interventions sequentially. Writing $\mathrm{reg}_I(a) = \mu_a(I) - \min_{a' \in \mathcal{A}} \mu_{a'}(I)$ for the suboptimality gap on instance I, an instance distribution $\mathcal{D}$ is *feature degenerate with joint margin* $\Delta > 0$ if there exist two instances $I, I' \in \mathrm{supp}(\mathcal{D})$ satisfying identical observable features $x(I) = x(I')$ on which no single intervention is simultaneously near optimal:

$$\min_{a \in \mathcal{A}} \big[\mathrm{reg}_I(a) + \mathrm{reg}_{I'}(a)\big] \;\geq\; \Delta.$$

Proposition 7 (The interactive search boundary)**.** Let instance distribution $\mathcal{D}$ be feature degenerate with joint margin Δ in the sense of Definition 2, witnessed by instances I, I'. Then the following two statements hold at the pair $\{I, I'\}$ that witnesses it.

(i) **Offline performance floor.** Every amortized policy incurs an expected regret of at least $\Delta/2$ averaged across the pair $\{I, I'\}$.

(ii) **Interactive search guarantee.** For any tolerance $\nu > 0$ and confidence level $\gamma \in (0, 1)$, a search policy evaluating each candidate intervention using $m = \mathcal{O}(\sigma^2 \log(K/\gamma)/\nu^2)$ interactive rollouts and playing the empirical minimizer is ν-optimal on both instances simultaneously with probability at least $1 - \gamma$.

Social homophily as formulated in Assumption 1 keeps the innate prior inside a shared rank r subspace that the structural intervention features resolve, which is the nondegenerate regime where offline amortization operates without the informational penalty of part (i).

Proof of Proposition 7. Part (i). Any deterministic amortized policy is a measurable mapping $\pi_{\mathrm{amort}} : x(I) \mapsto a$ of the observable features alone. Because $x(I) = x(I')$ by feature degeneracy, the policy must select the identical intervention $\hat{a} = \pi_{\mathrm{amort}}(x(I))$ on both instances. By the joint margin condition, $\mathrm{reg}_I(\hat{a}) + \mathrm{reg}_{I'}(\hat{a}) \geq \Delta$. The average regret across the pair satisfies

$$\frac{1}{2}\big[\mathrm{reg}_I(\hat{a}) + \mathrm{reg}_{I'}(\hat{a})\big] \;\geq\; \frac{\Delta}{2}.$$

For a randomized amortized policy, the action is drawn from distribution $p = \pi_{\mathrm{amort}}(x(I))$, which is identical across both instances. Averaging across action draws yields

$$\frac{1}{2}\,\mathbb{E}_{a \sim p}\big[\mathrm{reg}_I(a) + \mathrm{reg}_{I'}(a)\big] \;=\; \frac{1}{2}\sum_{a \in \mathcal{A}} p(a)\big[\mathrm{reg}_I(a) + \mathrm{reg}_{I'}(a)\big] \;\geq\; \frac{1}{2}\sum_{a \in \mathcal{A}} p(a)\Delta \;=\; \frac{\Delta}{2},$$

which establishes the universal offline lower bound.

Part (ii). Fix an arbitrary instance I. For each candidate intervention $a \in \mathcal{A}$, evaluating m independent rollouts produces empirical mean estimate $\hat{\mu}_a = \frac{1}{m}\sum_{i=1}^{m} y_{a,i}$. Under conditionally σ-sub-Gaussian observation noise, the sub-Gaussian tail bound guarantees that

$$\Pr\big[|\hat{\mu}_a - \mu_a| \geq \nu/2\big] \;\leq\; 2\exp\Big(-\frac{m\nu^2}{8\sigma^2}\Big).$$

Setting rollout sample size to

$$m \;\geq\; \frac{8\sigma^2 \log(2K/\gamma)}{\nu^2}$$

ensures that the deviation probability is bounded by γ/K. Applying a union bound across all K candidate arms, the event $\mathcal{E} = \{\max_{a \in \mathcal{A}} |\hat{\mu}_a - \mu_a| < \nu/2\}$ holds with probability at least $1 - \gamma$.

Conditioned on $\mathcal{E}$, let $\hat{a} = \arg\min_{a \in \mathcal{A}} \hat{\mu}_a$ denote the empirical minimizer and let $a^\star = \arg\min_{a \in \mathcal{A}} \mu_a$ denote the true optimal intervention. Then:

$$\mu_{\hat{a}} \;\leq\; \hat{\mu}_{\hat{a}} + \frac{\nu}{2} \;\leq\; \hat{\mu}_{a^\star} + \frac{\nu}{2} \;\leq\; \mu_{a^\star} + \nu \;=\; \min_{a \in \mathcal{A}} \mu_a + \nu,$$

proving that $\hat{a}$ is ν-optimal on instance I. Because this derivation holds universally without distributional assumptions on I, it applies simultaneously to I and I'.

Summing the sampling requirement across all K candidate interventions, the total interactive exploration budget evaluates to

$$\sum_{a=1}^{K} m \;=\; K \cdot \left\lceil \frac{8\sigma^2 \log(2K/\gamma)}{\nu^2} \right\rceil \;=\; \mathcal{O}\Big(\frac{K\sigma^2 \log(K/\gamma)}{\nu^2} \Big).$$

Selecting $\nu < \Delta/2$ ensures that interactive search attains regret strictly below the offline floor $\Delta/2$, completing the proof. ■

C LIMITATIONS

Our theoretical guarantees and algorithmic formulations build on the exact linear reward representation of the Friedkin-Johnsen equilibrium paired with low rank structural homophily, demonstrating where offline amortization replaces online interactive search and what the substitution costs. The operational scope of this framework follows the setting the structural intervention literature adopts, where network topology is static through a deployment and feedback is evaluated once social diffusion reaches equilibrium, as specified in equation 4. Innate opinions are private by construction, and no public dataset pairs the networks evaluated here with per individual opinion time series, so the real network results establish transfer across measured topology under a controlled opinion process. Alternative opinion dynamics, such as the bounded confidence and trust gated updates of Hegselmann & Krause (2002), one of the many variants surveyed by Shirzadi et al. (2025), possess distinct equilibrium geometries and invite complementary theoretical study. Promising future directions include deriving closed form representations for these nonlinear diffusion families and carrying U_p itself into the tracked state on co-evolving networks. Two quantities the filter already reads, the residual scale $\hat{\sigma}_W$ and the orthogonal complement term isolated by Proposition 5, are natural online signals for switching between the spectral window and the full lift of §4.1, a selection rule we leave to future work. The two regimes carry states of size $d_p = 21$ and $d_\star$.

D ALGORITHMS AND COMPUTATIONAL COMPLEXITY

This section details the algorithmic execution and computational complexity of our structure aware moderation methods. We provide complete specifications for the single instance Bayesian sampler Lift-TS in Algorithm 1, the streaming deployment of the Learned Dynamics Tracker (LDT) in Algorithm 2, and the offline trajectory pretraining procedure for the recurrent neural filter in Algorithm 3. In addition, the learned prior policy (LPP) analyzed in Appendix F, where it serves to measure the pretraining side of Theorem 2, is a decision pretrained transformer deployed without any dedicated per instance exploration phase, evaluating $a_t \in \arg\max_a g_\phi(a \mid \{\varphi_{a'}\}, h_t)$ on the running interaction context h_t of played actions and scalar rewards.

The collection policy. Algorithm 3 collects trajectories under a behaviour policy fixed before any stream is scored, while Algorithm 2 deploys the schedule of Lemma 4.

Recursive multiwindow updates. The windowed ridge estimates in Algorithm 2 update recursively from running sufficient statistics, ensuring that per round computational latency remains strictly constant across arbitrary horizons. For each sliding window of length W, the estimator indexes active rounds $\mathcal{I}_t^W = \{\max(1, t-W), \ldots, t-1\}$ with sample count $m_W = |\mathcal{I}_t^W|$. The internal state is completely characterized by the regularized Gram matrix, the cross moment vector, and the response energy scalar,

$$G_t^W \;=\; \lambda I + \sum_{\tau \in \mathcal{I}_t^W} \varphi_{a_\tau} \varphi_{a_\tau}^\top, \qquad b_t^W \;=\; \sum_{\tau \in \mathcal{I}_t^W} \varphi_{a_\tau} y_\tau, \qquad q_t^W \;=\; \sum_{\tau \in \mathcal{I}_t^W} y_\tau^2.$$

Upon executing intervention a_t and observing feedback y_t, the state advances via the rank one update $G \mathrel{+}= \varphi_{a_t}\varphi_{a_t}^\top$, $b \mathrel{+}= \varphi_{a_t} y_t$, $q \mathrel{+}= y_t^2$, and $m \mathrel{+}= 1$. For finite windows with $t > W$, the expiring observation $(\varphi_{a_{t-W}}, y_{t-W})$ is evicted by applying the matching rank one downdate (Horn

Algorithm 1 Lift-TS, structure aware Thompson sampling on a single instance

Require: Unintervened Laplacian $L \in \mathbb{S}^{|V|}$; candidate intervention pool $\mathcal{A}$, where arm a strengthens pairs E_a by amounts $\{\delta_e\}_{e\in E_a}$; observation noise scale σ; prior scale τ; sampling scale c; horizon T

Ensure: Sequence of executed interventions $a_1, \dots, a_T$

Phase one, deterministic offline precomputation, run once per network topology.

$M \leftarrow (I + L)^{-1}$ ▷ Compute unperturbed forest resolvent matrix

for $a \in \mathcal{A}$ **do**

  $B_a \leftarrow [\, b_e \,]_{e\in E_a}$ with $b_e = \mathbf{e}_i - \mathbf{e}_j$ for candidate pair $e = \{i, j\}$

  $X_a \leftarrow M - MB_a\big(\mathrm{diag}(\delta)^{-1} + B_a^\top M B_a\big)^{-1} B_a^\top M$ ▷ Closed form resolvent evaluation via Proposition 1

  $\psi_a \leftarrow \mathrm{vecsym}(X_a)$ ▷ Symmetric vectorization scaling off diagonal entries by $\sqrt{2}$, so $\psi_a^\top \psi_b = \langle X_a, X_b \rangle$

end for

$P \leftarrow$ orthonormal basis matrix for $\mathrm{span}\{\psi_a\}_{a\in\mathcal{A}}$, with dimension $d_\star \le \min(K, |V|(|V|+1)/2)$

$\psi_a \leftarrow P^\top \psi_a$ for every $a \in \mathcal{A}$, and let $\theta_P^\star := P^\top \theta^\star \in \mathbb{R}^{d_\star}$ denote coordinates in this basis ▷ Posterior inference is restricted to this $d_\star$ dimensional subspace

Phase two, subspace Bayesian initialization.

$\mu_0 \leftarrow \mathbf{0} \in \mathbb{R}^{d_\star}, \quad \Sigma_0 \leftarrow \tau^2 I_{d_\star}, \quad C_0 \leftarrow \tau I_{d_\star}$ ▷ Initialize mean, covariance, and lower Cholesky factor $\Sigma_0 = C_0 C_0^\top$

Phase three, online interactive moderation.

for $t = 1$ to T **do**

  Draw perturbation $\epsilon_t \sim \mathcal{N}(\mathbf{0}, I_{d_\star})$ and sample parameter $\tilde\theta_t \leftarrow \mu_{t-1} + c\, C_{t-1}\epsilon_t$ ▷ $\mathcal{O}(d_\star^2)$ sampling in the geometry Lemma 2 identifies with Theorem 1

  $a_t \leftarrow \arg\min_{a\in\mathcal{A}} \psi_a^\top \tilde\theta_t$, breaking ties uniformly at random

  Deploy a_t and observe feedback $y_t = \psi_{a_t}^\top \theta_P^\star + \zeta_t$ with $\zeta_t \sim \mathcal{N}(0, \sigma^2)$

  $v_t \leftarrow \Sigma_{t-1}\psi_{a_t}, \quad \kappa_t \leftarrow \sigma^2 + \psi_{a_t}^\top v_t$ ▷ Compute scalar predictive variance

  $\mu_t \leftarrow \mu_{t-1} + \kappa_t^{-1}\big(y_t - \psi_{a_t}^\top \mu_{t-1}\big)\, v_t$ ▷ Conjugate posterior mean update

  $\Sigma_t \leftarrow \Sigma_{t-1} - \kappa_t^{-1} v_t v_t^\top$ ▷ Rank one covariance downdate in $\mathcal{O}(d_\star^2)$ time

  Refresh C_t from Σ_t via rank one Cholesky downdate, with full refactorization every $d_\star$ rounds ▷ Preserves positive definiteness and machine precision

end for

& Johnson, 2012). These rank one transitions advance each window state from $\mathcal{I}_t^W$ to $\mathcal{I}_{t+1}^W$ in $\mathcal{O}(d_p^2)$ operations, independent of window memory W and horizon t. Stacking the Gram matrices and right hand sides across all windows in the grid $\mathcal{W}$ enables active parameter estimates $\hat\theta_t^W = (G_t^W)^{-1} b_t^W$ to be evaluated through a single batched linear solve in $\mathcal{O}(|\mathcal{W}| d_p^3)$ time. The horizon is absent from every count because the sufficient statistics have fixed size, so round t costs what round 1 costs, and the history the recursion reads back into is bounded by the longest finite window, $\max\{W < \infty\} = 150$ rounds, which is the furthest back any eviction reaches.

Recovering the estimates from these moments by batched solves keeps roundoff from accumulating over long sequences at the same complexity. The residual scale $\hat\sigma_W$ follows in closed form from the same moments, without storing individual historical observations:

$$m_W\, \hat\sigma_W^2 \;=\; \sum_{\tau\in\mathcal{I}_t^W} \big(y_\tau - \varphi_{a_\tau}^\top \hat\theta_t^W\big)^2 \;=\; q_t^W - 2\,\hat\theta_t^{W\top} b_t^W + \hat\theta_t^{W\top}\big(G_t^W - \lambda I\big)\hat\theta_t^W .$$

Evaluating this uncertainty summary costs only $\mathcal{O}(d_p^2)$ additional operations per window. The sliding window baselines read that same state object, so their estimates are the ones LDT filters.

Computational complexity. All of our learners exhibit per round computational costs that are independent of network size $|V|$ at a fixed spectral truncation p, once the arm features are precomputed. The independence is from $|V|$ and not from the pool size: scoring the candidates costs $\mathcal{O}(Kd_p)$ for LDT and $\mathcal{O}(Kd_\star)$ for Lift-TS, so a pool that grows with the network carries that growth into the per round cost, and the flat rows of Table 3 hold $K = 100$ fixed while $|V|$ grows eightfold.

