GRAPH-DISCOVERING FOR NETWORKED RESTLESS MULTI-ARMED BANDITS WITH NONLINEAR DELAYED PROPAGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study nonlinear graph-discovering restless multi-armed bandits, where budgeted interventions propagate through an unknown sparse directed graph and induce delayed, persistent, and nonlinear effects on future arm dynamics. Graph identification is challenging because reward-seeking actions may provide insufficient information, particularly when nonlinear responses are saturated. We propose *Nonlinear Graph-Discovery Optimistic Intervention* (N-GDOI), which couples derivative-weighted Fisher exploration and localized sparse nonlinear graph estimation with optimistic stochastic rollout planning. Fisher-aware exploration probes uncertain graph directions in informative response regions, while planning over graph confidence sets accounts for both model uncertainty and nonlinear trajectory noise. We establish high-probability graph-recovery guarantees under adaptively collected data and prove stability of stochastic rollout values under graph perturbations. N-GDOI achieves $\widetilde{\mathcal{O}}(T^{2/3})$ truncated rollout regret over time horizon T under scheduled Fisher-informative exploration, improving to $\widetilde{\mathcal{O}}(\sqrt{T})$ when reward-seeking actions provide persistent Fisher excitation. The bounds explicitly separate statistical graph uncertainty from finite graph-candidate, Monte-Carlo rollout, and approximate action-search errors. Experiments show improved graph recovery and long-horizon reward, with the largest gains under partially saturated nonlinear propagation.

1 INTRODUCTION

Restless multi-armed bandits (RMABs) provide a canonical framework for sequential resource allocation under limited intervention capacity. An RMAB consists of N arms, each modeled as a controlled Markov decision process with a local state, a binary active or passive action, a transition kernel, and a one-step reward. At each round, the learner may activate at most $M < N$ arms, while both active and passive arms continue to evolve. The goal is to maximize the expected discounted cumulative reward over a finite horizon. Classical RMABs assume that, conditional on their local states and actions, the arms evolve independently and are coupled only through the shared budget. This factorization enables arm-wise relaxations and underlies Whittle-index and online index-learning methods (Tekin & Liu, 2012; Nakhleh et al., 2021; Wang et al., 2023; Chen et al., 2026).

This conditional-independence assumption can be restrictive in intervention domains with network spillovers. Consider community-level public-health outreach (Liang et al., 2025): a planner can provide intensive support to only a few communities each week, yet an intervention in one community may subsequently affect others through mobility, social contact, or information diffusion. Such effects are often delayed, because an intervention first changes behavioral or health-related exposure, which then influences future states and rewards. They may also be nonlinear: responses can be weak below an activation threshold, substantial over an intermediate range, and saturated after repeated exposure (Centola, 2010). Moreover, the network governing these spillovers is typically unknown and cannot be queried directly (Wilder et al., 2018).

We study this setting as a *nonlinear graph-discovering restless bandit*. Budgeted interventions generate state-dependent source signals that propagate through an unknown sparse directed graph. Their effects accumulate in arm-specific exposures, persist across rounds, and nonlinearly alter future

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.

local-state transitions and rewards. Persistence and state feedback turn direct graph interactions into delayed multi-round effects: an intervention changes future exposure, the resulting state changes future source signals, and those signals propagate again through the network. Consequently, an arm's intervention value cannot be determined from its local state alone, and the unknown graph is part of the controlled dynamics rather than an auxiliary source of side information.

Existing RMAB methods do not directly address this combination of unknown propagation and nonlinear dynamic coupling. Online RMAB algorithms learn unknown local transition kernels but generally retain conditionally independent arm dynamics (Wang et al., 2023; Chen et al., 2026). Networked and coupled RMAB models allow externalities, global rewards, or combinatorial interactions, but typically assume that the interaction graph, coupling features, or transition model is known (Ou et al., 2022; Herlihy & Dickerson, 2023; Zhang et al., 2025; Luan et al., 2025; Xu et al., 2025; Raman et al., 2024). Their primary challenge is planning under a given coupled model, whereas our setting requires learning the coupling mechanism while simultaneously controlling the system. The graph in our model also has a different role from those in graph bandits and static causal bandits. Graph-feedback, graph-triggered, and mobility-constrained bandits use edges to encode side observations, evolution counters, or feasible movements (Genalti et al., 2026; Rege et al.; Gong & Zhang, 2025; Qi et al., 2026; Eldowa et al., 2023; Paschalidis et al., 2026). These edges do not represent an unknown weighted mechanism through which an intervention changes other arms' future transition laws. Unknown-graph causal bandits learn structural relations, but primarily under static structural causal models and one-step intervention samples (Zhao & Zhou, 2025; Durand et al., 2025; Peng et al., 2024; Yan & Tajer, 2024; Elahi et al., 2024). Taken together, these lines of work leave open the problem of learning an unknown propagation graph from a single, adaptively controlled restless trajectory with persistent nonlinear spillovers.

This problem couples statistical identification, exploration, and planning. The interventions selected by the learner determine the source signals used to estimate the graph, so reward-seeking behavior may repeatedly visit a narrow region and leave important propagation directions unidentified. More subtly, source-signal diversity alone does not guarantee information under a nonlinear link: when the response is saturated, even substantial input variation can have little observable effect on exposure transitions. Finally, graph-estimation error affects rewards only after propagating through stochastic multi-step dynamics. Planning with a single estimated graph can therefore be overconfident, while replacing random trajectories by their means can distort action values under nonlinear dynamics. These difficulties lead to our central question:

> *Can an online learner discover an unknown directed spillover graph while making effective budgeted interventions under delayed, persistent, stochastic, and nonlinear network effects?*

We answer this question affirmatively under a reachable-information condition: feasible interventions must be able to generate sufficiently diverse source signals in regions where the nonlinear response remains informative. This requirement is fundamental, since permanently saturated systems do not contain enough information to distinguish alternative propagation graphs. We propose *Nonlinear Graph-Discovery Optimistic Intervention* (N-GDOI), an online model-based framework that connects graph identification with long-horizon control. For identification, N-GDOI uses derivative-weighted Fisher exploration to collect informative transitions and fits the incoming graph neighborhoods through localized sparse nonlinear regression. For decision-making, it constructs a confidence set of plausible graphs and performs optimistic stochastic multi-step rollouts. Finite graph candidate sets, Monte-Carlo simulation, and approximate combinatorial search make the planner implementable, while their errors remain explicit in the analysis. Our contributions are fourfold:

- **Problem formulation.** We introduce a graph-coupled RMAB with an unknown sparse directed propagation graph and persistent nonlinear exposures. The model captures delayed inter-arm effects in the transition dynamics, rather than feedback, mobility, or action-enablement relations.
- **Algorithm design.** We develop N-GDOI to jointly learn the propagation graph and optimize budgeted interventions. It combines Fisher-aware exploration and localized sparse nonlinear estimation with confidence-set optimism and stochastic rollout planning.
- **Theoretical analysis.** We establish high-probability graph-recovery guarantees under adaptive data collection and prove stability of stochastic rollout values under graph per-

turbations. N-GDOI achieves $\widetilde{\mathcal{O}}(T^{2/3})$ truncated rollout regret over time horizon T under scheduled Fisher exploration and $\widetilde{\mathcal{O}}(\sqrt{T})$ regret when reward-seeking actions provide persistent Fisher excitation. The bounds also retain graph-discretization, Monte-Carlo, and action-search errors.

- **Empirical validation.** Experiments under identity, sigmoid, and softplus propagation show that N-GDOI remains competitive in the linear special case and provides its largest gains under partially saturated nonlinear dynamics. Graph-recovery and ablation results further demonstrate the roles of Fisher exploration, graph optimism, and stochastic rollouts.

2 Problem Formulation

We study a *nonlinear graph-discovering restless bandit* in which budgeted interventions generate state-dependent signals that propagate through an unknown sparse directed graph. The resulting effects persist over time, nonlinearly influence future states and rewards, and must be learned from adaptively collected exposure transitions.

Arms and interventions. There are N arms indexed by $[N]$. At round $t \in \{0, \dots, T-1\}$, arm i has state $s_i^t \in \mathcal{S}_i$, where $\mathcal{S}_i$ is finite, and the global state is $s^t = (s_1^t, \dots, s_N^t)$. The learner chooses $a^t \in \mathcal{A}_M := \{a \in \{0,1\}^N : \|a\|_1 \le M\}$, where $M < N$ and $a_i^t = 1$ activates arm i.

Unknown graph and source signals. Interactions are represented by an unknown nonnegative directed propagation matrix $P^\star \in \mathbb{R}_+^{N \times N}$, where $P_{ij}^\star$ is the influence of source arm i on target arm j. Each incoming column satisfies $\|P_{\cdot j}^\star\|_0 \le d$ and $\|P_{\cdot j}^\star\|_1 \le P_{\max}$. We exclude self-loops by setting $P_{ii}^\star = 0$; allowing them does not otherwise change the model. At round t, arm i emits $x_i^t = g_i(s_i^t) + a_i^t \Delta_i(s_i^t)$, where g_i is its endogenous signal and Δ_i is the intervention-induced increment. Thus, passive arms may also affect the network. We assume $0 \le x_i^t \le x_{\max}$ and $\|x^t\|_1 \le B_x$; setting $g_i \equiv 0$ recovers intervention-only propagation.

Nonlinear exposure dynamics. Each arm has an exposure $z_i^t \in \mathcal{Z}_i$, summarizing accumulated network influence. After a^t determines x^t, exposure evolves as

$$z_j^{t+1} = \sigma\big(\alpha z_j^t + \langle P_{\cdot j}^\star, x^t \rangle\big) + \eta_j^{t+1}, \qquad j \in [N], \tag{1}$$

where $\alpha \in [0,1)$ controls persistence and σ is a known nondecreasing link. The term αz_j^t carries past spillovers into future rounds, while $\langle P_{\cdot j}^\star, x^t \rangle$ is the new network input. Conditional on the post-action history, the innovation η_j^{t+1} is mean-zero and ν_z-sub-Gaussian, with support chosen so that $z_j^{t+1} \in \mathcal{Z}_j$. Its conditional law is graph-independent and available to the rollout simulator. The deterministic model sets $\eta_j^{t+1} = 0$, and the linear model additionally sets $\sigma(u) = u$.

Restless dynamics and rewards. Arms evolve under both active and passive actions. Conditional on the realized exposure, their transitions factorize: $\mathbb{P}^\star(s^{t+1} \mid s^t, a^t, z^{t+1}) = \prod_{i=1}^N \mathbb{P}_i^\star(s_i^{t+1} \mid s_i^t, a_i^t, z_i^{t+1})$. Thus, cross-arm dependence is induced by the shared propagation process rather than an unrestricted joint transition kernel. We treat the local kernels $\{\mathbb{P}_i^\star\}_{i=1}^N$ as known, so the unknown structural parameter is $P^\star$. The expected one-step reward is $r(s,z,a) = \sum_{i=1}^N r_i(s_i, a_i, z_i)$, and the observed reward is $R^t = r(s^t, z^t, a^t) + \xi^t$, where ξ^t is conditionally mean-zero and sub-Gaussian. The local reward functions are known. A non-anticipating policy π maximizes

$$V^\pi(s^0, z^0) = \mathbb{E}^\pi\left[\sum_{t=0}^{T-1} \gamma^t R^t\right], \qquad \gamma \in (0,1). \tag{2}$$

Observation protocol. We consider the instrumented setting. Before acting at round t, the learner observes (s^t, z^t) and the past history; after selecting a^t, it observes (R^t, s^{t+1}, z^{t+1}). Let $\mathcal{F}_t$ and $\mathcal{F}_t^+ = \sigma(\mathcal{F}_t, a^t)$ denote the pre-decision and post-action filtrations. Since x^t is determined by (s^t, a^t), it is $\mathcal{F}_t^+$-measurable. Hence, the graph-learning covariates are adaptively generated but predictable with respect to the next exposure innovation.

3 N-GDOI ALGORITHM

We propose *Nonlinear Graph-Discovery Optimistic Intervention* (N-GDOI), an online model-based algorithm that couples graph identification with long-horizon intervention planning. N-GDOI repeatedly performs three operations. It first collects informative exposure transitions through a short burn-in phase and scheduled exploration. It then estimates the sparse propagation graph and maintains a confidence set of statistically plausible models. On reward-seeking rounds, it selects an intervention by optimizing stochastic multi-step rollout values over this confidence set. Algorithm 1 summarizes the complete procedure.

The central difficulty is that source-signal diversity alone does not ensure graph identifiability. A transition is informative only when its source signal excites uncertain graph directions and the nonlinear response has a non-negligible local derivative. N-GDOI therefore weights exploration by plug-in Fisher information. Moreover, because nonlinear propagation does not generally commute with expectation, the planner simulates complete stochastic trajectories rather than propagating only mean exposures.

Fisher-aware graph learning. Let $\mathcal{T}_{\mathrm{exp}} \subseteq [T]$ be a predictable exploration schedule that contains the first t_{burn} rounds and satisfies $n_{\mathrm{exp}}(t) := |\mathcal{T}_{\mathrm{exp}} \cap \{0, \dots, t-1\}| \asymp t^{2/3}$. The burn-in procedure collects diverse transitions and returns a localizer $\widetilde{P}_{\cdot j}$ for each incoming graph column. Subsequent estimates are restricted to $\mathcal{W}_j^{\mathrm{loc}} := \{w \in \mathbb{R}_+^N : \|w - \widetilde{P}_{\cdot j}\|_1 \le r_{\mathrm{loc}}/2\}$. The localizer guarantee $\|\widetilde{P}_{\cdot j} - P^\star_{\cdot j}\|_1 \le r_{\mathrm{loc}}/2$ keeps the true graph feasible and confines nonlinear estimation to a locally identifiable region. This localization is unnecessary when the estimation loss is globally restricted convex.