Algorithm 2 Deployment of the Learned Dynamics Tracker on a streaming episode

Require: Spectral features $\{\varphi_a\}_{a\in\mathcal{A}} \subset \mathbb{R}^{d_p}$ from equation 8; window grid $\mathcal{W}$ with reference window $W_{\mathrm{med}} \in \mathcal{W}$; frozen recurrent filter g_ω with state h; warmup length n_0; ridge regularizer λ; exploration constant c with the G optimal design π of Lemma 4, computed over the whole pool $\{\varphi_a\}_{a\in\mathcal{A}}$ in $\mathbb{R}^{d_p}$; safeguard radius ε_0

Ensure: Interventions $a_1, a_2, \ldots$ executed online; every rule below reads the round index alone, so no horizon is supplied as input

Episode initialization, state resets at boundaries and carries only the frozen weights ω.

for $W \in \mathcal{W}$ **do**
  $G^W \leftarrow \lambda I_{d_p}, \quad b^W \leftarrow \mathbf{0} \in \mathbb{R}^{d_p}, \quad q^W \leftarrow 0, \quad m_W \leftarrow 0$ ▷ Gram matrix, cross moment, response energy, sample count
end for
$h \leftarrow \mathbf{0}; \quad S \leftarrow 0; \quad$ history buffer $\mathcal{H} \leftarrow [\,]$ storing recent tuples $(\varphi_{a_\tau}, y_\tau)$ up to length $\max\{W \in \mathcal{W} : W < \infty\}$ ▷ S accumulates the exploration density $\rho_t = ct^{-1/3}$
for $t = 1, 2, \ldots$ **do**
  if $t \le n_0$ **then**
    $a_t \leftarrow$ an intervention drawn uniformly at random from $\mathcal{A}$ ▷ Standard warmup matched to reactive baselines
  else
    $S_- \leftarrow S; \quad S \leftarrow S + \rho_t$ with $\rho_t = c\,t^{-1/3}$ ▷ Running exploration budget of the schedule of Lemma 4
    if $\lfloor S \rfloor > \lfloor S_- \rfloor$ **then**
      $a_t \leftarrow$ an intervention drawn from the G optimal design π ▷ Forced design round, at most $\frac{3}{2}cT^{2/3}$ of them through round T
    else
      Step one, multiscale parameter estimation across the grid $\mathcal{W}$. Every window already holds $m_W = \min\{t-1, W\} \ge 1$ observations here, so each summary below is well defined.
      for $W \in \mathcal{W}$ **do**
        $\hat\theta^W \leftarrow (G^W)^{-1} b^W$ ▷ Batched linear system solve across $|\mathcal{W}|$ right hand sides
        $\hat\sigma_W \leftarrow \big[\max\big\{0,\, m_W^{-1}\big(q^W - 2\,\hat\theta^{W\top} b^W + \hat\theta^{W\top}(G^W - \lambda I)\hat\theta^W\big)\big\}\big]^{1/2}$ ▷ Exact residual variance without caching raw residuals
      end for
      Assemble multiwindow diagnostic context:
      $x_t \leftarrow \big[(\hat\theta^W)_{W\in\mathcal{W}},\, \big(m_W^{-1/2},\, \hat\sigma_W,\, \|\hat\theta^W - \hat\theta^{W_{\mathrm{med}}}\|_2\big)_{W\in\mathcal{W}}\big]$
      Step two, causal neural filtering, safeguard projection and matched greedy execution.
      $(u_t, h) \leftarrow g_\omega(x_t, h)$ ▷ Zero initialized output head gives $u_t = 0$ at initialization
      $\hat\theta_{\mathrm{play}} \leftarrow \hat\theta^{W_{\mathrm{med}}} + u_t \cdot \min\big\{1,\, \varepsilon_t / \|u_t\|_2\big\}$ with $\varepsilon_t = \varepsilon_0 t^{-1/3}$ ▷ Safeguard projection $\Pi_{\mathcal{B}(\hat\theta^{W_{\mathrm{med}}}, \varepsilon_t)}(\hat\theta^{W_{\mathrm{med}}} + u_t)$, the estimate Theorem 4 certifies
      $a_t \leftarrow \arg\min_{a\in\mathcal{A}} \varphi_a^\top \hat\theta_{\mathrm{play}}$, breaking ties uniformly at random
    end if
  end if
  Deploy a_t and observe scalar reward $y_t = \varphi_{a_t}^\top \theta_t + \zeta_t$ with $\zeta_t \sim \mathcal{N}(0, \sigma^2)$
  Step three, recursive moment updates independent of $|V|$ and T.
  for $W \in \mathcal{W}$ **do**
    $G^W \mathrel{+}= \varphi_{a_t}\varphi_{a_t}^\top, \quad b^W \mathrel{+}= \varphi_{a_t} y_t, \quad q^W \mathrel{+}= y_t^2, \quad m_W \mathrel{+}= 1$
    if $W < \infty$ and $t > W$ **then**
      Read $(\varphi_{a_{t-W}}, y_{t-W})$ from $\mathcal{H}$ and apply matching downdate:
      $G^W \mathrel{-}= \varphi_{a_{t-W}}\varphi_{a_{t-W}}^\top, \quad b^W \mathrel{-}= \varphi_{a_{t-W}} y_{t-W}, \quad q^W \mathrel{-}= y_{t-W}^2, \quad m_W \mathrel{-}= 1$
    end if
  end for
  Append current observation (φ_{a_t}, y_t) to $\mathcal{H}$ and evict pairs older than every finite window
end for

Lift-TS executes conjugate Gaussian updates in the $d_\star$ dimensional span in $\mathcal{O}(d_\star^2)$ time and draws each posterior sample from a Cholesky factor of that covariance with no subspace estimation phase,

Algorithm 3 Offline pretraining of the Learned Dynamics Tracker

Require: Streaming distribution generator $\mathcal{D}$; window grid $\mathcal{W}$; behaviour exploration rate ε; optimizer learning rate η; epoch budget N_{ep}; episode budget N_{traj}; warmup length n_0

Ensure: Trained parameter weights ω for recurrent neural filter g_ω

Initialize network weights ω with the output layer of g_ω set strictly to zero, ensuring $\hat{\theta}_{\mathrm{play}} \equiv \hat{\theta}^{W_{\mathrm{med}}}$ prior to gradient steps

The first stage, trajectory collection under a fixed behaviour policy.

for $n = 1$ to N_{traj} **do**

  Draw a graph topology, an intervention pool, and an innate opinion trajectory $s_{1:T} \sim \mathcal{D}$

  Form eigenvector basis U_p, precompute spectral features $\{\varphi_a\}$ from equation 8, and evaluate targets $\theta_t = \mathrm{vecsym}(U_p^\top s_t s_t^\top U_p)$ from equation 9

  Run the moment recursions of Algorithm 2 under a fixed behaviour policy: uniform over $\mathcal{A}$ for $t \le n_0$, then $a_t \leftarrow \arg\min_a \varphi_a^\top \hat{\theta}_t^{W_{\mathrm{med}}}$ with probability $1 - \varepsilon$ and uniform otherwise ▷ Algorithm 2 with the correction held at its zero initialization, so collection is a single pass that does not depend on ω; the schedule and the safeguard of Algorithm 2 enter at deployment, where the pathwise bound of Proposition 2 reads the estimate actually played

  Store multiwindow context sequence $x_{n_0+1:T}$ paired with parameter targets $\theta_{n_0+1:T}$ ▷ Supervision is recorded purely in parameter space, decoupled from arm identities

end for

The second stage, supervised sequence regression in parameter space.

for $e = 1$ to N_{ep} **do**

  for each minibatch of stored sequences **do**

    Execute forward causal pass of g_ω initialized at $h_{n_0} = \mathbf{0}$, yielding corrections $u_{n_0+1:T}$ ▷ Causal pass matching the recursion and the state initialization of Algorithm 2

    $\hat{\theta}_{\mathrm{play},t} \leftarrow \hat{\theta}_t^{W_{\mathrm{med}}} + u_t$ for every $t > n_0$

    Evaluate mean squared parameter error $\mathcal{L}(\omega)$ from equation 11

    Update weights ω from $\nabla_\omega \mathcal{L}(\omega)$ at learning rate η using AdamW (Loshchilov & Hutter, 2019)

  end for

  Evaluate prediction error on held out validation streams of $\mathcal{D}$ and checkpoint optimal weights

end for

Freeze weights ω; the resulting filter deploys across arbitrary network sizes and candidate pools without retraining

at $\mathcal{O}(d_\star^3)$ per round in the implementation the runtimes are measured on and at an amortized $\mathcal{O}(d_\star^2)$ under the rank one downdate of Algorithm 1, then scores candidate interventions in an $\mathcal{O}(Kd_\star)$ pass. LDT solves $|\mathcal{W}|$ windowed ridge systems of size $d_p = p(p+1)/2$, executes one causal GRU forward step, and scores candidate interventions against the updated estimate in $\mathcal{O}(|\mathcal{W}|d_p^3 + h^2 + Kd_p)$ time per round, with hidden dimension $h = 128$ and spectral dimension $d_p = 21$. Advancing a single recurrent step per round eliminates window re-encoding. The G optimal design π of Lemma 4 is solved once per candidate pool ahead of the stream, so a forced round replaces the scoring pass by one $\mathcal{O}(K)$ categorical draw. The state the recursion carries is free of $|V|$ as well, holding the projected pool at Kd_p numbers, one Gram matrix and right hand side per window at $|\mathcal{W}|d_p^2$, and the expiring pairs the finite windows downdate, bounded by the longest finite window at $150\,d_p$; the $|V| \times |V|$ resolvents are released once the projections are formed. Offline pretraining is a single investment of 669 s, comprising 3 s to generate 1500 episodes and 666 s for 60 training epochs over 431,450 interaction steps after the warmup, independent of network size and amortized across every subsequent streaming deployment. LPP requires a single transformer forward pass over candidate arm tokens and the context, with no estimator state to maintain between rounds.

In contrast, OPD-Min-ESTR executes an initial subspace estimation over the $|V| \times |V|$ matrix followed by reduced space OFUL updates scaling as $\mathcal{O}((2|V|-1)^2)$, which compounds quadratically with network size. Table 3 measures these scaling rates across an eightfold expansion in network size, focusing on $|V| \ge 128$ where algorithmic updates dominate interpreter overhead. Every structure aware learner maintains decision latency flat within 1.0%, with LDT deciding in 0.11 ms at all sizes. Meanwhile, the baseline decision time rises from 0.179 to 16.0 ms, a factor of 89, matching

the theoretical quadratic prediction with a fitted scaling exponent of $|V|^{2.15}$. At $|V| = 1024$, LDT decides $142.7\times$ faster than OPD-Min-ESTR, a decision stage ratio that the end to end accounting below completes with the offline stage and the pretraining.

Timing protocol. To ensure a rigorous and conservative runtime comparison, OPD-Min-ESTR is evaluated using an optimized rank one Sherman-Morrison update schedule for maintaining A^{-1}, $\hat{\theta}$, arm scores $\{x_a^\top \hat{\theta}\}_{a\in\mathcal{A}}$, and confidence widths $x_a^\top A^{-1} x_a$ instead of refactorizing each round, with periodic exact refreshes to control rounding error. This optimization strictly accelerates the baseline while producing bit identical action sequences and regret trajectories to the full refactorization version at $|V| \in \{32, 64\}$ and T up to 800. The unaccelerated baseline runs $4.1\times$ to $15.6\times$ slower over $|V| \in [32, 512]$. Timing against it would understate the baseline throughput by roughly an order of magnitude, so all runtime comparisons are benchmarked against this accelerated version.

Table 3: Per round cost in ms (decision plus state update) over a $J = 30$-episode stream at $T = 200$ with $K = 100$ arms on one CPU thread. Each entry is the smallest of three repetitions of that stream, applied identically to every method, and the last row reports fitted scaling exponents time $\propto |V|^{\alpha}$.

$\lvert V\rvert$	**LDT (ours)**	OPD-Min-ESTR	Speedup
128	**0.113**	0.179	$1.6\times$
256	**0.112**	0.609	$5.4\times$
512	**0.113**	2.629	$23.3\times$
1024	**0.112**	15.956	$142.7\times$
Fitted α	0.00	2.15	

Scale extrapolation and regret. Lift-TS, SW-Ridge, and Vel-Extrap stay flat to the same 1.0% tolerance as LDT because their state dimension is set by the feature construction and not by $|V|$, while OPD-Min-ESTR acts in $2|V| - 1$ dimensions, which is the separation the fitted exponents of Table 3 record. Table 4 evaluates the statistical performance achieved under this computational efficiency. Of the 24 paired bootstrap comparisons against competing methods, 21 intervals strictly exclude zero in LDT's favour. The physical reward scale contracts as networks expand: the streaming generator holds the innate opinion norm fixed, so algebraic connectivity increases, the resolvents X_a of candidate interventions move closer together, and the reward spread $\Delta_{\max}$ listed beside each row falls by three orders of magnitude across this range, which is why the absolute regret of every method falls with it. Table 5 therefore reads the same runs in units of that spread. Dividing a row by its own $\Delta_{\max}$ is an exact rescaling, so every comparison within a size is the one above, and what the second reading supplies is the comparison across sizes: over a 32 fold expansion the normalized regret of LDT rises by $1.96\times$, the smallest of the five, with the alternatives between $2.16\times$ and $2.34\times$. The four largest configurations are $2.7\times$ to $21\times$ larger than the largest graph in LDT's pretraining prior, which caps at $|V| = 48$, so this is at the same time a zero shot scale extrapolation test.

Table 4: Cumulative dynamic regret against network size over 20 paired streams of $J = 20$ episodes at $T = 200$, matching the configuration of Table 3, in objective units alongside the empirical reward spread $\Delta_{\max}$. Best per row in bold.

$\lvert V\rvert$	$\Delta_{\max}$	**LDT (ours)**	Lift-TS (ours)	SW-Ridge	Vel-Extrap	OPD-Min-ESTR
32	1.39×10^{-1}	**100.5**	135.4	113.6	140.3	231.7
64	3.45×10^{-2}	**31.9**	50.6	34.0	40.9	68.0
128	8.79×10^{-3}	**9.12**	15.11	9.53	11.25	27.53
256	1.87×10^{-3}	2.36	3.44	**2.05**	2.78	7.12
512	4.37×10^{-4}	**0.57**	0.87	0.64	0.87	1.67
1024	1.01×10^{-4}	**0.14**	0.21	0.19	0.23	0.39

Offline initialization cost. Our offline stage consists of extracting the spectral basis U_p and projecting candidate intervention resolvents onto it, while the baseline executes rank one projections into its $2|V| - 1$ dimensional subspace. As detailed in Table 6, projecting onto our invariant spectral representation is faster at every network size, executing in 0.271 s at $|V| = 1024$ against 4.058 s for

Table 5: The runs of Table 4 in units of the reward spread, $\mathrm{Regret}/\Delta_{\max}$, which counts how many maximally suboptimal rounds the regret at that size is worth. Best per row in bold.