For a feasible action a, let $x_t(a) = g(s^t) + a \odot \Delta(s^t)$ and define $\widehat{\omega}_{j,t}(a) := [\sigma'(\alpha z_j^t + \langle \widehat{P}_{\cdot j,t}, x_t(a) \rangle)]^2$. On an exploration round, N-GDOI favors actions with high derivative-weighted information gain:

$$\mathsf{IG}_t(a) = \sum_{j=1}^{N} \log\Big[1 + \widehat{\omega}_{j,t}(a)\, x_t(a)^\top \widehat{V}_{j,t}^{-1} x_t(a)\Big], \quad a_t^{\mathrm{exp}} \in \arg\max_{a \in \mathcal{A}_M} \mathsf{IG}_t(a). \tag{3}$$

Here $\widehat{V}_{j,t}$ is a regularized plug-in Fisher proxy. The leverage term identifies poorly explored source-signal directions, while $\widehat{\omega}_{j,t}(a)$ discounts actions predicted to place response in a saturated region.

The routine APPROXFISHERORACLE approximately solves Eq. 3 using greedy or beam search under the budget constraint. During initial burn-in, before a reliable nonlinear estimate is available, it uses the current preliminary estimate and a derivative floor to avoid degenerate scores.

Let $\mathcal{E}_t = \mathcal{T}_{\mathrm{exp}} \cap \{0, \dots, t-1\}$ be the observed exploration rounds. After burn-in, each incoming graph column is estimated by localized sparse nonlinear regression:

$$\widehat{P}_{\cdot j,t} \in \arg\min_{w \in \mathcal{W}_j^{\mathrm{loc}}} \left\{ \frac{1}{n_{\mathrm{exp}}(t) \vee 1} \sum_{\tau \in \mathcal{E}_t} \big[z_j^{\tau+1} - \sigma\big(\alpha z_j^\tau + \langle w, x^\tau \rangle\big)\big]^2 + \lambda_t \|w\|_1 \right\}, \tag{4}$$

where $\lambda_t \asymp \nu_z L_\sigma B_x \sqrt{\log(2N^2T/\delta)/(n_{\mathrm{exp}}(t) \vee 1)}$. The Fisher proxy is $\widehat{V}_{j,t} = \lambda_{\mathrm{rid}} I + \sum_{\tau \in \mathcal{E}_t} \widehat{\omega}_{j,\tau,t} x^\tau (x^\tau)^\top$, with $\widehat{\omega}_{j,\tau,t} = [\sigma'(\alpha z_j^\tau + \langle \widehat{P}_{\cdot j,t}, x^\tau \rangle)]^2$. We use $\lambda_{\mathrm{rid}} > 0$ for Fisher regularization, distinguishing it from the sparse-regression penalty λ_t.

The resulting confidence set is

$$\mathcal{C}_t = \Big\{ P \in \mathcal{P} : \|P_{\cdot j} - \widehat{P}_{\cdot j,t}\|_1 \le \varepsilon_t^{\mathrm{conf}},\ \forall j \in [N] \Big\}, \tag{5}$$

where $\varepsilon_t^{\mathrm{conf}} = \widetilde{\mathcal{O}}(d/\sqrt{n_{\mathrm{exp}}(t)})$. Its calibration ensures that $P^\star \in \mathcal{C}_t$ uniformly with high probability. The columnwise geometry is chosen because the rollout sensitivity is controlled by $\|P - \widetilde{P}\|_{1,\infty} = \max_j \|P_{\cdot j} - \widetilde{P}_{\cdot j}\|_1$.

Optimistic stochastic planning. On a reward-seeking round, N-GDOI selects the first action of an optimistic H-step rollout:

$$a_t^{\mathrm{opt}} \in \arg\max_{a \in \mathcal{A}_M} \sup_{P \in \mathcal{C}_t} Q_{t,H}(a; P), \tag{6}$$

Algorithm 1 N-GDOI: Nonlinear Graph-Discovery Optimistic Intervention

Require: Horizon T, budget M, burn-in length t_{burn}, rollout horizon H, rollout counts $\{K_t\}$, exploration schedule $\mathcal{T}_{\mathrm{exp}}$, continuation policy μ

Initialize $\mathcal{E}_0 \leftarrow \emptyset$, $\widehat{P}_0 \leftarrow P^{(0)}$, and $\widehat{V}_{j,0} \leftarrow \lambda_{\mathrm{rid}} I$ for all $j \in [N]$
for $t = 0, \ldots, t_{\mathrm{burn}} - 1$ **do**
  Observe (s^t, z^t) and set $a^t \leftarrow \textsc{ApproxFisherOracle}(s^t, z^t, \widehat{P}_t, \widehat{V}_t)$
  Apply a^t and observe (R^t, s^{t+1}, z^{t+1})
  Set $\mathcal{E}_{t+1} \leftarrow \mathcal{E}_t \cup \{t\}$ and update the preliminary graph and Fisher estimates
end for
Construct localizers $\{\widetilde{P}_{\cdot j}\}_{j=1}^{N}$ from burn-in transitions
for $t = t_{\mathrm{burn}}, \ldots, T - 1$ **do**
  Observe (s^t, z^t)
  Estimate $\widehat{P}_t$ by equation 4; update $\widehat{V}_t$ and construct $\mathcal{C}_t$ by equation 5
  if $t \in \mathcal{T}_{\mathrm{exp}}$ **then**
    Set $a^t \leftarrow \textsc{ApproxFisherOracle}(s^t, z^t, \widehat{P}_t, \widehat{V}_t)$
  else
    Construct candidate graphs $\mathcal{G}_t \subseteq \mathcal{C}_t$
    Set $a^t \leftarrow \textsc{ApproxOptimisticPlanner}(s^t, z^t, \mathcal{G}_t, H, K_t, \mu)$
  end if
  Apply a^t and observe (R^t, s^{t+1}, z^{t+1})
  Set $\mathcal{E}_{t+1} \leftarrow \mathcal{E}_t \cup \{t\}$ if $t \in \mathcal{T}_{\mathrm{exp}}$, and $\mathcal{E}_{t+1} \leftarrow \mathcal{E}_t$ otherwise
end for

where $Q_{t,H}(a; P)$ is the expected discounted return obtained by taking a first and following a feasible continuation policy μ thereafter. Rollouts sample the graph-independent exposure innovations and known local transition kernels. This accounts for nonlinear stochastic effects that a mean-state recursion may miss.

Direct optimization of Eq. 6 is generally intractable. The routine ApproxOptimisticPlanner first constructs a finite set $\mathcal{G}_t \subseteq \mathcal{C}_t$ containing the point estimate and sparse boundary perturbations along uncertain directions. It then uses K_t Monte-Carlo rollouts to evaluate graph–action pairs and beam search to build a budget-feasible intervention set. Its approximation is characterized by the graph covering radius ρ_t, Monte-Carlo error $b_{t,K}$, and action-search residual $\varepsilon_{\mathrm{opt},t}$, all retained explicitly in the regret analysis.

Algorithm 1 separates the burn-in, graph-learning, and planning roles without requiring separate estimators for exploration and exploitation. The conservative theoretical version fits the graph using only transitions indexed by $\mathcal{E}_t$, whose information content is explicitly controlled. When reward-seeking transitions also provide persistent Fisher excitation, the all-transition variant may use them after burn-in and attains the faster rate in Corollary 1.

4 Theoretical Analysis

We analyze how N-GDOI converts informative interventions into graph accuracy and, subsequently, into decision quality. The argument has three steps. Fisher-aware exploration first supplies curvature for sparse nonlinear graph estimation. A stochastic simulation lemma then controls how graph error perturbs multi-step rollout values. Finally, optimism over the graph confidence set yields sublinear rollout regret while retaining graph-net, Monte-Carlo, and approximate action-search errors.

Truncated rollout benchmark. For graph $P \in \mathcal{P}$, let $Q_{t,H}(a; P)$ be the expected H-step discounted value obtained by taking action a at round t and following feasible continuation policy μ thereafter. The true-graph H-step rollout oracle selects $a_t^\star \in \arg\max_{a \in \mathcal{A}_M} Q_{t,H}(a; P^\star)$. We evaluate performance using the *truncated rollout regret*, defined as

$$\mathfrak{R}_T^{(H)} = \sum_{t=0}^{T-1} \left[Q_{t,H}(a_t^\star; P^\star) - Q_{t,H}(a^t; P^\star) \right]. \tag{7}$$

This benchmark evaluates intervention selection under the same rollout horizon and continuation policy used by the planner. It captures delayed network effects without assuming that a general constrained restless MDP can be solved exactly.

The benchmark differs from the unrestricted optimal-policy criteria used in known-model coupled RMAB planning (Xu et al., 2025; Raman et al., 2024; Zhang et al., 2025). Our result isolates the additional loss caused by graph uncertainty and approximate rollout planning; it does not claim regret against an exact optimal policy for the full graph-coupled MDP.

Adaptive sparse graph recovery. Let $\mathcal{E}_t$ be the explicit exploration rounds observed before t and $n_{\exp}(t) = |\mathcal{E}_t|$. For target arm j, write $\theta_j^\tau = \alpha z_j^\tau + \langle P_{\cdot j}^\star, x^\tau \rangle$. The estimator in equation 4 uses the predictable set $\mathcal{E}_t$, with $\lambda_t \asymp \nu_z L_\sigma B_x \sqrt{\log(2N^2T/\delta)/(n_{\exp}(t) \vee 1)}$.

Theorem 1 (Adaptive sparse graph recovery)**.** *Suppose the regularity, localization, and Fisher-curvature conditions in Appendix B hold. Then, after burn-in, with probability at least* $1-\delta$*, simultaneously for every* $j \in [N]$ *and* $t \in [T]$*,*

$$\begin{aligned} \|\widehat{P}_{\cdot j,t} - P_{\cdot j}^\star\|_2 &\le \frac{C_1 \nu_z L_\sigma B_x}{\kappa_{\mathrm{FI}}} \sqrt{\frac{d \log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}}, \\ \|\widehat{P}_{\cdot j,t} - P_{\cdot j}^\star\|_1 &\le \frac{C_2 d \nu_z L_\sigma B_x}{\kappa_{\mathrm{FI}}} \sqrt{\frac{\log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}}. \end{aligned} \tag{8}$$

The proof combines martingale score concentration with localized sparse restricted curvature. Adaptivity is permitted because x^τ is determined before the next exposure innovation. Localization is needed because nonlinear least-squares losses need not be globally convex.

The theorem calibrates $\mathcal{C}_t$ in equation 5. With a sufficiently large constant, $P^\star \in \mathcal{C}_t$ uniformly with high probability and $\mathrm{diam}_{1,\infty}(\mathcal{C}_t) \le 2\varepsilon_t^{\mathrm{conf}}$. Under additional local irrepresentability and beta-min conditions, thresholding also recovers the directed support; see Appendix C.

Remark 1. Unknown-graph causal bandits establish structure-learning and regret results for static linear or nonparametric structural causal models (Peng et al., 2024; Yan & Tajer, 2024; Elahi et al., 2024; Zhao & Zhou, 2025). Theorem 1 instead concerns a single adaptively controlled restless trajectory: observations are temporally dependent, the graph acts through persistent exposures, and identifiability varies with the local derivative of the nonlinear link. The result therefore addresses a different statistical object from static intervention-effect estimation.

Fisher excitation. The effective sample size is the number of informative transitions, rather than the total number of rounds. Define the oracle Fisher design $V_{j,t}^\star = \lambda_{\mathrm{rid}} I + \sum_{\tau \in \mathcal{E}_t} [\sigma'(\theta_j^\tau)]^2 x^\tau (x^\tau)^\top$.

Theorem 2 (Excitation-induced Fisher curvature)**.** *Suppose the scheduled exploration sequence satisfies the reachable sparse Fisher-design condition in Appendix B. Then, with probability at least* $1-\delta$*, for every* $j \in [N]$ *and every* t *with* $n_{\exp}(t) \ge C d \log(2NT/\delta)$*,*

$$\inf_{v \in \mathcal{K}_j \setminus \{0\}} \frac{v^\top V_{j,t}^\star v}{n_{\exp}(t) \|v\|_2^2} \ge \frac{\kappa_{\exp}}{2}. \tag{9}$$

Consequently, Theorem 1 applies with $\kappa_{\mathrm{FI}} \asymp \kappa_{\exp}$*.*

Remark 2 (What exploration can guarantee)**.** The information-gain objective in equation 3 is an implementable surrogate for constructing an informative design; the theorem does not assert that every greedy maximizer automatically satisfies sparse restricted excitation. Feasible interventions must generate both diverse source signals and non-saturated responses. If all reachable responses have nearly zero derivative, graph recovery is statistically impossible.

From graph error to rollout error. Graph recovery improves control only if rollout values vary continuously with the graph. The next result couples complete stochastic trajectories under two graphs using common exposure innovations and coupled local transitions.

Theorem 3 (Nonlinear stochastic rollout stability)**.** *Under the regularity conditions in Appendix B, for every finite* H *there exists* $C_H < \infty$ *such that, for all feasible* a *and* $P, \widetilde{P} \in \mathcal{P}$*,*

$$\left| Q_{t,H}(a; P) - Q_{t,H}(a; \widetilde{P}) \right| \le C_H \|P - \widetilde{P}\|_{1,\infty}. \tag{10}$$

If the coupled exposure–state recursion is contractive, C_H is uniformly bounded in H.

Remark 3 (Relation to coupled-RMAB planning)**.** Known-network RMAB methods focus on allocating actions under a specified externality, coupling matrix, global reward, or cascade model (Ou et al., 2022; Herlihy & Dickerson, 2023; Luan et al., 2025; Xu et al., 2025; Raman et al., 2024; Zhang et al., 2025). Theorem 3 supplies the missing statistical link for an unknown dynamical graph: it converts graph-confidence diameter into multi-step value uncertainty.

Regret with approximate planning. Let ρ_t be the covering radius of $\mathcal{G}_t \subseteq \mathcal{C}_t$, K_t the rollout count, and $\varepsilon_{\mathrm{opt},t}$ the certified action-search residual. Define $R_H = R_{\max}(1-\gamma^H)/(1-\gamma)$ and $b_{t,K} = R_H\sqrt{\log(2TL_t/\delta)/(2K_t)}$, where L_t is the number of graph–action pairs covered by the concentration event.