$\lvert V\rvert$	**LDT (ours)**	Lift-TS (ours)	SW-Ridge	Vel-Extrap	OPD-Min-ESTR
32	**722**	973	816	1008	1665
64	**925**	1468	986	1188	1975
128	**1038**	1720	1085	1281	3132
256	1262	1835	**1096**	1482	3795
512	**1315**	1987	1475	1996	3828
1024	**1413**	2103	1909	2321	3858

Table 6: One time setup cost in seconds against network size, measured under the protocol of Table 3. Both methods evaluate resolvent matrices $X_a = (I + L_a)^{-1}$, listed once. The pretraining cost of LDT is 669 s, separate and incurred once globally.

$\lvert V\rvert$	Arm matrices	Spectral features, LDT	Rank one projection, OPD-Min-ESTR
128	0.041	**0.004**	0.011
256	0.237	**0.013**	0.072
512	1.483	**0.056**	0.529
1024	8.747	**0.271**	4.058

OPD-Min-ESTR, a $15.0\times$ speedup. These setup times place our structural parametrization as the faster route throughout, in offline initialization as well as in online streaming execution.

End to end cost and break even. The $142.7\times$ of Table 3 is a decision stage figure: it compares per round latencies and excludes the one time 669 s pretraining, which Table 3 does not carry, as well as the per network arm resolvents of Table 6, which both methods pay identically. Counting everything, LDT is ahead of OPD-Min-ESTR on a given network by $s_{\mathrm{off}}(\lvert V\rvert)$ seconds of offline projection plus $s_{\mathrm{on}}(\lvert V\rvert)$ milliseconds per decision round, and the global pretraining is repaid once $s_{\mathrm{off}} + 10^{-3} s_{\mathrm{on}} T_{\mathrm{tot}} \geq 669$ s across all deployments. At $\lvert V\rvert = 1024$, $s_{\mathrm{off}} = 4.058 - 0.271 = 3.787$ s and $s_{\mathrm{on}} = 15.956 - 0.112 = 15.844$ ms, so a single network repays the pretraining after 4.2×10^4 decision rounds, which is under three of the $J{=}20$ episode streams at $T{=}800$ we report, or after 177 deployments with no online decisions at all. The baseline's per round cost grows as $\lvert V\rvert^{2.15}$, so the amortization argument strengthens with network size, the regime the method targets.

Setup cost on the real networks. Table 6 measures the setup against synthetic size, and Table 7 measures the same stages on the seven real networks over the arm pool their streaming evaluations use. The candidate resolvents dominate and are the stage both methods pay alike, and the spectral representation built on top of them takes 9.564 s at $\lvert V\rvert = 4039$, 2.6% of that pool. Each X_a is reduced to its $d_p = 21$ coordinates as it is formed, so the stage holds one dense resolvent at a time and its working set stays at 838 MB where the whole pool would occupy 12.2 GB, and what the decision stage carries forward is the feature block alone, 8.2 kB at every size in the table. That block is what Table 3 reads each round, which is why the per round cost is set by K and d_p while the setup is set by $\lvert V\rvert$.

Design conditioning under the deployed policy. Theorem 3(ii) and Corollary 2(ii) are stated in the $d_{\mathrm{eff}} = r(r+1)/2 = 6$ dimensional block that carries θ_t under Assumption 1, and Lemma 4 supplies their design condition from the schedule §5.2 deploys. Table 8 measures the condition itself, replaying the benchmark streams and recording $\lambda_{\min}$ of the normalized window covariance $\hat{\Sigma}_W = (G_t^W - \lambda I)/m_W$ restricted to that block, at every post warmup round of every episode, 71,600 rounds in each arm of the comparison. The schedule widens what every window sees. It raises the median distinct arm count at all four lengths, lifts $\lambda_{\min}$ at each of them, and carries the relative threshold on 86% of rounds at $W = 150$ and on every round at the full history window, the two grid members whose memory matches the $\Theta(T^{7/9})$ Theorem 3(ii) prescribes over the reported horizons. In the ambient $d_p = 21$ coordinates the filter estimates in, the same comparison moves from 0.08 to 0.16 at $W = 150$ and from 0.10 to 0.24 at the full history window. Because the count m that Lemma 4 bounds below grows as $Wt^{-1/3}$, it is the long windows that supply it, and the

Table 7: Setup cost in seconds on the seven real networks, over the arm pool of 100 candidate interventions modifying $|V|/2$ edges each. Spectral features is the basis U_p together with the projection of the pool onto it. Working set is the additional resident memory in megabytes the stage holds.

Network	$\lvert V\rvert$	$\lvert E\rvert$	Arm matrices	Spectral features	Working set
Florentine	15	20	0.002	0.002	16
Davis	32	89	0.004	0.002	18
Karate	34	78	0.004	0.002	18
Les Misérables	77	254	0.016	0.002	17
Political blogs	1222	16,714	12.751	0.329	111
Hamsterster	2000	16,098	51.571	1.114	231
ego-Facebook	4039	88,234	367.327	9.564	838

played estimate reads all four windows at once, so the decision is anchored where the condition is measured. Proposition 2 certifies the same policy from the estimation error alone, bounding its regret by that error along the direction each decision exposes and, once that alignment is relaxed, by the feature diameter times its norm, which is why the implementation carries the ridge $\lambda = 1$ and estimates in the ambient coordinates, where $(\hat{\Sigma}_W + \lambda I)^{-1}$ is well posed at every window length.

Table 8: Conditioning of the windowed design in the effective block, on identical streams and seeds. Design draws is the mean number of scheduled draws inside the window, magnitudes below 10^{-15} are numerically zero, and the last column is the fraction of rounds with $\lambda_{\min}/\lambda_{\max} > 10^{-6}$.

Policy	Window W	Distinct arms	Design draws	$\lambda_{\min}$ median	$\lambda_{\min}$ 5th pct.	Rounds above
Schedule inactive	20	1	0.0	$< 10^{-15}$	$< 10^{-15}$	0.08
	60	2	0.0	$< 10^{-15}$	$< 10^{-15}$	0.31
	150	20	0.0	1.9×10^{-5}	$< 10^{-15}$	0.81
	∞	21	0.0	2.3×10^{-5}	5.5×10^{-6}	1.00
Deployed policy	20	**2**	0.3	$< 10^{-15}$	$< 10^{-15}$	0.08
	60	**3**	1.0	$< 10^{-15}$	$< 10^{-15}$	**0.32**
	150	**21**	1.7	$\mathbf{2.2\times10^{-5}}$	$< 10^{-15}$	**0.86**
	∞	**22**	1.7	$\mathbf{2.6\times10^{-5}}$	$\mathbf{7.2\times10^{-6}}$	1.00

What the schedule and the safeguard spend. Table 9 counts both devices along the streams every reported table is run on. A draw is placed whenever the running sum $\sum_{s\le t} cs^{-1/3}$ passes an integer, which over an episode of length T is at most $\frac{3}{2}cT^{2/3}$ draws by Lemma 4; the measured counts are 2, 3, 5, 7, 9 and 11 per episode at the six horizons against bounds of 3.2, 5.1, 6.7, 8.1, 10.7 and 12.9, so the exploration the certificate charges for is the exploration the deployment spends and it grows at the $T^{2/3}$ the lemma states. The projection is anchored at the fixed grid member $W_{\text{med}} = 60$. Consecutive members of the deployed grid $\{20, 60, 150, \infty\}$ stand at a ratio of at most 3, so by Corollary 6 the grid holds a member within a factor 1.09 of the bound at the balancing memory equation 22 for every budget and horizon placing that memory in $[20, 150]$, the anchor alone stays within 1.39 across the threefold band around it, and equation 23 is the rate a fixed length carries at every other $(\mathcal{B}, T)$. The radius $\varepsilon_t = \varepsilon_0 t^{-1/3}$ contracts as the round index grows, so it is active on 3.8% of decision rounds at $T = 100$ and on 81.3% at $T = 800$: the $\frac{3}{2}D\varepsilon_0 T^{2/3}$ term of Theorem 4 is the term the deployment pays, and the estimate that theorem certifies is the estimate the greedy decision reads. The pair $(c, \varepsilon_0) = (0.1, 2.0)$ is fixed on the validation streams of Appendix G.

E Experimental details

Every guarantee of §4 is matched here by the measurement that grades it: Theorems 5 and 1 by Table 24, Theorem 2 with Proposition 4 by Table 32, Corollary 2 by Table 30, Theorem 3 by Table 15 and Figure 3, Theorem 6 with Corollary 4 by Tables 27, 31 and 10, Lemma 4 and Theorem 4 by Tables 8 and 9, Propositions 1 and 2 with Corollary 1 by Tables 23 and 22, and Proposition 5 with

Table 9: The deployed schedule and safeguard along the streams of Table 1, over 20 paired streams of $J = 20$ episodes at $c = 0.1$ and $\varepsilon_0 = 2.0$. Design draws are per episode, beside the $\frac{3}{2}cT^{2/3}$ bound of Lemma 4. Decision rounds are the post warmup rounds the greedy rule acts on, and the last column is the fraction of them with the projection active.

T	Decision rounds	Design draws	Lemma bound	Projection active
100	30,800	2	3.2	0.038
200	70,400	3	5.1	0.287
300	109,600	5	6.7	0.491
400	148,800	7	8.1	0.616
600	228,000	9	10.7	0.748
800	307,200	11	12.9	0.813

Corollary 3 by Table 11. Each constant is evaluated on the rollouts the regret beside it is read from, so a bound and the quantity it bounds come out of one experiment.

Evaluation environments. The empirical benchmark comprises two complementary problem classes. First, for stationary structural intervention and amortization analyses, we evaluate on Erdős-Rényi random graphs (Erdős & Rényi, 1959) and real social networks. Each candidate intervention applies rank one weight adjustments to $k = |V|/2$ node pairs drawn uniformly without replacement from all $\binom{|V|}{2}$ unordered pairs, so an intervention both reweights existing edges and creates new links, following the structural link intervention formulation of Zhu et al. (2021) and Cinus et al. (2026). Innate opinions are supported on an r dimensional subspace $\mathcal{S}_r \subset \mathbf{1}^\perp$ under social homophily (full rank otherwise), mean centered following Musco et al. (2018), with observation noise scale $\sigma = 0.1$. Second, for streaming evaluations under nonstationary drift, experiments run on a fixed network topology sharing a rank r spectral prior. Within each streaming episode, latent coordinates follow a nonlinear oscillator with position dependent angular velocities drawn from a shared dynamical regime, so the instantaneous reward preserves the exact rank one Friedkin-Johnsen structure while the trajectories admit no trivial linear extrapolation. The per step angular scale is held invariant across horizons, so a longer episode is a longer excursion along the same oscillator, and because the velocity field has a stable zero that the trajectory approaches for most draws of the regime, path length grows sublinearly in T. Measuring the empirical path length of θ_t across the twenty paired streams and twenty episodes per horizon yields variation budgets 2.83, 3.59, 3.99, 4.53, 5.51 and 6.43 at $T \in \{100, 200, 300, 400, 600, 800\}$, a log-log growth of $\mathcal{B} \propto T^{0.39}$ at $R^2 = 0.98$. This is the input Theorem 3 needs: it turns the certified leading term $\mathcal{B}^{1/3}T^{7/9}$ into the concrete reference $T^{0.91}$ that the measured exponents of Table 15 are read against.

The six evaluated streaming horizons span an eightfold range ($T \in [100, 800]$) and are disjoint from the horizon set $\{150, 250, 350, 500\}$ seen during pretraining, which they bracket on both sides, so every horizon we report is one the filter has never trained on. Running each horizon as its own set of streams lets every reactive baseline take its best window at that horizon. Four of the real networks are the ones Cinus et al. (2026) evaluate on, the Florentine families network (Padgett & Ansell, 1993), Davis southern women network (Davis et al., 1941), Zachary karate club (Zachary, 1977) and Les Misérables co-appearance network (Knuth, 1993) at $|V| \in \{15, 32, 34, 77\}$, all loaded from NetworkX (Hagberg et al., 2008). To those we add three larger ones: the political blogosphere of Adamic & Glance (2005), the standard benchmark for polarization on a measured network, symmetrized and reduced to its largest connected component at $|V| = 1222$ and 16,714 edges; the Hamsterster friendship network of Kunegis (2013), reduced the same way to $|V| = 2000$ and 16,098 edges; and the ego-Facebook network of McAuley & Leskovec (2012), distributed already connected and simple at $|V| = 4039$ and 88,234 edges. All three enter the head-to-head of Table 29, and Table 12 compiles all environment and algorithmic hyperparameters.

Homophily subspace instantiations. Assumption 1 posits that the innate opinion prior is supported on an r dimensional subspace of $\mathbf{1}^\perp$. Every experiment in the paper encodes arms the same way, by the exact projection $\varphi_a = \mathrm{vecsym}(U_p^\top X_a U_p)$ of §5.2, whose width $d_p = 21$ is a function of p alone, which is what lets one checkpoint serve every network size. What the studies vary is the subspace the prior occupies, in two realizations, both with $r = 3$. In the streaming, nonstationary and arm pool evaluations it is the graph spectral basis $\mathcal{S}_r = \mathrm{span}(U_r)$, where $U_r \in \mathbb{R}^{|V| \times r}$ contains

the r leading nontrivial Laplacian eigenvectors and innate beliefs are $s = U_r c$ for a time varying coordinate vector c; the inclusion $\mathcal{S}_r \subseteq \mathrm{span}(U_p)$ then makes expected rewards exactly linear in φ_a. The transfer study of Table 31 and the generative scope sweep of Table 27 move the prior off that window deliberately, the first to a random r dimensional subspace of $\mathbf{1}^\perp$ pinned per network, the second to priors that push innate opinion mass outside $\mathrm{span}(U_p)$, and those are the instances Theorem 6 is stated for. Ridge fitting the expected rewards μ_a on the encoding over 40 held out tasks with $K = 400$ arms measures what it retains in each case. Under the spectral instantiation the projection is exact to solver precision, with a 95th percentile relative residual of 3.0×10^{-11}, 1.4×10^{-9} and 2.5×10^{-7} at $|V| = 16, 32, 64$, which is the inclusion $\mathcal{S}_r \subseteq \mathrm{span}(U_p)$ of Theorem 2 holding numerically; under the random subspace instantiation the inclusion fails and the same projection leaves a median residual of 0.40, 0.67 and 0.92 over those sizes, the regime Theorem 6 prices and Table 31 measures. Both satisfy Assumption 1, and the proofs of Theorem 2 and Corollary 2 use only the support condition $s \in \mathcal{S}_r$, never the identity of the subspace.