Theorem 4 (Per-round optimistic planning error)**.** *Suppose $P^\star \in \mathcal{C}_t$. On each exploitation round, with probability at least $1-\delta/T$,*

$$Q_{t,H}(a_t^\star; P^\star) - Q_{t,H}(a^t; P^\star) \le C_H \operatorname{diam}_{1,\infty}(\mathcal{C}_t) + C_H\rho_t + 2b_{t,K} + \varepsilon_{\mathrm{opt},t}. \tag{11}$$

The four terms respectively represent statistical graph uncertainty, finite graph-candidate approximation, Monte-Carlo error, and combinatorial action-search error. Thus, the guarantee does not hide practical planning costs behind an exact optimization oracle.

Theorem 5 (Regret under scheduled Fisher exploration)**.** *Suppose $n_{\exp}(t) \asymp t^{2/3}$ and Theorem 2 applies. If $\rho_t = \widetilde{\mathcal{O}}(t^{-1/3})$, $K_t = \widetilde{\Omega}(t^{2/3})$, and $\varepsilon_{\mathrm{opt},t} = \widetilde{\mathcal{O}}(t^{-1/3})$, then, with probability at least $1-\delta$,*

$$\mathfrak{R}_T^{(H)} \le \widetilde{\mathcal{O}}\left(\left[R_H + \frac{C_H d \nu_z L_\sigma B_x}{\kappa_{\exp}} + C_H c_\rho + \frac{R_H}{\sqrt{c_K}} + c_{\mathrm{opt}}\right] T^{2/3}\right). \tag{12}$$

Corollary 1 (Fast rate under persistent excitation)**.** *Suppose reward-seeking transitions satisfy persistent Fisher restricted excitation after burn-in, explicit exploration occurs on at most $O(\sqrt{T})$ rounds, and $\rho_t = \widetilde{\mathcal{O}}(t^{-1/2})$, $K_t = \widetilde{\Omega}(t)$, and $\varepsilon_{\mathrm{opt},t} = \widetilde{\mathcal{O}}(t^{-1/2})$. Then $\mathfrak{R}_T^{(H)} \le \widetilde{\mathcal{O}}\left(\sqrt{T}\right)$.*

Remark 4 (Interpretation and comparison of rates)**.** The scheduled rate balances $\Theta(T^{2/3})$ deliberate exploration rounds against graph uncertainty of order $\widetilde{\mathcal{O}}(t^{-1/3})$. The faster rate requires reward-seeking actions to supply Fisher information at linear frequency; optimism alone does not imply this condition. Although $T^{2/3}$ and $\sqrt{T}$ rates also appear in graph-triggered, nonstationary, causal, and independent-arm bandits (Genalti et al., 2026; Migali et al., 2026; Yan & Tajer, 2024; Wang et al., 2023), these rates are not directly comparable: those works use different graph semantics, observation models, and regret benchmarks. Here the rate accounts jointly for learning a nonlinear dynamical spillover graph and planning under its uncertainty.

5 Experiments

We evaluate whether N-GDOI improves sequential intervention quality and online graph discovery under nonlinear network dynamics. We ask three questions: (i) Does N-GDOI improve long-horizon return over independent-arm and network-aware baselines? (ii) Does Fisher-aware exploration improve recovery of the directed propagation graph? (iii) Which components of N-GDOI account for its performance?

Experimental setup. We evaluate on synthetic networked RMABs ($N = 30$, $M = 5$, $T = 500$, $\alpha = 0.4$) featuring a directed 4-community SBM graph ($d = 4$, no self-loops, weights in $[0.08, 0.28]$) and identity, sigmoid (partially saturated), and softplus links. We compare N-GDOI against Myopic, UCWhittle (Wang et al., 2023), GINO-Q (Chen et al., 2026), Linear-Graph (a misspecified linear estimator with a nonlinear planner), GRETA (Herlihy & Dickerson, 2023), and our multi-hop extension MH-GRETA. Graph-recovery evaluation includes only methods with explicit online graph estimators (Appendix D). All methods share 10 environment seeds. Performance is measured by post-burn-in discounted return ($\sum_{t=0}^{T-1}\gamma^t R^t$), directed support F1 (threshold 0.05), and Frobenius error ($\|\widehat{P} - P^\star\|_F$). Tables report mean $\pm$ standard error, and figures depict 95% confidence intervals.

Decision performance. Figure 1 shows that the relative benefit of N-GDOI depends on the propagation mechanism. Under identity propagation, where the derivative is constant and the linear model is correctly specified, N-GDOI is close to LINEAR-GRAPH and MH-GRETA; at $\alpha = 0.4$, its gain over MH-GRETA is approximately 1.0%. Under sigmoid propagation, the gain is 10.5% at $\alpha = 0.4$ and increases to 28.2% at $\alpha = 0.8$. This is consistent with the role of Fisher exploration: as persistence increases, reward-driven policies more frequently visit low-derivative regions, whereas N-GDOI explicitly seeks informative reachable regimes. Softplus produces an intermediate pattern, with a 7.3% gain over MH-GRETA at $\alpha = 0.4$. These trends agree qualitatively with the dependence of Theorems 1 and 5 on Fisher curvature.

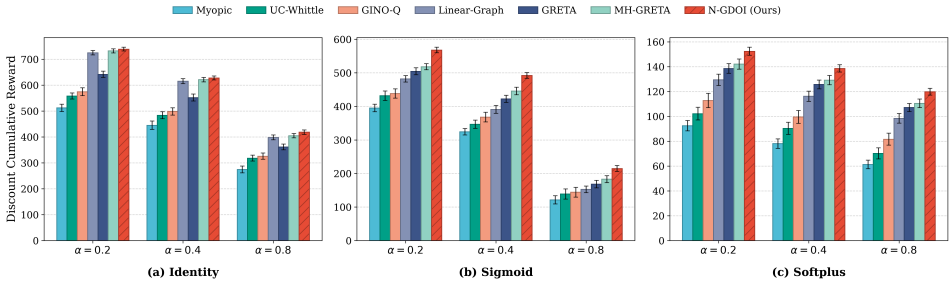


Figure 1: **Return across links and persistence levels.** N-GDOI has its largest advantage under partially saturated sigmoid propagation. Error bars show 95% confidence intervals over 10 seeds.

Table 1: **Graph recovery across links and persistence levels.** Each entry reports directed support F1 / Frobenius error; higher F1 and lower error are better. Results are mean ± standard error.