Spectral truncation and the eigengap. Of the two spectral quantities only p is a deployment choice. The prior rank r describes the instance and enters no algorithm: LDT reads $\varphi_a \in \mathbb{R}^{d_p}$, and the single frozen filter at $p = 6$ covers every rank the benchmark spans, which Table 26 varies from $r = 2$ to $r = 5$. What a practitioner sets is p, at the first sizeable eigengap above the ranks a deployment expects, with Theorem 6 pricing a window that turns out too narrow. Corollary 4 expresses the transfer penalty of Theorem 6 through the Laplacian eigengap $\gamma_p = \lambda_{p+1} - \lambda_p$ at the truncation, so Table 10 reports that gap at $p = 6$ for every graph the paper evaluates, together with the largest coincident eigenvalue cluster inside the retained block. The gaps span two orders of magnitude, from 0.95 on the Florentine families network to 0.016 on Hamsterster, and on Les Misérables the retained block ends inside an exactly degenerate cluster, $\lambda_6 = \lambda_7 = 1$, the eigenvalue at which star like neighbourhoods pile up, so $\gamma_6 = 0$. That case is why the paper proves Proposition 5(ii) in place of assuming a gap: a zero gap leaves the coordinates ambiguous while the instance stands, since the reward $\varphi_a^\top \theta_t$ is exactly invariant to which basis of the degenerate eigenspace is returned, leaving the bandit instance, its gaps and its optimal arm unchanged and moving only the vector the filter reads. Table 29 accordingly shows the transfer to Les Misérables behaving like the transfer to the graphs with wide gaps, and Table 11 prices what the residual coordinate ambiguity costs.

Table 10: Laplacian eigengap $\gamma_p = \lambda_{p+1} - \lambda_p$ at the deployed truncation $p = 6$, over the nontrivial spectrum, with the largest group of coincident eigenvalues inside the retained block. The Erdős-Rényi entry is the median over the twenty benchmark streams.

Graph	$\lvert V\rvert$	γ_p at $p = 6$	Largest cluster in U_p
Erdős-Rényi (benchmark)	32	0.296	1
Cycle C_{32}	32	0.249	2
Florentine families	15	0.952	1
Davis southern women	32	0.093	1
Zachary karate club	34	0.585	1
Les Misérables	77	0.000	2
Political blogs	1222	0.066	1
Hamsterster	2000	0.016	1
ego-Facebook	4039	0.515	1

Canonicalization of the spectral basis. An eigensolver pins U_p only up to a column sign, and inside a coincident cluster only up to a rotation of the eigenspace. The rule we ship resolves what can be resolved from the graph alone: cluster the retained eigenvalues at tolerance 10^{-6}; inside each cluster of size greater than one, rotate to the eigenbasis of the degree form $U_c^\top \mathrm{diag}(\deg) U_c$ restricted to that eigenspace, taken in ascending order, falling back to $\mathrm{diag}(A^3)$ and $\mathrm{diag}(A^4)$ when the degree form is itself a multiple of the identity there; then make the largest magnitude entry of every column positive with the lowest index breaking ties. Replacing U_c by $U_c Q$ sends the degree form to $Q^\top A Q$, whose eigenvectors are $Q^\top V$, so $U_c Q\, Q^\top V = U_c V$ is recovered whatever Q was, and the output is a function of the graph. Table 11 replays identical streams with identical seeds under bases differing only by an element of the ambiguity group, on the benchmark Erdős-Rényi graph, whose spectrum is simple so only the 2^p sign group is live, and on the cycle C_{32}, vertex transitive with nontrivial eigenvalues $2 - 2\cos(2\pi k/n)$ exactly doubly degenerate, so three full rotations are live at $p = 6$

on top of the signs. On the Erdős-Rényi graph the rule is exact, reproducing one basis from every sign draw to machine precision. On the cycle the 60 degenerate clusters it meets are left to the sign convention, which is exactly the content of Corollary 3: the rotation group acts on each plane without a fixed direction, so no relabelling equivariant rule separates them. What the theory leaves open the measurement finds to be small. Sign draws move mean regret by $+0.98\%$ and $+0.11\%$ on the two families, and the rotations that are live on the cycle move it by $+4.98\%$, while fixing the basis by the rule lands at $+3.91\%$ and -0.21%, all against per stream spreads of 9.1% and up to 13.4%. Canonicalization therefore buys a reproducible feature map with a shift inside the spread the ambiguity itself carries, and the rule runs once per graph ahead of the stream.

Table 11: Regret under the spectral basis ambiguity over 20 paired streams at $T = 200$, with 6 draws per stream for the randomized conditions. All conditions replay identical tasks, seeds and noise, and relative columns are means over streams of the per stream change against the baseline basis.

Graph	Basis condition	Regret	Rel.	Mean |rel.|	Spread per stream
Erdős-Rényi simple spectrum	Baseline	5.02 ± 0.30			
	Random signs	5.09 ± 0.32	$+0.98\%$	2.66%	9.1%
	Canonicalized	5.21 ± 0.31	$+3.91\%$	4.41%	
Cycle C_{32} three degenerate pairs	Baseline	13.75 ± 0.38			
	Random signs	13.75 ± 0.37	$+0.11\%$	3.14%	11.6%
	Signs and rotations	14.42 ± 0.40	$+4.98\%$	6.01%	13.4%
	Canonicalized	13.65 ± 0.31	-0.21%	6.41%	

Tracker architecture and deployment. At each interaction step, LDT computes windowed ridge estimates of the spectral parameter $\theta_t \in \mathbb{R}^{p(p+1)/2}$ ($d_p = 21$ for $p = 6$) across sliding windows $W \in \{20, 60, 150, \infty\}$ with regularization $\lambda = 1$. These estimates are concatenated with three uncertainty scalars per window, the effective sample count $m_W^{-1/2}$, the residual standard deviation $\hat{\sigma}_W$ and the divergence from the reference estimate $\|\hat{\theta}^W - \hat{\theta}^{W_{\mathrm{med}}}\|_2$, in a deterministic order: the first $4 \times 21 = 84$ coordinates hold the multiscale ridge estimates by window and the remaining $4 \times 3 = 12$ hold the uncertainty summaries, for a 96 dimensional diagnostic vector. Those scalars are what let the filter separate nonstationary parameter motion from observation noise across noise environments. A single layer GRU g_ω (hidden dimension 128, dropout 0.1, equipped with a projection head consisting of LayerNorm, Linear, ReLU, Dropout, and a zero initialized linear output layer) maps this representation to an inductive residual correction on the medium reference window estimate ($W_{\mathrm{med}} = 60$), predicting parameter $\hat{\theta}_{\mathrm{play}}$. Because the output projection is initialized to zero, LDT coincides exactly with the analytical reference tracker at initialization, so every parameter it learns is measured against that reference.

Because the recurrent architecture is causal, training over complete trajectories with per step regression targets and deploying online by incremental hidden state updates compute the identical function, at constant per round latency over unbounded contexts. Pretraining minimizes mean squared error against true parameter targets θ_t collected under ε-greedy exploration ($\varepsilon = 0.1$). On held out validation streams, the frozen checkpoint reduces parameter prediction error by 21.3% relative to the reference window estimator. On the streaming benchmark, it reduces parameter estimation error by 10.2% over the uncorrected reference estimator and by 8.9% over the oracle window tracker (Table 21), each of the latter two measured on the data the deployed policy collects for itself. Online decision making matches the protocol of the reactive baselines, a fixed warmup of $\max\{6, d_p\} = 21$ random pulls followed by greedy execution $\arg\min_a \varphi_a^\top \hat{\theta}_{\mathrm{play}}$, so the head-to-head isolates the learned dynamics and those baselines additionally receive their oracle window. That warmup is the first round at which a design in $d_p = 21$ dimensions can have full rank, and pretraining uses the same count, so every history length the filter meets at deployment also appeared in training, and the recurrent state carries the rest of the stream forward.

Arm space capacity control. To isolate the inductive benefits of our parameter space formulation from raw model capacity, we benchmark against an in context transformer operating directly in arm space (4 layers, 8 attention heads, $d_{\mathrm{model}} = 256$, containing 3,307,778 parameters). It is permutation invariant with respect to candidate interventions, observing each arm through its spectral features φ_a

Table 12: Experimental parameters and hyperparameters across evaluation domains. Parenthetical notes indicate parameter variations specific to individual studies.

Component	Parameter	Values
Graphs	Topologies	Erdős-Rényi (edge probability 0.2); real networks (Florentine, Davis, Karate, Les Misérables, political blogs, Hamsterster, ego-Facebook)
	Network size $\|V\|$	Synthetic $\{8, 16, 32\}$ (structure), 32 (streaming), $\{40, 80, 160\}$ (support recovery), $\{16, 32, 64\}$ (robustness), 32–1024 (computational scaling); real 15–4039
Opinions	Prior rank	Rank r ($r = 3$) under homophily (spectral or random subspace); full rank otherwise; polarized transformation $\text{sign}(s_i)\|s_i\|^{1/3}$ (robustness sweep)
	Noise scale σ	0.1 (nominal); swept across $\{0.05, 0.1, 0.2, 1.0\}$ (robustness sweep)
Arms	Topology modification	Rank one weight increments on $k = \|V\|/2$ pairs drawn uniformly from unordered pairs, increments $\delta_e \sim \mathcal{U}[0.5, 1.5]$
	Pool cardinality K	100 (nominal benchmark); $\{100, 200, 400, 800, 1600\}$ (cardinality study); $\{50, 200\}$ (robustness sweep)
Bandit	Horizon T	3000 (structural verification); $\{500, \ldots, 4000\}$ (drift tracking); $\{100, \ldots, 800\}$ per episode (streaming benchmark)
	Stream configuration	20 streams $\times$ J episodes, with $J = 20$
	Lift-TS parameters	Prior variance scale $\tau = 1$, warmup horizon $n_0 = 6$, posterior sampling scale $c = 0.5$ on the streaming benchmark and $c = 0.6$ in the rate study of Table 24
	Reactive trackers	Oracle window selected at each horizon from $\{15, 30, 60, 120, 200, 300, 400, 600, 800\}$ bounded by T, with T itself a candidate
Training	Optimization	AdamW (Loshchilov & Hutter, 2019), learning rate 3×10^{-4}, weight decay 10^{-4}, gradient norm clipping 1.0, cosine annealing schedule without warmup
	Training epochs	LDT 60 epochs (batch size 32); LPP at most 40 epochs (batch size 128); arm space transformer at most 30 epochs (batch size 128), checkpointed via early stopping
	LDT window grid	$\mathcal{W} = \{20, 60, 150, \infty\}$ with reference window $W_{\text{med}} = 60$; unbounded recurrent execution
	LDT training prior	1500 synthetic episodes on Erdős-Rényi graphs; $\|V\| \in \{16, 24, 32, 48\}$, $r \in \{2, 3, 4\}$, $K \sim \mathcal{U}\{60, \ldots, 239\}$, angular drift rate in $[0.3, 2.5]$, σ log-uniform on $[0.03, 0.5]$, $T \in \{150, 250, 350, 500\}$; validated on 100 held out episodes

Table 13: Arm space capacity control in three forms, across 20 paired streams and horizons $T \in \{200, 400\}$. Columns report cumulative dynamic regret and mean polarization-disagreement index, with paired differences relative to LDT in 95% bootstrap intervals, positive values favouring LDT.

	$T = 200$			$T = 400$		
Architecture variant	Regret	Index	Δ vs. LDT	Regret	Index	Δ vs. LDT
LDT (ours)	**100.5**	**0.224**	—	**160.5**	**0.219**	—
Arm space transformer	125.6	0.231	+25.1 [18.7, 31.4]	247.3	0.230	+86.8 [71.7, 101.5]
forgetting gate held off	125.3	0.230	+24.9 [18.4, 31.2]	247.0	0.230	+86.5 [71.5, 101.0]
current step target	116.4	0.228	+15.9 [10.1, 21.6]	227.8	0.227	+67.3 [54.0, 80.8]

alongside a per arm sufficient statistic channel, with an exponential forgetting gate on observation age, parameterized by a learned decay rate and age embeddings, and it is pretrained by masked cross entropy to predict the one-step-ahead optimal arm on the streaming distribution, drawing its own regimes so that they reach the deployment size and bracket its drift rate (sizes $\{8, 16, 32\}$, drift in $[0.7, 1.5]$). Table 13 evaluates it in three forms, the pretrained policy, the same weights with forgetting held off at inference, and a second pretraining that targets the instantaneous optimal arm.

Table 14: Ablation of LDT components on the benchmark of Table 1, added one at a time. Lower is better, best per column in bold. The row with no component is SW-Ridge at its oracle window.

Multiscale estimation	Uncertainty summaries	Learned dynamics	$T = 200$		$T = 400$	
			Regret	Index	Regret	Index
			113.6	0.228	191.7	0.223
		✓	107.1	0.226	185.1	0.222
✓		✓	104.6	0.225	169.4	0.220
✓	✓	✓	**100.5**	**0.224**	**160.5**	**0.219**

The parameter space filter g_ω tracks a compact 21 dimensional dynamical state in 106,261 parameters against the transformer's $31\times$ larger budget, and in all three forms the transformer records higher regret, its strongest trailing LDT by $+15.9$ at $T = 200$ and $+67.3$ at $T = 400$ with every paired interval excluding zero. The comparison holds the streams and the arm features fixed and gives both learners an offline stage on the streaming family, the control drawing 6000 episodes against LDT's 1500, which rules out parameter count or offline budget as the explanation of the ordering; what it varies is our design as a whole, the projection onto parameter coordinates, the multiscale preprocessing and the zero initialized residual regression.

Window grid and residual design. The regret analysis of Theorem 3 prescribes an effective memory of $W = \Theta\big((d_{\text{eff}}^{1/2} T^{7/6}/\mathcal{B})^{2/3}\big)$. Because the leading constant depends on latent parameters and the variation budget is unknown online, a single fixed window cannot cover the horizons and drift speeds deployment presents. The multiscale grid is the constructive answer: four windowed estimators and their uncertainty summaries let the recurrent filter synthesize an effective memory from the stream itself, with no online hyperparameter search. The unbounded member of that grid is the full history estimate, which is the estimate Lift-TS carries in these same coordinates.

The inductive residual is anchored to the fixed grid member $W_{\text{med}} = 60$, a window the reactive baselines select from at their oracle setting. Table 14 quantifies the empirical contribution of the window grid and uncertainty summaries. The three ablation variants are trained under the identical recipe, one with the window grid collapsed to $W_{\text{med}} = 60$, one with uncertainty summaries omitted, and one with both inputs removed. In paired bootstrap tests, collapsing the grid incurs regret increases of $+8.7$ $[5.3, 12.2]$ at $T = 200$ and $+24.6$ $[16.8, 32.9]$ at $T = 400$; withholding uncertainty summaries adds $+4.1$ $[1.7, 6.6]$ and $+9.0$ $[4.5, 14.0]$; removing both inputs adds $+6.6$ $[3.3, 9.8]$ and $+24.6$ $[16.0, 33.6]$; and removing all three, which leaves SW-Ridge at its oracle window, yields $+13.1$ $[5.6, 20.7]$ and $+31.2$ $[16.5, 45.9]$. Every interval strictly excludes zero, with performance gaps expanding at the longer horizon. Among the single component removals the multiscale grid carries the largest contribution at both horizons, confirming the value of multiscale memory.

Baseline reproduction and fairness. Every evaluation of OPD-Min-ESTR runs the authors' released code without modification. Under their default benchmark configuration (polarization polynomial order 3, horizon $T = 10^4$, first stage exploration length $T_1 = \sqrt{T}$), it attains cumulative regret of 409 at $|V| = 8$ and 1164 at $|V| = 16$, replicating its documented advantages over unstructured OFUL baselines, and reproducing their stochastic block model environment (Holland et al., 1983) with the official code yields optimal objective values of approximately 2.6 at $|V| = 16$ and 3.0 at $|V| = 32$ against their published 2.64 and 3.05. Our reward computation therefore coincides with the published forest matrix objective $s^\top (I + L_a)^{-1} s = \langle X_a, ss^\top \rangle$.

In comparative evaluations we reuse its exact building blocks, subspace estimation, rank one projection and reduced space OFUL (Abbasi-Yadkori et al., 2011), and score every method on cumulative dynamic regret against the per step optimum. OPD-Min-ESTR is reported in its strongest configuration, oracle initialization with the $t = 0$ subspace and minimal exploration $T_1 = 1$, alongside per episode restart and within episode re-estimation variants. The reactive baselines are SW-Ridge (Garivier & Moulines, 2011; Cheung et al., 2019; Zhao et al., 2020), the standard sliding window ridge estimator, and Vel-Extrap, a velocity extrapolator we construct as a stronger anticipatory control, which fits the drift velocity from two half windows and extrapolates one step ahead. Vel-Extrap realizes forward state tracking under a dynamical model (Hall & Willett, 2013) and smooth trend

extrapolation (Jia et al., 2023), capturing what a linear velocity assumption recovers on oscillating trajectories, and every reactive baseline receives an oracle window chosen from a candidate grid. All algorithms share identical streams, seeds, metrics and scalar feedback, intervention features remain opinion free, and every comparison reports 95% paired bootstrap intervals. Each interval resamples the 20 stream level paired differences with replacement 5000 times, so the resampling unit is the independent stream and every stream enters as a single value already averaged over its J episodes. Every estimator resets internal statistics at episode boundaries, so no algorithm carries inter episode state unless designed to, as LDT's frozen filter and the pooled subspace variant of OPD-Min-ESTR are. Every stream carries $J = 20$ episodes.

Horizon scaling exponents. Table 15 reports cumulative dynamic regret at the longest horizon ($T = 800$) alongside empirical scaling exponents α in $\mathrm{Regret}_T \propto T^{\alpha}$, fitted via log-log least squares over the six evaluation horizons with 95% bootstrap intervals over 20 streams. For the empirical variation budget $\mathcal{B} \propto T^{0.39}$, the theoretical tracking rate certified by Theorem 3 is $\frac{1}{3}(0.39) + \frac{7}{9} = 0.91$. LDT attains an empirical exponent of 0.67 [.63, .72], lying well within this theoretical certificate while attaining the lowest cumulative regret across all horizons, and Table 23 reads the same deployed policy through Proposition 2 alone at 0.77 [.73, .81] on these streams, inside it as well. Vel-Extrap and Lift-TS share that exponent at larger constants, and SW-Ridge sits at 0.76. OPD-Min-ESTR is near linear at 0.95 [.94, .96], above the certificate, which is the penalty Corollary 5 attaches to rebuilding an estimate at every episode boundary.

Evaluating paired exponent differences on identical stream resamples places SW-Ridge above LDT by 0.089 $[0.059, 0.125]$ and OPD-Min-ESTR above LDT by 0.281 $[0.239, 0.327]$. The eightfold horizon span provides sufficient dynamic range to resolve these exponents with high confidence. The candidate window grid is truncated at each horizon to windows satisfying $W \leq T$ alongside T itself, ensuring that unbounded windowing remains available at every horizon. SW-Ridge selects interior windows across all horizons (60 at $T = 100$, 120 through $T = 400$, and 200 at $T = 600, 800$), so its exponent reflects the optimal effective memory. Vel-Extrap selects the longest available window at four horizons and interior windows at the remaining two (300 at $T = 400$ and 400 at $T = 800$).

Table 15: Cumulative dynamic regret at $T = 800$ alongside empirical horizon scaling exponents α in $\mathrm{Regret}_T \propto T^{\alpha}$, accompanied by 95% bootstrap intervals over 20 paired streams. Lowest regret in bold. All methods share identical streams and seeds.

Method	**LDT (ours)**	Lift-TS (ours)	SW-Ridge	Vel-Extrap	OPD-Min-ESTR
Regret at $T = 800$	**270.9**	327.9	343.2	377.3	879.0
Exponent α	0.67	0.67	0.76	0.67	0.95
95% bootstrap interval	[.63, .72]	[.63, .70]	[.72, .81]	[.62, .72]	[.94, .96]

Statistical paired significance. Table 16 is the paired difference matrix behind the benchmark of Table 1. All 48 method-by-horizon 95% paired bootstrap intervals lie strictly below zero; LDT attains the lowest regret on at least 12 of the 20 streams in every cell and on all 20 in six of the eight comparisons; and the absolute margin at $T = 800$ is at least 3.8 times the margin at $T = 100$, so the tracking advantage widens with the horizon. The 48 comparisons share the same 20 streams and form a single family, so we control the family wise error rate directly. Recomputing each comparison as an exact paired sign flip test, enumerating all 2^{20} sign assignments, and applying the Holm step down procedure across the family leaves all 48 significant at a family wise level of 0.05: the largest adjusted p value is 2.5×10^{-2}, and 36 of the 48 attain 1.9×10^{-6}, the smallest value the test can return at this sample size.

Additional baseline comparisons. Table 17 evaluates three additional published algorithms alongside OPD-Min-ESTR with per episode restarts, using matched streams, seeds, and structural feature representations. D-LinUCB (Russac et al., 2019) is the discounted linear learner, implemented in numerically stable recursive form with design matrix $V \leftarrow \gamma V + \varphi\varphi^{\top} + (1-\gamma)\lambda I$, companion matrix $\tilde{V} \leftarrow \gamma^2 \tilde{V} + \varphi\varphi^{\top} + (1-\gamma^2)\lambda I$, and confidence width $\|\varphi\|_{V^{-1}\tilde{V}V^{-1}}$, selecting interventions by the lower confidence bound since the objective is minimized. BOB (Cheung et al., 2019) is the parameter free meta algorithm running exponential weights over a discrete set of discount factors across temporal blocks, restarting its base learner at block boundaries. LowPopArt

Table 16: Paired significance matrix for Table 1. Mean paired regret differences, LDT minus baseline, across 20 paired streams with 95% bootstrap intervals. The final row in each block records the number of streams LDT wins.

T	vs. Lift-TS (ours)	vs. SW-Ridge	vs. Vel-Extrap	vs. OPD-Min-ESTR
100	-14.7 [-19.0, -9.8]	-3.5 [-6.4, -0.8]	-27.6 [-32.5, -23.2]	-54.0 [-62.3, -46.6]
200	-34.9 [-43.3, -26.2]	-13.1 [-20.7, -5.6]	-39.8 [-47.7, -32.7]	-131.3 [-149.4, -114.8]
300	-46.1 [-58.1, -33.6]	-21.2 [-32.5, -10.3]	-51.3 [-63.2, -40.6]	-210.5 [-240.0, -183.8]
400	-52.8 [-68.2, -36.8]	-31.2 [-45.9, -16.5]	-62.8 [-78.7, -48.0]	-289.4 [-329.5, -253.1]
600	-59.1 [-82.9, -34.6]	-53.0 [-74.9, -30.9]	-85.8 [-108.9, -63.6]	-448.6 [-510.4, -392.6]
800	-57.0 [-89.8, -22.5]	-72.3 [-102.0, -42.7]	-106.4 [-137.6, -76.5]	-608.1 [-690.6, -533.1]
Wins	$\geq 15/20$	$\geq 12/20$	$20/20$	$20/20$

T	vs. D-LinUCB	vs. BOB	vs. LowPopArt	vs. OPD-Min-ESTR restarted
100	-27.4 [-31.5, -23.8]	-54.7 [-61.4, -48.2]	-69.3 [-78.6, -60.6]	-83.3 [-96.4, -70.6]
200	-65.0 [-72.4, -58.8]	-122.8 [-134.0, -112.3]	-160.5 [-179.0, -143.1]	-191.0 [-218.6, -163.7]
300	-100.4 [-111.0, -91.2]	-202.7 [-222.8, -184.7]	-251.1 [-280.5, -224.0]	-302.2 [-346.3, -258.4]
400	-131.6 [-146.5, -118.9]	-278.9 [-306.1, -254.2]	-340.9 [-380.5, -304.0]	-413.5 [-472.7, -355.5]
600	-190.4 [-212.3, -170.8]	-432.4 [-471.7, -395.4]	-517.2 [-576.6, -461.9]	-637.0 [-726.3, -548.7]
800	-241.0 [-269.9, -214.3]	-578.1 [-634.1, -525.8]	-691.6 [-769.7, -618.8]	-859.4 [-978.0, -743.1]
Wins	$20/20$	$20/20$	$20/20$	$20/20$

(Jang et al., 2024) is the arm set dependent low rank bandit whose first stage optimizes a Frank-Wolfe design π over vectorized candidate interventions, estimates $\hat{\Theta} = \mathrm{mat}(Q(\pi)^{+} \frac{1}{n_1} \sum_i y_i v_i)$ with $Q(\pi) = \sum_a \pi_a v_a v_a^{\top}$, extracts the leading eigenvector as the recovered opinion direction, and hands it to the same second stage reduced optimism that OPD-Min-ESTR uses, so the comparison isolates the estimation stage. Every method receives the same candidate pool.

Every baseline is tuned under an oracle protocol: optimistic learners evaluate confidence intervals according to their analytical certificates with tight constants, algorithms with forgetting parameters receive oracle parameters chosen from candidate grids, and BOB receives the oracle best block length from $\{\lceil\sqrt{T}\rceil, \lceil T^{2/3}\rceil, \lceil d_p^{2/3}\sqrt{T}\rceil\}$. The oracle discount evaluates to $\gamma^{\star} = 1.0$ for D-LinUCB at every horizon. LDT leads every one of them at every horizon, as the lower block of Table 16 documents, the closest at $T = 800$ being D-LinUCB at $+241.0$ $[214.3, 269.9]$. That value reflects how smooth the stream is inside one episode, where geometric forgetting spends more variance than the drift it recovers, while the multiscale grid holds several effective memories at once.

Table 17: Three additional published algorithms across 20 paired streams of $J = 20$ episodes, reporting cumulative dynamic regret with 95% bootstrap intervals and best per row in bold.

T	**LDT (ours)**	D-LinUCB	BOB	LowPopArt	OPD-Min-ESTR restarted
100	**66.9**	94.3 [87, 101]	121.6 [112, 131]	136.2 [122, 150]	150.2 [133, 168]
200	**100.5**	165.5 [154, 177]	223.3 [206, 241]	261.0 [236, 287]	291.5 [257, 326]
300	**131.1**	231.5 [216, 247]	333.8 [307, 361]	382.2 [345, 420]	433.3 [382, 485]
400	**160.5**	292.0 [272, 312]	439.4 [404, 475]	501.3 [454, 550]	574.0 [507, 642]
600	**215.8**	406.2 [380, 432]	648.1 [598, 700]	732.9 [665, 802]	852.7 [755, 952]
800	**270.9**	511.8 [481, 543]	849.0 [781, 917]	962.5 [875, 1054]	1130.3 [1001, 1260]

Model based trackers. Every tracker in Table 1 is model free about the drift, fitting a linear model in a window and reading it off. Table 18 adds two that model the drift explicitly, in increasing order of what they are told, on the same streams, seeds, features and greedy control. The Kalman filter runs on $\theta_t \in \mathbb{R}^{d_p}$ with a random walk state model and scalar observation $\varphi_{a_t}^{\top}\theta_t + \zeta_t$, seeing exactly what LDT and SW-Ridge see. The extended Kalman filter runs on the rank r coordinate c_t with observation $c^{\top} G_a c$ for $G_a = U_r^{\top} X_a U_r$, linearized at the current mean; it is told the innate subspace $\mathcal{S}_r$, so it works in $r = 3$ dimensions where LDT works in $d_p = 21$. Process noise and jitter are selected on held out streams.

LDT leads the feature information block at every horizon, and by a widening margin. The Kalman filter trails by 298.4 $[231.7, 370.8]$ at $T = 800$, moving from $1.14\times$ LDT at $T = 100$ to $2.10\times$, since the drift is a smooth oscillation while a random walk state model carries a diffusion prior, so its innovations go into following a curve whose shape that prior leaves open. Handing the extended Kalman filter the innate subspace lets it work in $r = 3$ dimensions where LDT works in $d_p = 21$, and by $T = 800$ LDT is ahead by 83.4 $[58.5, 108.3]$: a subspace supplies a smaller state, while pretraining supplies the shape of the drift, and the second is what compounds as the stream runs. Table 19 repeats the comparison on the `structured` and `arc` generators, where LDT's frozen checkpoint and the filters' state models are off distribution together. LDT has the lowest regret in every column there, ahead of the subspace informed filter by 63.9 on `structured` and by 58.2 on `arc` at $T = 800$, so the ordering carries beyond the generator the checkpoint was trained on.

Table 18: Model based trackers on the `dynamical` streams (20 paired streams, $J = 20$ episodes, matched seeds), reporting cumulative dynamic regret with 95% bootstrap intervals. Best under feature information alone in bold.

	Feature information only			Given $\mathcal{S}_r$
T	**LDT (ours)**	SW-Ridge	Kalman	EKF
100	**66.9** [60, 74]	70.4 [63, 78]	76.3 [67, 86]	62.9 [56, 70]
200	**100.5** [89, 112]	113.6 [100, 127]	148.7 [129, 169]	111.2 [98, 124]
300	**131.1** [115, 148]	152.3 [134, 171]	220.5 [191, 251]	152.3 [135, 169]
400	**160.5** [140, 181]	191.7 [167, 216]	290.1 [251, 331]	194.1 [172, 216]
600	**215.8** [188, 245]	268.8 [232, 306]	430.5 [371, 493]	275.1 [245, 307]
800	**270.9** [234, 310]	343.2 [294, 394]	569.3 [491, 652]	354.3 [317, 393]

Table 19: The same trackers under generator shift, where the filters' state models and LDT's frozen checkpoint are all off distribution. Cumulative dynamic regret with 95% bootstrap intervals over 20 paired streams. Best in bold. No method is retrained for the shifted generator.