Method	Identity			Sigmoid			Softplus		
	$\alpha = .2$	$\alpha = .4$	$\alpha = .8$	$\alpha = .2$	$\alpha = .4$	$\alpha = .8$	$\alpha = .2$	$\alpha = .4$	$\alpha = .8$
Linear-Graph	.88±.04/.45±.05	.81±.05/.55±.08	.65±.07/.88±.12	.52±.09/2.13±.22	.41±.08/2.75±.28	.25±.06/3.42±.35	.72±.06/1.27±.15	.63±.07/1.65±.18	.47±.08/2.20±.25
GRETA	.79±.05/.92±.10	.72±.06/1.19±.12	.55±.08/1.65±.18	.67±.07/1.85±.15	.58±.06/2.10±.20	.38±.07/2.85±.25	.76±.05/1.15±.12	.69±.06/1.43±.15	.51±.07/1.95±.22
MH-GRETA	.85±.04/.65±.08	.78±.05/.82±.10	.61±.06/1.24±.15	.72±.06/1.72±.14	.62±.05/2.05±.18	.42±.08/2.70±.22	.81±.04/1.04±.10	.73±.05/1.35±.14	.56±.06/1.80±.20
N-GDOI	**.89±.03/.42±.04**	**.82±.04/.51±.06**	**.67±.05/.87±.10**	**.95±.02/.45±.07**	**.94±.03/.58±.08**	**.84±.05/.92±.12**	**.93±.03/.48±.06**	**.87±.04/.62±.08**	**.74±.03/.95±.12**

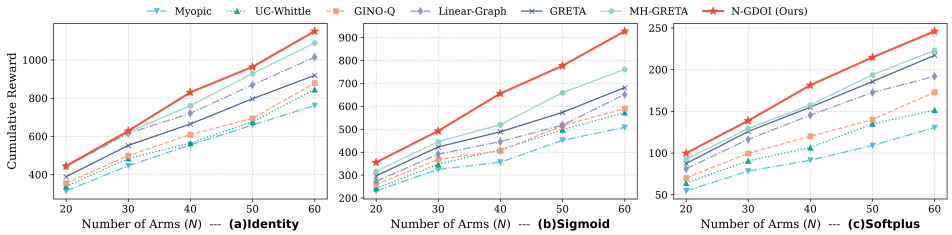
Online graph discovery. Table 1 reports structural and weight recovery jointly. Under identity propagation, N-GDOI and LINEAR-GRAPH are statistically close, providing a useful linear-model sanity check. At sigmoid $\alpha = 0.4$, N-GDOI increases F1 from 0.62 to 0.94 relative to MH-GRETA and reduces Frobenius error from 2.05 to 0.58, a reduction of approximately 72%. At $\alpha = 0.8$, its F1 remains 0.84, compared with 0.42 for MH-GRETA. Softplus is less strongly saturated, and accordingly the gap is smaller but remains substantial. These results support the theoretical mechanism: graph error decreases with derivative-weighted Fisher curvature, while feature variation alone need not be informative in low-derivative regions.

Table 2: **N-GDOI ablations.**

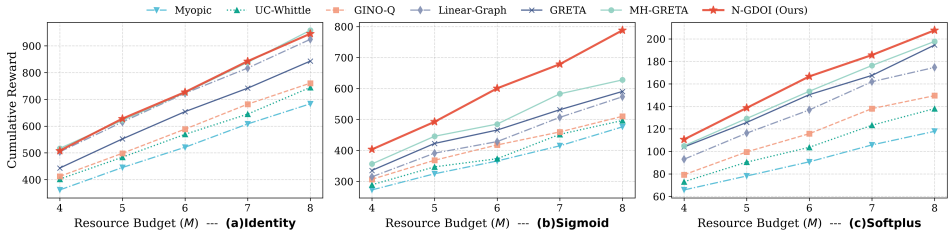
Variant	Return ↑	Frobenius error ↓	F1 ↑
Full N-GDOI	**492.2 ± 2.7**	**0.58 ± 0.08**	**0.94 ± 0.03**
w/o Fisher exploration	445.6 ± 4.0	2.10 ± 0.06	0.62 ± 0.02
Nominal planning	412.3 ± 4.7	0.85 ± 0.04	0.78 ± 0.02
Mean-state rollout	465.1 ± 3.3	0.62 ± 0.03	0.88 ± 0.01
Linear propagation	391.0 ± 3.4	2.75 ± 0.28	0.41 ± 0.05

Component ablations. Table 2 isolates the main design choices in the default sigmoid environment. Removing post-burn-in Fisher exploration reduces F1 from 0.94 to 0.62 and increases graph error from 0.58 to 2.10, confirming that reward-seeking data alone are insufficient in the partially saturated regime. Mean-state rollout reduces return by approximately 5.5%, consistent with the certainty-equivalence bias excluded by Theorem 3. Nominal planning also degrades graph recovery because exploitation decisions change the states encountered at later exploration rounds; thus, planning and identification remain coupled even when the estimator uses only designated exploration

transitions. The linear variant performs worst, demonstrating the combined cost of nonlinear model misspecification and inaccurate graph recovery.



(a) Varying N with $M = 5$.



(b) Varying M with $N = 30$.

Figure 2: **Sensitivity to network size and budget.** N-GDOI remains robust as N and M vary.

Sensitivity to scale and budget. Figure 2 evaluates the scalability of N-GDOI by varying the network size $N \in [20, 60]$ at a fixed budget $M = 5$ and varying the resource budget $M \in [4, 8]$ at a fixed $N = 30$ across three link functions. N-GDOI consistently outperforms all baselines across all tested configurations, exhibiting robust scalability as system dimensions increase. Notably, the performance gap between N-GDOI and network-aware baselines widens most significantly under the *sigmoid* propagation setting. This occurs because sigmoid dynamics introduce non-trivial saturation and low-derivative regions where unguided or linearly-approximated exploration fails, whereas N-GDOI's Fisher-aware exploration and local sparse recovery mechanism accurately identify and target informative transitions, maintaining high decision efficiency even in larger or more resource-constrained networks.

Additional evaluations and ablations. Additional results and extended evaluations are deferred to the appendices. Appendix D.3 presents continuous learning curves illustrating the temporal evolution of total return, graph recovery, and confidence set contraction. Appendix D.4 evaluates structural and algorithmic robustness under diverse graph topologies and varying exposure noise levels. Finally, Appendix D.5 reports empirical truncated rollout pseudo-regret relative to a high-precision true-graph oracle, validating the sublinear growth of N-GDOI compared to all baselines.

6 Conclusion

We study budgeted restless intervention under unknown directed network spillovers that are delayed, persistent, and nonlinear. We proposed N-GDOI, which combines Fisher-aware exploration and sparse nonlinear graph estimation with confidence-set optimism and stochastic multi-step planning. Our analysis establishes adaptive graph-recovery guarantees, connects graph error to rollout-value error, and gives $\widetilde{\mathcal{O}}(T^{2/3})$ truncated rollout regret, improving to $\widetilde{\mathcal{O}}(\sqrt{T})$ under persistent Fisher excitation. Experiments show that N-GDOI remains competitive in the linear special case and is particularly effective under partially saturated nonlinear dynamics. Our results focus on observed exposures, known local transition kernels, and reachable informative regimes. Extending the framework to latent exposures, jointly unknown local dynamics, and decentralized or time-varying propagation graphs provides a natural direction for future work.

REFERENCES

Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996): 1194–1197, 2010.