	`structured`		`arc`	
Method	$T = 200$	$T = 800$	$T = 200$	$T = 800$
LDT (ours)	**107.9** [97, 120]	**302.1** [270, 336]	**99.6** [91, 109]	**278.5** [249, 311]
SW-Ridge	113.6 [102, 125]	344.9 [301, 388]	113.7 [100, 128]	338.0 [292, 385]
Kalman	137.3 [121, 155]	514.2 [456, 574]	135.2 [120, 151]	504.2 [448, 564]
Given the innate subspace $\mathcal{S}_r$				
EKF	112.0 [100, 124]	366.0 [327, 404]	107.9 [94, 122]	336.7 [299, 376]

A tracker with the same offline budget. LDT reads 1500 offline episodes before deployment, and Kalman-Pre reads the same ones, drawn from the identical prior under the identical seed, so the two enter the stream off the same 462,950 recorded steps of θ_t. On them Kalman-Pre fits a linear Gaussian state model by pooled least squares, at order one, $\theta_{t+1} = A_1\theta_t + w_t$, and at order two, $\theta_{t+1} = A_1\theta_t + A_2\theta_{t-1} + w_t$, with $w_t \sim \mathcal{N}(0, Q)$. The family suits this stream closely, since a rotation through angle ω carries $x_{t+1} = 2\cos(\omega)\, x_t - x_{t-1}$ in whichever plane it acts, so the order two fit recovers the generating recursion itself, with one step residual 3.3×10^{-5}. Online the fit runs as an exact Kalman filter in companion form under the same greedy control. The order one fit is the one carried into the stream. The shared offline budget places Kalman-Pre within the paired interval at $T = 200$ in Table 20, and the margin in favour of LDT then opens to 14.5 at $T = 400$ and 42.2 at $T = 800$, both intervals clear of zero, the horizon dependence the amortization argument asks for.

Parameter space mechanism verification. Table 21 is the measurement behind §6. Replaying evaluation streams, we record at each interaction step the true parameter θ_t, the uncorrected reference estimate ($W_{\text{med}} = 60$), LDT's filtered estimate and the oracle window SW-Ridge estimate, and we define the window lag bias $b_t = \theta_t - \text{mean}_{u \in [t - W_{\text{med}} + 1, t]}\theta_u$, the systematic lag that forward trajectory extrapolation would target. Three findings characterize the filter. LDT reduces parameter estimation error by 10.2% relative to the uncorrected medium window and by 8.9% relative to oracle SW-Ridge; the cosine similarity between the learned correction u_t and the lag bias is -0.022, so

Table 20: A tracker given the offline budget LDT is given. Kalman-Pre reads the same 1500 offline episodes, with order and process noise scale selected on held out streams. Cumulative dynamic regret with 95% bootstrap intervals over 20 matched streams, paired in the last column.

T	**LDT (ours)**	Kalman-Pre	Paired difference
200	**100.5** [89, 112]	102.3 [91, 114]	-1.8 [-10.1, $+6.1$]
400	**160.5** [140, 181]	175.0 [156, 194]	-14.5 [-29.4, -1.1]
800	**270.9** [234, 310]	313.1 [278, 348]	-42.2 [-70.0, -16.2]

corrections act orthogonally to it; and the lag bias accounts for 0.340 of the residual error, under half of what remains. LDT therefore acts as a causal filter removing observation noise from the current estimate, which is both the quantity equation 11 asks it to reduce and the quantity Proposition 2 converts into a regret bound that needs no excitation condition.

Table 21: Parameter space mechanism verification ($T = 200$, 20 paired streams, 95% bootstrap intervals). Estimates are recorded on the streams the deployed policy collects.

Diagnostic metric	Measured value	95% bootstrap interval
Estimation error, LDT	**0.810**	[0.803, 0.816]
Estimation error, uncorrected reference window ($W = 60$)	0.901	[0.898, 0.904]
Estimation error, SW-Ridge ($W = 120$)	0.889	[0.884, 0.892]
Error ratio, LDT vs. uncorrected reference window	0.898	[0.891, 0.905]
Error ratio, LDT vs. SW-Ridge	0.911	[0.905, 0.917]
Cosine similarity between correction and lag bias	-0.022	[-0.040, -0.004]
Lag bias norm, relative to total parameter error 0.810	0.340	[0.318, 0.361]

The certificate the pretraining loss supplies. Proposition 2 turns the estimation error of Table 21 into a regret bound for the shipped policy, evaluated here term by term on the same rollouts, per episode at $T = 200$ over 400 episodes, so the bound and the regret it bounds are read off identical runs. The warmup rounds and the scheduled design draws are charged at the gap the algorithm actually pays on them, so every column bounds the same total regret. Two constants govern the chain. The feature diameter is $D = 0.285$ against $2\Phi_{\max} = 1.411$, a factor of 5.0, because Proposition 1 writes every X_a as a rank $|E_a|$ update of one common resolvent, so the arm features sit in a tight cluster and an $\arg\min$ sees only their spread. The chain then reads 7.74 at the realized arm pair, 20.44 at the worst case over arm pairs, and 42.53 once Cauchy-Schwarz replaces the arm geometry by $\|e\|_2$, against a measured regret of 5.02 at an operating point where the worst any policy can do is $\sum_t \Delta_{\max} = 27.8$. The realized pair form is within a factor of 1.5 of the regret it bounds and the pairwise form stays below that worst case, which locates what the remaining slack buys: the last step trades the alignment between estimation error and the directions the arms separate along for a certificate that no longer refers to the arms at all. That final form is the one worth having, because the Cauchy-Schwarz step costs 0.3% over $D \sum_t \|e_t\|_2$ and leaves $\bar{\mathcal{L}} = 0.663$ as its only data dependent factor, the empirical value of the pretraining objective equation 11. Minimizing the offline loss is minimizing the deployed regret certificate: §5.2 and §4.3 meet at that number.

What the correction moves inside it. Theorem 4 holds at every setting of ω, which is what makes it a safeguard and what keeps it silent on which of two estimates lies closer to the truth. Corollary 1 answers that in the currency of the certificate, and Table 22 evaluates both sides. Two policies are run on identical streams under common random numbers, differing at one line, the estimate the greedy rule reads: the deployed filter under the safeguard, and the anchor window estimate alone. The ratio the corollary predicts is $\sqrt{0.663/0.802} = 0.909$ against a measured certificate ratio of 0.910, so the reduction of the offline objective passes into the certificate at the square root on all 20 streams, and realized regret moves with it, 5.02 against 5.92 per episode at a paired difference of 0.90 [0.51, 1.32] in favour of the deployed filter.

The certificate across horizons. Table 23 evaluates the same chain at every horizon of Table 1, on identical streams, seeds and frozen filter, reproducing those per episode values in its $T = 200$

Table 22: The deployed correction against the zero correction policy that plays on the anchor estimate, run under common random numbers on the streams, seeds and frozen filter of Table 23 at $T = 200$, per episode over 400 episodes. The certificate is the leading form $D\,T\sqrt{\bar{\mathcal{L}}}$ of Proposition 2 along each policy's own rollout, warmup and scheduled draws charged at their realized gap. Differences are paired with 95% bootstrap intervals over the 20 streams.

	Deployed	Zero correction	Difference
Pretraining objective $\bar{\mathcal{L}}$	**0.663**	0.802	0.139 [0.130, 0.147]
Estimation error $\|e_t\|_2$	**0.810**	0.895	0.085 [0.080, 0.091]
Certificate $D\,T\sqrt{\bar{\mathcal{L}}}$	**42.53**	46.72	4.19 [3.92, 4.45]
Measured regret	**5.02**	5.92	0.90 [0.51, 1.32]

column. The two ends answer different questions. The pretraining objective is a tracking level, holding between 0.660 and 0.674 across an eightfold span with a slope whose interval covers zero, so the forms that reach e_t through its norm alone fix the constant and grow at $T^{1.06}$. The leading form keeps the arm geometry and grows at $T^{0.77}$ [.73, .81], below the $T^{0.91}$ that Theorem 3 certifies for the schedule this policy runs, so the two routes to the same policy land on the same side of it; it stays above the realized regret at every horizon, within a factor of 1.4 at $T = 100$ widening to 1.7 at $T = 800$, and its exponent exceeds the realized one by 0.098 [0.074, 0.129] on identical stream resamples. What separates the two ends is alignment. The last row reports the decision projected loss $\bar{\mathcal{G}} = \frac{1}{T}\sum_t (\varphi_{a_t^\star} - \varphi_{a_t})^\top e_t$, the estimation error read along the direction each round's decision exposes, and it falls at $T^{-0.15}$ while $\bar{\mathcal{L}}$ holds flat: a fixed tracking error converts into less regret as the horizon grows, because the trajectory settles along the oscillator and leaves less of that error in the directions the arms separate. All twenty streams are held out from pretraining and the upper limit row is their one sided 95% bound, so each column bounds a fresh stream.

Table 23: Proposition 2 at every horizon, on the streams, seeds and frozen filter of Table 1, per stream over 20 episodes, so the first row is the LDT column of Table 15. Certificates appear in proof order, and α is the log-log slope in T with 95% bootstrap intervals over 20 paired streams.

T	100	200	300	400	600	800	α
Measured regret	**66.9**	**100.5**	**131.1**	**160.5**	**215.8**	**270.9**	0.67 [.63, .72]
Certificate, realized pair	92.0	154.8	209.7	262.0	362.1	458.1	0.77 [.73, .81]
upper 95% limit over streams	101.1	171.5	233.1	291.8	404.5	513.8	
Certificate, pairwise	200.7	408.8	617.5	826.6	1246.3	1672.1	1.02 [1.01, 1.03]
Certificate, pretraining form	395.9	851.5	1303.8	1759.2	2681.1	3607.6	1.06 [1.06, 1.06]
Pretraining objective $\bar{\mathcal{L}}$	0.674	0.663	0.660	0.660	0.663	0.666	−0.006 [−.014, +.002]
Decision projected loss $\bar{\mathcal{G}}$	0.0362	0.0333	0.0308	0.0294	0.0277	0.0266	−0.15 [−.19, −.11]

Baseline variants. We run OPD-Min-ESTR in four configurations, so the streaming comparison is read against its strongest available form: per episode restart, which re-runs subspace exploration on each episode; within episode re-estimation, which restarts it periodically inside an episode; pooled subspace estimation, which pools exploration data across the episodes of a stream and is legitimate since those episodes do share the prior; and oracle subspace initialization, which is handed the exact $t = 0$ subspace with no exploration spent. Over 20 paired streams at $T = 200$ the four read 291.5, 305.2, 246.3 and 231.7 in dynamic regret against 100.5 for LDT, so the last is the strongest of them and is the configuration reported everywhere else in the paper.

Bayesian regret bound verification. Table 24 measures Lift-TS at horizon $T = 3000$ against the tighter of the two certificates in §4.1. Theorem 5 is the benchmark the lifted structure hands to exact posterior sampling, while Theorem 1 is what the implemented Gaussian sampler carries through Lemma 2, looser by $\tilde{\mathcal{O}}(d_\star)$ at the sufficient setting of its two constants, so the table is graded against the first line, the tighter of the two. Because Theorem 5 is normalized by reward range, the analytical bound evaluates to $\bar{\Delta}_{\max}\sqrt{\frac{1}{2}d_\star \log K \cdot T}$ with $\bar{\Delta}_{\max}$ the empirical range averaged across the same instances, the constant a Bayes bound averaged over instances carries. The noise level $\sigma = 0.1$ sits

below $\Delta_{\max}/2$ on every instance, the smallest observed range being 0.655. Since posterior sampling explores stochastically, we report 95% bootstrap intervals over 24 seeds, with growth exponents α fitted across eight log-spaced checkpoints. The analytical certificate holds with a safety margin of at least $6.0\times$ across all graph sizes, and it still holds with at least a factor of 3.7 when the mean range is replaced by the smallest range seen at that size.

Table 24: Lift-TS against the bound of Theorem 5 at $T = 3000$. Here $\bar{\Delta}_{\max}$ is the average reward range, the analytical bound is $\bar{\Delta}_{\max}\sqrt{\frac{1}{2}d_\star \log K \cdot T}$, and α the empirical growth exponent. Brackets denote 95% bootstrap intervals over 24 random seeds.

$\lvert V\rvert$	$d_\star$	Regret at $T = 3000$	$\bar{\Delta}_{\max}$	Analytical bound	Exponent α
8	29	74.1 [30.9, 131.5]	1.32	590	0.57 [0.36, 0.70]
16	100	231.2 [110.6, 368.6]	1.91	1585	0.74 [0.56, 0.83]
32	100	145.4 [62.4, 257.3]	1.05	876	0.64 [0.40, 0.80]

The sampler the certificate speaks about. Lemma 2 identifies the conjugate pair (μ_t, Σ_t) that Algorithm 1 carries with the ridge estimate and the design matrix in which Theorem 1 is written, $\Sigma_t = \sigma^2 V_t^{-1}$ and $\mu_t = V_t^{-1} b_t$ at ridge $\lambda = \sigma^2/\tau^2$, so the played draw is the analysed one at inflation $c_{\text{eff}} = c\sigma$. The deployed constants $(\tau, \sigma, c) = (1, 0.1, 0.6)$ place the runs of Table 24 at $\lambda = 10^{-2}$ and $c_{\text{eff}} = 6 \times 10^{-2}$, the pair Theorem 1 reads through κ_λ and $p_{\text{opt}}(c, \delta)$. Table 25 evaluates both sides of that identity on those same runs, rebuilding V_t and b_t alongside the recursion and comparing at eight logarithmically spaced checkpoints; the relative departure stays at the accuracy of the linear solve, below 10^{-8} at every checkpoint, size and seed. The one hypothesis the inflation governs is measured beside it. Concentration and anti concentration are properties of the isotropic Gaussian and hold at every inflation, while the inflation itself sets the probability of the optimistic event $\min_a \psi_a^\top \tilde{\theta}_t \le \min_a \mu_a$, which is the quantity $p_{\text{opt}}(c, \delta)$ Theorem 1 carries on the event $\mathcal{E}_\delta$ of its statement, where any inflation $c \ge \beta_T(\delta)$ holds it at or above $\bar{\Phi}(1)$. The deployed c_{eff} sits below that threshold, and there Lemma 3 evaluates the same quantity as a frequency, through the tail $\bar{\Phi}(G_t/c)$ of the pointwise error G_t at the optimal arm: whenever the ridge estimate already values that arm at least as favourably as the truth does, half of the draws keep it there at every inflation, so the frequency of the optimistic event is at least half the frequency of $\{G_t \le 0\}$, which the same runs record. That floor reads 0.267, 0.168 and 0.156 across the three sizes, the realized frequency 0.497, 0.315 and 0.330 clears it at each, and across seeds that frequency rank correlates with realized regret at -0.59, -0.72, -0.52, the dependence the analysis predicts.

Table 25: The implemented sampler of Algorithm 1 against the sampler Theorem 1 analyses, on the runs of Table 24 at the ridge $\lambda = 10^{-2}$ and inflation $c_{\text{eff}} = 6 \times 10^{-2}$ of Lemma 2. The last column is the largest relative departure of (Σ_t, μ_t) from $(\sigma^2 V_t^{-1}, V_t^{-1} b_t)$, and intervals are over the 24 seeds.

$\lvert V\rvert$	Optimism $\hat{p}$	Certified floor $\hat{q}/2$	Rank correlation	Relative departure
8	0.497 [0.356, 0.637]	0.267	-0.59	3.2×10^{-9}
16	0.315 [0.187, 0.447]	0.168	-0.72	8.3×10^{-9}
32	0.330 [0.207, 0.455]	0.156	-0.52	3.4×10^{-9}

Task parameter robustness. Table 26 evaluates thirteen one factor variations of the streaming benchmark around the base configuration ($T = 200$, $J = 20$, 20 paired streams, oracle windows for reactive trackers, and the identical pretrained LDT checkpoint). In each setting, the strongest alternative is the method of lowest mean regret there, Lift-TS included, so LDT is always compared against the best available method. The last two settings match the evaluation range of Cinus et al. (2026), with observation noise at $\sigma = 1.0$, ten times the nominal level, and the polarized innate opinions of their Appendix G.3, $s_i \mapsto \text{sign}(s_i)|s_i|^{1/3}$ renormalized to unit norm. Under polarized opinions the innate vectors move off $\text{span}(U_r)$, so every structure aware method there reads the spectral features at a relative misfit of 1.9%, while Θ_t stays exactly rank one throughout, which is the form the baseline assumes. LDT leads there by 5.3 with interval $[-8.8, -1.9]$, so the spectral representation holds its margin at that misfit.

Five of the thirteen configurations reside outside the support of LDT's pretraining prior, and the evaluation horizon $T = 200$ is absent from pretraining across all rows, so every row is a transfer across task parameters. LDT is significantly below OPD-Min-ESTR in all thirteen settings, and its paired interval against the strongest alternative favours LDT in nine of them.

Table 26: One factor variations around the nominal configuration ($|V| = 32$, $r = 3$, $K = 100$, drift rate 1.0, $\sigma = 0.1$). Δ is the 95% paired bootstrap interval on LDT minus the strongest alternative, † marks settings outside the pretraining prior, and bold marks LDT where it is lowest.

		Strongest alternative			
Setting	**LDT (ours)**	Method	Regret	OPD-Min-ESTR	Δ
Nominal	**100.5**	SW-Ridge	113.6	231.7	-13.1 $[-20.7, -5.9]$
$\|V| = 16$	**218.8**	Lift-TS (ours)	226.5	424.4	-7.7 $[-21.9, +7.1]$
$\|V| = 64^\dagger$	**31.9**	SW-Ridge	34.0	68.0	-2.1 $[-4.0, -0.5]$
$r = 2$	**104.9**	SW-Ridge	116.0	256.9	-11.1 $[-24.2, +0.2]$
$r = 5^\dagger$	108.5	SW-Ridge	105.9	192.6	$+2.6$ $[-3.8, +9.0]$
$K = 50^\dagger$	**94.5**	SW-Ridge	108.0	215.4	-13.5 $[-20.1, -7.2]$
$K = 200$	**113.1**	SW-Ridge	127.3	248.0	-14.2 $[-24.2, -5.2]$
Drift rate 0.5	**102.7**	SW-Ridge	114.4	229.5	-11.8 $[-19.1, -4.7]$
Drift rate 2.0	**98.6**	SW-Ridge	112.1	232.9	-13.5 $[-20.2, -6.8]$
$\sigma = 0.05$	**95.1**	Lift-TS (ours)	101.4	204.0	-6.4 $[-14.1, +1.2]$
$\sigma = 0.2$	**112.4**	SW-Ridge	122.3	251.2	-9.9 $[-17.2, -2.9]$
$\sigma = 1.0^\dagger$	**145.4**	SW-Ridge	158.6	271.5	-13.2 $[-18.9, -7.3]$
Polarized opinions $s^\dagger$	**70.5**	SW-Ridge	75.7	140.2	-5.3 $[-8.8, -1.9]$

Generative scope beyond the one factor sweep. Table 27 varies parts of the generative model that Table 26 holds fixed, namely the shape of the innate opinion law, the subspace the innate opinions occupy, and whether the graph is static. The three prior rows replace the Gaussian vertex marginal by the coordinatewise quantile map $F^{-1} \circ \Phi$, giving Student t with three degrees of freedom, Laplace at equal variance, and an equal weight mixture of $\mathcal{N}(-1.5, 0.5^2)$ and $\mathcal{N}(+1.5, 0.5^2)$ that places the population in two camps. The two graph rows rewire a Poisson number of edges per round at mean rate 0.05 and 0.2, so L, U_p and every φ_a go stale while the deployment keeps the features it built at round zero. Everything else follows Table 26, frozen checkpoint included.

Theorem 6 predicts the outcome and the last two columns grade it. Each configuration is measured for its leakage $\|q_t\|$ with $q_t = (I - U_p U_p^\top) s_t$ and for the realized model error $\varepsilon_{\mathrm{mis}} = \max_a |\mu_a(t) - \varphi_a^\top \theta_t|$, reported in units of the reward spread $\max_a \mu_a(t) - \min_a \mu_a(t)$, which is the fraction of the signal no learner reading φ_a can represent. Two mechanisms generate that error and the table separates them. The prior rows keep the graph fixed and push innate opinion mass out of the spectral window, up to $\|q\| = 0.340$ under the two camp mixture, so the residual enters $\varepsilon_{\mathrm{mis}}$ through truncation and part *(ii)* of the theorem converts $\|q\|$ into it. The graph rows leave leakage at machine zero, since s_t never leaves $\mathrm{span}(U_p)$, and the error comes from features built at round zero against a graph that has since moved, which part *(i)* prices from the measured $\varepsilon_{\mathrm{mis}}$ itself, that part reading the scalar and not its origin. One penalty therefore covers both paths and is linear in $\varepsilon_{\mathrm{mis}}$, and the sweep traces it from 2.3×10^{-7} to 0.660 of the reward spread. Across that range LDT leads in every configuration, by 8.1 to 13.2 with all six intervals excluding zero: a third of the innate opinion outside the spectral window, a bimodal population, heavy tails and a graph that rewires under the deployment each cost margin without reaching the point where the structural prior stops paying.

The candidate pool. Table 26 varies the environment around a fixed pool and Table 30 varies how many interventions the pool holds, so Table 28 varies what one intervention is: the number of edges it modifies, at 4 and 31 against the nominal 16; the pairs it may draw from, restricted to the edges the graph already carries out of all 496 vertex pairs, which concentrates the pool on the existing structure; and the weight range, moved from $[0.5, 1.5]$ to $[1.5, 3.0]$. The same frozen filter reads all four through the same 21 coordinates and keeps the lowest regret in every row, ahead of the strongest alternative by 2.3 to 22.4 and of OPD-Min-ESTR by 62.5 to 176.6.

Table 27: Generative scope beyond Table 26, same protocol and column meanings, with $\|q\|$ the innate opinion mass outside $\mathrm{span}(U_p)$ and the last column the realized model error of Theorem 6 in units of the reward spread. Lowest mean regret per row in bold.

		Strongest alternative				
Setting	**LDT (ours)**	Method	Regret	Δ	$\|q\|$	$\varepsilon_{\mathrm{mis}}$/spread
Nominal	**100.5**	SW-Ridge	113.6	-13.1 $[-20.7, -5.6]$	$< 10^{-15}$	2.3×10^{-7}
Student t prior	**133.0**	SW-Ridge	143.7	-10.8 $[-19.5, -2.2]$	0.244	0.151
Laplace prior	**119.2**	SW-Ridge	130.4	-11.2 $[-20.5, -2.5]$	0.153	0.077
Two camp mixture prior	**75.8**	SW-Ridge	83.9	-8.1 $[-13.3, -2.8]$	0.340	0.294
Graph rewiring rate 0.05	**98.1**	SW-Ridge	111.2	-13.2 $[-20.7, -5.9]$	$< 10^{-15}$	0.248
Graph rewiring rate 0.2	**94.8**	SW-Ridge	103.2	-8.4 $[-16.4, -0.3]$	$< 10^{-15}$	0.660

Table 28: The candidate pool, protocol and columns as in Table 26. The nominal pool draws $k = 16$ of the 496 vertex pairs uniformly for each of $K = 100$ interventions, at weights in $[0.5, 1.5]$.

		Strongest alternative			
Setting	**LDT (ours)**	Method	Regret	OPD-Min-ESTR	Δ
Nominal	**100.5**	SW-Ridge	113.6	231.7	-13.1 $[-20.7, -5.6]$
$k = 4$ modified edges	**98.4**	SW-Ridge	102.4	182.8	-4.0 $[-10.3, +2.3]$
$k = 31$ modified edges	**86.6**	SW-Ridge	100.7	236.5	-14.1 $[-24.6, -6.7]$
Candidates on existing edges	**57.0**	SW-Ridge	59.2	119.5	-2.3 $[-7.8, +2.5]$
Weights in $[1.5, 3.0]$	**113.5**	SW-Ridge	135.9	290.1	-22.4 $[-32.0, -13.0]$

Transfer to real networks. Table 29 evaluates streaming transfer by replacing synthetic topologies with the seven real networks of Table 12, holding all other parameters fixed ($T = 200$, $J = 20$, 20 paired streams, oracle windows, and OPD-Min-ESTR with oracle initialization). The seven networks contribute measured topology, while innate opinions and their drift come from the same spectral oscillator as every other experiment here, so what this measures is transfer across real topology under the controlled opinion process of §6. LDT is pretrained only on synthetic graphs of at most 48 nodes, so this is a transfer test as well as a robustness test. LDT significantly outperforms Vel-Extrap and OPD-Min-ESTR on all seven real networks and SW-Ridge on six of them, 20 of the 21 paired intervals strictly excluding zero in its favour, and attains the lowest cumulative regret among all five evaluated methods on Davis, Karate, Les Misérables and ego-Facebook.

Read against Lift-TS, the same seven networks trace the amortization boundary of §4, and what locates it is the measured reward spread. A wide spread preserves the ranking of candidate interventions under delayed parameter estimates, which is the regime in which averaging over the whole history is already close to optimal; the $|V| = 16$ setting of Table 26 sits there with a spread of 0.156, as do Florentine and Les Misérables at 0.188 and 0.152 against 0.080 on nominal synthetic graphs. The political blogosphere ($|V| = 1222$, $25\times$ the largest pretraining graph) is the extreme case: sparse and strongly clustered, algebraic connectivity 0.169, it carries the widest spread of the seven at 0.209, and there our training free sampler leads, so the two learners of §5 take the first two places. LDT still outperforms SW-Ridge by 42.4, Vel-Extrap by 129.7 and OPD-Min-ESTR by 537.3 on that topology, all intervals excluding zero, cutting the baseline's regret by $3.6\times$.

As the spread contracts the tracking margin widens, and it widens against the direction of network size. On Hamsterster ($|V| = 2000$, $40\times$ pretraining) the spread is 0.131, below the 0.152 at which the wide spread regime begins across these topologies, the two learners of §5 again take the first two places, and LDT is ahead of SW-Ridge by 36.6, Vel-Extrap by 94.8 and OPD-Min-ESTR by 380.0, a factor of 3.9. On ego-Facebook ($|V| = 4039$, $80\times$ pretraining) the spread is 0.040, the ranking of candidate interventions shifts continuously, and LDT again takes the lowest regret, ahead of SW-Ridge by 5.3, Vel-Extrap by 17.8, Lift-TS by 31.9 and OPD-Min-ESTR by 114.8, a factor of 3.4, with the intervals against Vel-Extrap, Lift-TS and OPD-Min-ESTR excluding zero. Across the three largest networks size grows $1222 \to 2000 \to 4039$ while the spread contracts $0.209 \to 0.131 \to 0.040$, so the boundary §4 describes is the one the measurements follow.

Table 29: Streaming performance head-to-head on seven held out real networks ($T = 200$, $J = 20$, 20 paired streams). The upper block reports dynamic regret, best per row in bold; the lower gives 95% paired bootstrap intervals on regret differences relative to LDT (negative values favour LDT).

Network ($\lvert V\rvert$)	**LDT (ours)**	Lift-TS (ours)	SW-Ridge	Vel-Extrap	OPD-Min-ESTR
Physical cumulative dynamic regret					
Florentine (15)	256.7	**251.7**	276.3	328.5	486.0
Davis (32)	**129.9**	156.3	142.3	165.7	270.0
Karate (34)	**100.1**	128.2	112.0	124.5	184.7
Les Misérables (77)	**167.6**	178.7	210.2	255.9	460.5
Political blogs (1222)	206.8	**161.5**	249.2	336.5	744.0
Hamsterster (2000)	132.4	**131.8**	169.0	227.2	512.3
ego-Facebook (4039)	**48.6**	80.6	53.9	66.4	163.4
Paired bootstrap difference intervals relative to LDT					
Florentine (15)	—	$[-7.6, +17.4]$	$[-28.6, -10.3]$	$[-83.7, -60.4]$	$[-247.6, -211.0]$
Davis (32)	—	$[-32.9, -19.9]$	$[-20.9, -4.2]$	$[-42.3, -29.1]$	$[-147.3, -133.8]$
Karate (34)	—	$[-35.3, -21.2]$	$[-21.7, -2.5]$	$[-32.6, -16.2]$	$[-92.9, -76.5]$
Les Misérables (77)	—	$[-27.0, +5.2]$	$[-63.7, -23.9]$	$[-104.4, -74.6]$	$[-314.1, -272.1]$
Political blogs (1222)	—	$[+33.5, +56.4]$	$[-64.9, -20.8]$	$[-152.6, -106.6]$	$[-556.1, -517.5]$
Hamsterster (2000)	—	$[-8.1, +9.7]$	$[-56.4, -19.8]$	$[-112.5, -79.1]$	$[-396.0, -365.0]$
ego-Facebook (4039)	—	$[-39.3, -24.8]$	$[-12.4, +0.7]$	$[-24.2, -12.3]$	$[-124.3, -105.3]$

F THE LEARNED PRIOR POLICY

Theorem 2 and Corollary 2 characterize a policy that is pretrained for decision making and subsequently frozen. §5 translates the core design insight of that result into two operational learners, a posterior sampler for stationary instances and a causal filter for drifting streams. This appendix implements the literal object that the theoretical result describes, enabling the pretraining side of the amortization boundary to be evaluated on its own terms. The learned prior policy is a complementary design. It chooses interventions from the prior it carries together with accumulated interaction history without reserving rounds for exploration, which is the exact regime certified by those theoretical guarantees. Because it dedicates no horizon rounds to exploration and fits no per instance estimator, any performance margin it achieves is carried by the learned prior alone.