Gongpu Chen, Soung Chang Liew, and Deniz Gündüz. Gino-q: Learning an asymptotically optimal index policy for restless multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20032–20040, 2026.

Jean Durand, Yashas Annadani, Stefan Bauer, and Sonali Parbhoo. Causal bayesian optimization with unknown graphs. *arXiv preprint arXiv:2503.19554*, 2025.

Muhammad Qasim Elahi, Mahsa Ghasemi, and Murat Kocaoglu. Partial structure discovery is sufficient for no-regret learning in causal bandits. *Advances in Neural Information Processing Systems*, 37:109066–109100, 2024.

Khaled Eldowa, Emmanuel Esposito, Tom Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 36:46122–46133, 2023.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto Maria Metelli. Bridging rested and restless bandits with graph-triggering: Rising and rotting. *Journal of Machine Learning Research*, 27(151):1–46, 2026.

Xueping Gong and Jiheng Zhang. Efficient graph bandit learning with side-observations and switching constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16871–16879, 2025.

Christine Herlihy and John P Dickerson. Networked restless bandits with positive externalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11997–12004, 2023.

Biyonka Liang, Lily Xu, Aparna Taneja, Milind Tambe, and Lucas Janson. Context in public health for underserved communities: A bayesian approach to online restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28195–28203, 2025.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.

Yao Luan, Ni Mu, and Qing-Shan Jia. Addressing coupling in restless multi-armed bandits by finetuning whittle index. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pp. 2456–2461. IEEE, 2025.

Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Cristiano Migali, Marco Mussi, Gianmarco Genalti, and Alberto Maria Metelli. Tightening regret lower and upper bounds in restless rising bandits. *Advances in Neural Information Processing Systems*, 38:18437–18478, 2026.

Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I Hou, Srinivas Shakkottai, et al. Neurwin: Neural whittle index network for restless bandits via deep rl. *Advances in Neural Information Processing Systems*, 34:828–839, 2021.

Han-Ching Ou, Christoph Siebenbrunner, Jackson Killian, Meredith B Brooks, David Kempe, Yevgeniy Vorobeychik, and Milind Tambe. Networked restless multi-armed bandits for mobile interventions. *arXiv preprint arXiv:2201.12408*, 2022.

Phevos Paschalidis, Aryan Naveen, Runyu Zhang, and Na Li. Cooperative multi-agent graph bandits: Ucb algorithm and regret analysis. *IEEE Transactions on Automatic Control*, 2026.

Chen Peng, Di Zhang, and Urbashi Mitra. Graph identification and upper confidence evaluation for causal bandits with linear models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7165–7169. IEEE, 2024.

Han Qi, Fei Guo, Li Zhu, and Qiaosheng Zhang. Graph feedback bandits on similar arms: With and without graph structures. *IEEE Transactions on Information Theory*, 2026.

Naveen Raman, Zheyuan R Shi, and Fei Fang. Global rewards in restless multi-armed bandits. *Advances in Neural Information Processing Systems*, 37:24625–24658, 2024.

Amit Kiran Rege, Sourav Chakraborty, Lijun Chen, and Claire Monteleoni. Bandit learning on dynamic graphs. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

Sho Tsugawa and Hiroyuki Ohsaki. Exploring unknown social networks for discovering hidden nodes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 1937–1951, 2025.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic whittle index policy: Online learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10131–10139, 2023.

Bryan Wilder, Nicole Immorlica, Eric Rice, and Milind Tambe. Maximizing influence in an unknown social network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Guojun Xiong, Shufan Wang, and Jian Li. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems*, 35: 17911–17925, 2022.

Lily Xu, Bryan Wilder, Elias Khalil, and Milind Tambe. Reinforcement learning with combinatorial actions for coupled restless bandits. In *International Conference on Learning Representations*, volume 2025, pp. 61875–61896, 2025.

Zirui Yan and Ali Tajer. Linear causal bandits: Unknown graph and soft interventions. *Advances in Neural Information Processing Systems*, 37:23939–23987, 2024.

Hanmo Zhang, Zenghui Sun, and Kai Wang. Networked restless multi-arm bandits with reinforcement learning. *arXiv preprint arXiv:2512.06274*, 2025.

Yijia Zhao and Qing Zhou. Causal bandit over unknown graphs: Upper confidence bounds with backdoor adjustment. *arXiv preprint arXiv:2502.02020*, 2025.

Mengqiu Zhou, Vincent YF Tan, and Meng Zhang. Timely best arm identification in restless shared networks. *arXiv preprint arXiv:2603.15078*, 2026.

Appendix Contents

A Related Work **13**

B Notation and Assumptions **14**

B.1 Notation 14

B.2 Assumptions 14

C Proofs of the Main Theoretical Results **17**

C.1 Proof of Adaptive Score Concentration 17

C.2 Proof of the Basic Inequality and Sparse Cone 18

C.3 Proof of Theorem 1 19

C.4 Proof of Confidence-Set Validity 20

C.5 Proof of Matrix-Level Graph Recovery 20

C.6 Proof of Directed Support Recovery 21

C.7 Proof of Theorem 2 21

C.8 Proof of Theorem 3 23

C.9 Proof of Graph-Net Approximation 25

C.10 Proof of Uniform Monte-Carlo Concentration 25

C.11 Proof of Theorem 4 26

C.12 Proof of Theorem 5 27

C.13 Proof of Corollary 1 28

C.14 Proof of the Information–Regret Tradeoff 28

D Additional Experimental Details and Results **29**

D.1 Environment Construction and Evaluation Protocol 29

D.2 Baseline Implementations and Hyperparameters 29

D.3 Learning Curves for Return and Graph Recovery 29

D.4 Robustness to Graph Topology and Exposure Noise 30

D.5 Empirical Truncated Rollout Regret 30

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, including grammar correction, improving clarity and conciseness, reorganizing exposition, and suggesting LaTeX phrasing and formatting. These tools were used to improve the presentation of the manuscript and response materials. All AI-assisted content was reviewed, edited, and verified by the authors. The authors manually checked mathematical notation, assumptions, derivations, theorem statements, experimental descriptions, and citations against the final manuscript, and supporting materials. We take responsibility for the final content of this work, including all claims, methods, theoretical results, experiments, and artifacts produced with the aid of generative AI.

A RELATED WORK

Our work relates to three lines of research: learning and planning in networked RMABs, causal bandits with unknown graphs, and bandits with feedback, triggering, or mobility graphs. We review their closest connections below. Table 3 compares representative methods by their structural assumptions and objectives.

Table 3: **Comparison with related bandit and RMAB frameworks.** ✓, ✗, and △ denote yes, no, and partial or structurally different support, respectively. "Learns coupling" means that the inter-arm structure governing decisions is estimated online, rather than merely being unnecessary to the algorithm. "Dynamic spillover" means that an intervention on one arm changes another arm's state or future transition law. "Persistent nonlinear spillover" requires both temporal memory and a nonlinear propagation mechanism. "Dynamic control" indicates optimization of multi-period, rather than purely one-step, performance. Formal guarantees may concern regret, sample complexity, approximation, graph recovery, or asymptotic convergence, and are therefore not directly comparable across rows.