Two predictions follow and are evaluated here against both OPD-Min-ESTR and Explore-Commit, one member of the structure blind class Corollary 2(i) analyzes. The corollary predicts a margin over Explore-Commit that widens with candidate pool size, and Theorem 6 bounds excess regret under subspace misspecification, so one frozen checkpoint should stay effective on networks drawn outside its pretraining distribution. Explore-Commit follows the canonical design of Lattimore & Szepesvári (2020), round-robin exploration for $\max\{K, \lceil T^{2/3} \rceil\}$ rounds followed by commitment to the empirical best arm, the budget its own regret analysis prescribes.

Policy model architecture. While Lift-TS constructs posterior beliefs from scratch on every graph, instances sampled from a shared low rank prior (Assumption 1) share structural properties that can be learned offline. The learned prior policy (LPP) is an in context transformer (Vaswani et al., 2017; Lee et al., 2023), pretrained once and then frozen, that maps candidate arm features and interaction histories to optimal interventions. Each intervention enters through the opinion free spectral feature φ_a of equation 8, so that an index, which carries no structural meaning from one graph to another, is never read and one checkpoint remains portable across graphs.

Sufficient statistic channel. Each arm token also carries the mean reward and pull count accumulated within the current context, which under the Gaussian observation noise of §3.2 are a sufficient statistic for that arm's observations. Each token therefore holds everything a posterior update would draw from that arm, and the policy can infer what the shared prior implies for unpulled candidates. Tokens enter as an unordered set, so the mapping is invariant to arm ordering and to pool size K: one frozen model serves pools of arbitrary cardinality, which is how the $\sqrt{\log K}$ dependence certified by Corollary 2 is realized in practice.

Table 30: Arm count separation aligned with Corollary 2(i). Final regret at $T = 2000$ over 20 paired instances as the candidate pool grows. Each Δ is the paired difference of LPP against that method with a 95% bootstrap interval, negative values favouring LPP.

		Explore-Commit		OPD-Min-ESTR	
K	**LPP (ours)**	Regret	Δ	Regret	Δ
100	**246.6**	393.3	-146.7 [-233.4, -56.9]	903.2	-656.6 [-870.9, -468.0]
200	**381.8**	517.8	-136.0 [-248.6, -6.6]	1054.4	-672.7 [-1053.7, -363.2]
400	**466.9**	1159.4	-692.5 [-909.0, -460.2]	1209.5	-742.6 [-1002.3, -519.8]
800	**653.3**	2439.9	-1786.6 [-2203.0, -1388.4]	1373.7	-720.5 [-1028.9, -422.1]
1600	**635.5**	5008.6	-4373.1 [-5123.2, -3585.5]	1219.2	-583.7 [-778.3, -394.8]

Supervised training objective. The policy g_ϕ is trained by supervised imitation. Given instance distribution $\mathcal{D}$ and interaction history h, training minimizes cross entropy against the optimal arm $a^\star(I) = \arg\min_a \mu_a(I)$:

$$\mathcal{L}_{\text{amort}}(\phi) \;=\; \mathbb{E}_{I\sim\mathcal{D},\,h}\big[\,-\log g_\phi\big(a^\star(I) \mid \{\varphi_a\}_{a=1}^{K},\, h\big)\big]. \tag{28}$$

Supervising optimal arm selection trains g_ϕ to approximate in context posterior sampling (Lee et al., 2023; Lin et al., 2024), which is exactly the hypothesis Theorem 2 prices, so the policy selects from accumulated context with no dedicated exploration phase.

Architecture and pretraining. LPP is parameterized as a transformer over arm tokens with full self attention and no positional codes, containing 4 layers, 8 heads, hidden dimension $d_{\text{model}} = 256$, feedforward width 1024, GELU activations (Hendrycks & Gimpel, 2016), and dropout 0.1, totaling 3,303,681 parameters. Arms enter as the spectral feature $\varphi_a \in \mathbb{R}^{21}$ of §5.2, the same encoding every other result in the paper reads. A context token embeds a pulled arm together with its observed reward, while a candidate token embeds an arm together with its empirical mean and pull fraction in the current context. A learned type embedding separates the two token families, and a scalar prediction head reads each candidate token so that a softmax over candidate arms yields the predicted optimal arm, and the pool size enters only through the number of candidate tokens.

Pretraining samples 25,000 synthetic tasks with 1,000 held out for checkpoint selection, on Erdős-Rényi graphs with $|V| \in \{8, 16, 32, 64\}$, $k = |V|/2$ pairs per arm, rank $r = 3$ and noise $\sigma = 0.1$. Candidate pool size is drawn uniformly from $\{50, \dots, 250\}$ per task, a range the two smallest pools of Table 30 sit inside and the three largest exceed, and context length uniformly from $\{1, \dots, 200\}$ rounds of randomly pulled arms labelled by the closed form optimal arm, so the policy meets the range of history budgets deployment presents. The loss is masked cross entropy in equation 28 over the candidate set, and optimizer hyperparameters match Table 12.

Context management at deployment. The maximum context length seen during pretraining is 200 rounds, whereas Table 30 extends to $T = 2000$. At each round the policy reads a rolling buffer of the 200 most recent rounds, with every candidate token's empirical mean and pull fraction computed strictly over that buffer, so inputs stay inside the pretraining domain and the context block never exceeds 200 tokens. The candidate block instead scales with the pool, reaching 1600 tokens at the largest evaluated size; because arms carry no index embeddings, candidate attention is permutation equivariant and each candidate is read from its own features contextualized by the shared history. One frozen checkpoint therefore serves pools far larger than the range $\{50, \dots, 250\}$ used during pretraining, and the same weights serve every horizon in Table 30.

Amortization predictions under homophily. Table 30 and Table 31 evaluate two predictions of §4 directly, and Figure 3 carries two more, the identification cost of Corollary 2(ii) in panel (a) and, in panel (b), the power law separation of Theorem 3 on a single drifting episode at a fixed variation budget, where a windowed tracker given its best window at every horizon scales as $T^{0.74}$ against $T^{1.00}$ for a policy that commits once. Corollary 2(i) proves that a policy reading the index reward history alone pays a term growing as $\Omega(\min\{K, T\})$ in the pool, while a learner operating under the shared prior pays only $\sqrt{\log K}$, and Table 30 measures both sides on the same instances. Explore-Commit, one member of that class, grows as $K^{1.06}$ at $R^2 = 1.00$ over the three largest pools, the

Table 31: Zero shot transfer. Regret at $T = 100$ on held out real networks under LPP pretrained on synthetic graphs alone and frozen. The last column is the realized model error of Theorem 6 in units of the reward spread, and each Δ is as in Table 30, over eight instances per graph.

			OPD-Min-ESTR		Explore-Commit		
Graph	$\lvert V\rvert$	**LPP (ours)**	Regret	Δ	Regret	Δ	ε_{mis}/spread
Florentine	15	**78.9**	110.6	-31.7 $[-50.5, -14.8]$	109.0	-30.1 $[-49.5, -12.8]$	1.14
Davis	32	**43.0**	76.5	-33.5 $[-45.6, -22.5]$	75.8	-32.8 $[-44.8, -22.1]$	2.72
Karate	34	**31.2**	49.6	-18.4 $[-23.4, -13.9]$	50.5	-19.3 $[-25.4, -13.8]$	3.23
Les Misérables	77	**132.0**	251.9	-119.9 $[-182.6, -60.0]$	253.0	-121.0 $[-182.1, -61.6]$	2.42

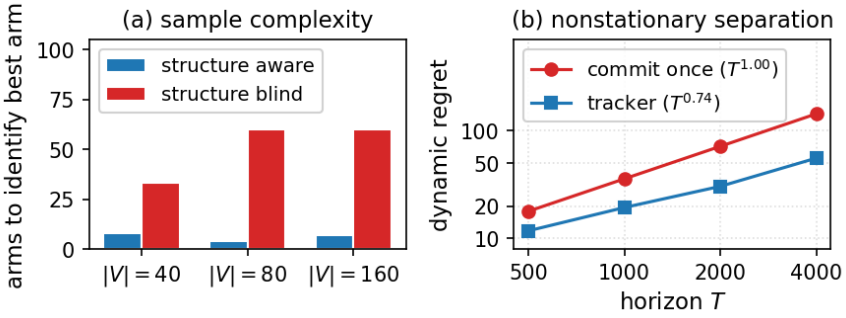

Figure 3: Supporting analyses. *(a)* Structure aware identification is flat in $|V|$ while the structure blind baseline grows (Corollary 2(ii)). *(b)* Nonstationary dynamic regret on log-log axes, the tracker at $T^{0.74}$ against commit once at $T^{1.00}$ (Theorem 3).

linear rate the corollary names once the exploration stage dominates its regret, while the frozen policy grows as $K^{0.35}$ across the sweep and is flat over the final doubling. LPP attains the lowest regret at every pool size, all ten paired intervals excluding zero, and its margin over Explore-Commit widens from -146.7 at $K = 100$ to -692.5 at $K = 400$ and -4373.1 at $K = 1600$.

This behaviour reflects the mechanism established in Corollary 2(i). Explore-Commit must sample every arm at least once, causing its exploration budget to scale as $\Omega(K)$ and consume 1600 of the 2000 rounds at the largest pool, whereas the frozen policy reads the shared prior and incurs only the constant per round imitation penalty R_{pre} of Theorem 2, which is independent of K. The two smallest pools lie inside the pretraining cardinality range and the three largest exceed it, with one frozen checkpoint serving every row. Furthermore, the transfer evaluation in Table 31 holds out network topology, network size and innate opinion distribution, and runs on the four real networks without adaptation. Each of those networks draws its innate opinions from a subspace pinned off the spectral window, leaving leakage $\|q\|_2$ between 2.9 and 8.3 and a realized model error between 1.1 and 3.2 times the reward spread, and the frozen policy leads on all four with every paired interval excluding zero. Both evaluations demonstrate that an offline structural prior eliminates the exploration penalty of large arm pools and transfers across novel graph families, which is the pretrained side of Theorem 2 measured on pools and networks the model never saw, in the misspecified regime Theorem 6 bounds. LPP isolates offline amortization on its own, the component Lift-TS and LDT each pair with online estimation, so it exercises the closed form features underpinning both learners.

Measuring the distance to posterior sampling. Theorem 2 is stated for a policy whose per round action distribution lies within total variation ε_{pre} of exact posterior sampling, while equation 28 is cross entropy against the optimal arm, so ε_{pre} is a quantity to measure. Under Assumption 1 it is measurable without approximation. Writing $s = U_r c$ on the leading nontrivial Laplacian eigenvectors of the instance, the basis the streaming benchmark reads, and normalizing the innate opinions, μ_a depends on c only through $\hat{c} = c/\|c\|_2$ and is even in it, so the posterior lives on the projective space of directions whatever the history, with a uniform prior and a likelihood Gaussian in μ_{a_t}. Quadrature on a Fibonacci lattice of 40,000 directions then evaluates the exact posterior sampling action distribution at every round; refining it fourfold moves the gap weighted deviation the bound is stated on by 1.3×10^{-4} per round on average and by 1.2×10^{-3} at the 95th percentile, so the

Table 32: The pretraining gap of Theorem 2, measured against exact posterior sampling computed by quadrature, over 40 held out instances at $K = 100$ and $T = 200$. Per round quantities and the bounds are summarized over the 195 post warmup rounds of every instance.

Per round quantity	Mean	Median	95th pct.
Total variation to exact posterior sampling, $\varepsilon_{\mathrm{pre}}$	0.233	0.024	1.000
Mass the exact posterior places on its own mode	0.977	1.000	1.000
Selection disagreement with that mode	0.218	0.000	1.000
Gap weighted deviation $\sum_a \lvert\pi_t(a) - p_t(a)\rvert\Delta_t(a)$	0.144	0.010	0.872
Regret of exact posterior sampling	0.002	0.000	0.004
Regret of the deployed policy	0.143	0.008	0.872
Largest gap $\Delta_{\max}$	4.429	3.988	8.289
Bound over the 195 policy rounds		**Value**	
Total variation relaxation		373.2	
Theorem 2, gap weighted R_{pre}		**28.5**	
Measured regret of the deployed policy		27.8	

lattice is converged in the currency of Theorem 2. Table 32 reports the comparison on 40 held out instances at $|V| = 32$, $r = 3$, $K = 100$ and $T = 200$, running the deployed policy verbatim, at its softmax temperature, with its five warmup rounds and its rolling context.

The deployed policy sits at $\varepsilon_{\mathrm{pre}} = 0.23$, and the same table shows what that number is made of. The exact posterior concentrates, placing on average 0.98 of its mass on one arm, so total variation against it is close to the indicator that a different arm was chosen, and the measured selection disagreement of 0.22 accounts for nearly all of it. Total variation counts disagreements and regret counts what the disagreement cost: across the phases $t < 50$, $50 \le t < 150$ and $t \ge 150$ it reads 0.38, 0.19 and 0.18, while the per round regret of the very same policy falls from 0.380 to 0.072 and 0.071, so on a nearly tied intervention pool the two come apart. This is why R_{pre} is stated in gap weighted form, which is the same proof stopped one inequality earlier (Proposition 4): over the 195 policy rounds Theorem 2 certifies 28.5 against a measured 27.8, a slack of 2.3%, and it does so against a strong benchmark, exact posterior sampling paying 0.002 per round, so the whole bound is carried by the deviation term.

G VALIDATION STREAMS

The released checkpoint, the deployment constants (c, ε_0) and the pretraining collection rule were settled on validation streams disjoint from every reported experiment. Generative configurations are drawn from the pretraining prior instead of the evaluation robustness grid, horizons are $\{150, 250, 350, 500\}$ and share no value with any reported table, graph topologies are random geometric and power law cluster families absent from pretraining, real networks are excluded so that Table 29 stays an independent transfer evaluation, and seeds are disjoint from every evaluation stream. Scored there by the quantity the paper reports, total dynamic regret over 48 streams of 20 episodes with the trackers given their oracle windows exactly as in the benchmark tables, the released checkpoint records 203.3 against 227.0 for the strongest tracker.