Methods	Learns Coupling	Markov Restless	Dynamic Spillover	Persistent Nonlinear Spillover	Saturation-Aware Design	Dynamic Control
Independent-arm and nonstationary bandits						
UCWhittle (Wang et al., 2023)	✗	✓	✗	✗	✗	✓
GINO-Q (Chen et al., 2026)	✗	✓	✗	✗	✗	✓
AoI-LUCB (Zhou et al., 2026)	✗	✓	✗	✗	✗	△
RC-BE (Migali et al., 2026)	✗	△	✗	✗	✗	✗
Networked and coupled RMABs						
ENGAGE (Ou et al., 2022)	✗	✓	△	✗	✗	✓
GRETA (Herlihy & Dickerson, 2023)	✗	✓	△	✗	✗	△
Networked RMAB–RL (Zhang et al., 2025)	✗	✓	✓	✗	✗	✓
WiFiP (Luan et al., 2025)	✗	✓	△	✗	✗	✓
SEQUOIA (Xu et al., 2025)	✗	✓	✗	✗	✗	✓
RMAB-G (Raman et al., 2024)	✗	✓	✗	✗	✗	✓
Unknown-graph causal learning and active search						
CS-UCB (Peng et al., 2024)	✓	✗	△	✗	✗	✗
GA-LCB (Yan & Tajer, 2024)	✓	✗	△	✗	✗	✗
POMIS-UCB (Elahi et al., 2024)	✓	✗	△	✗	✗	✗
BA-UCB (Zhao & Zhou, 2025)	△	✗	△	✗	✗	✗
CBO-U (Durand et al., 2025)	✓	✗	△	✗	✗	✗
Active Search (Tsugawa & Ohsaki, 2025)	✓	✗	✗	✗	✗	✗
Feedback, triggering, and traversal graphs						
q-FTRL (Eldowa et al., 2023)	✗	✗	✗	✗	✗	✗
Similar-arm feedback (Qi et al., 2026)	△	✗	✗	✗	✗	✗
Graph-triggered bandits (Genalti et al., 2026)	△	△	✗	✗	✗	✗
Dynamic-graph bandits (Rege et al.)	✗	✗	✗	✗	✗	△
Switching-graph bandits (Gong & Zhang, 2025)	✗	✗	✗	✗	✗	△
Multi-G-UCB (Paschalidis et al., 2026)	✗	✗	✗	✗	✗	△
N-GDOI (ours)	✓	✓	✓	✓	✓	✓

Learning and coupling in restless bandits. Classical RMABs exploit conditional independence across arms to obtain Lagrangian relaxations and index policies (Liu & Zhao, 2010; Mate et al.,

2020; Xiong et al., 2022; Wang et al., 2023). Recent online methods learn unknown arm-level dynamics while preserving this factorization. For example, UCWHITTLE constructs optimistic transition models and Whittle indices (Wang et al., 2023), whereas GINO-Q learns gain indices through three-timescale stochastic approximation without requiring indexability (Chen et al., 2026). Related work also studies pure exploration in restless systems (Zhou et al., 2026) and arms with exogenous rising reward profiles (Migali et al., 2026). These methods do not model actions on one arm as changing another arm's transition law through an unknown propagation process.

Networked and coupled RMAB models introduce richer interactions. Existing formulations include known commuting networks (Ou et al., 2022), positive externalities through graph-enabled message actions (Herlihy & Dickerson, 2023), known cascade dynamics (Zhang et al., 2025), and policy adaptation using given coupling features (Luan et al., 2025). Other approaches address coupling induced by combinatorial action constraints or global reward functions (Xu et al., 2025; Raman et al., 2024). These works primarily study planning or policy approximation under a specified interaction structure. In contrast, our transition coupling is generated by an unknown weighted directed graph and persistent nonlinear exposures. The learner must therefore identify the coupling mechanism from the same adaptive trajectory used for control.

Causal bandits and unknown-graph learning. Causal bandits exploit structural information to select interventions . Recent methods relax prior knowledge of the graph through local subgraph identification (Peng et al., 2024), graph-agnostic linear causal learning (Yan & Tajer, 2024), backdoor-adjusted confidence bounds (Zhao & Zhou, 2025), partial structure discovery under latent confounding (Elahi et al., 2024), or Bayesian optimization over uncertain causal structures (Durand et al., 2025). Active search methods likewise reveal parts of an unknown static network through explicit node queries (Tsugawa & Ohsaki, 2025). These settings differ from ours in both the data-generating process and the decision objective. They principally consider static structural models or direct graph queries, whereas N-GDOI learns from temporally dependent transitions of a controlled restless system. Persistence also creates a derivative-dependent identification problem: interventions can move the process into low-sensitivity regions in which additional samples contain little information about the propagation graph.

Graph-structured bandits. Graphs play several other roles in online learning. Feedback graphs encode which neighboring losses become observable after selecting an action (Eldowa et al., 2023; Qi et al., 2026); graph-triggered bandits use edges to determine which arms advance their evolution clocks (Genalti et al., 2026); and traversal or switching graphs restrict the actions that an agent can reach (Rege et al.; Gong & Zhang, 2025; Paschalidis et al., 2026). In these formulations, an edge represents information availability, triggering, or mobility. In our model, an edge is instead an unknown dynamical parameter: intervening on a source arm changes a target arm's future exposure and transition distribution. This distinction necessitates joint graph estimation and control. N-GDOI addresses it by combining derivative-weighted Fisher exploration with confidence-set optimistic stochastic rollouts, while explicitly accounting for graph, simulation, and action-search approximation errors.

B NOTATION AND ASSUMPTIONS

This section collects the notation and sufficient conditions used to prove the results in Section 4. We write $[N] = \{1, \dots, N\}$ and use $1\{\cdot\}$ for the indicator function. Vector norms are standard ℓ_p norms unless otherwise stated, and graph columns are indexed by target arms.

B.1 NOTATION

Throughout the paper and the technical proofs, we adopt the unified notation conventions shown in Table 4.

B.2 ASSUMPTIONS

We group the sufficient assumptions into model regularity, local sparse identifiability, reachable Fisher excitation, and controlled planning approximation. They are not claimed to be necessary. In

Table 4: **Notation used in the paper.**

Symbol	Meaning
System and environment	
N, T, M	Number of arms, decision horizon, and per-round intervention budget.
$\mathcal{S}_i$, s_i^t, s^t	Local state space, state of arm i, and global state at round t.
a_i^t, a^t, $\mathcal{A}_M$	Local action, joint action, and feasible set $\mathcal{A}_M = \{a \in \{0,1\}^N : \|a\|_1 \le M\}$.
x_i^t, x^t	Source signal and source-signal vector, where $x_i^t = g_i(s_i^t) + a_i^t \Delta_i(s_i^t)$.
z_i^t, z^t, $\mathcal{Z}_i$	Exposure of arm i, exposure vector, and exposure domain.
$P^\star$, $P^\star_{\cdot j}$, $\mathcal{P}$	True propagation graph, its incoming column for target j, and the admissible class of nonnegative sparse graphs.
d, $P_{\max}$	Maximum column sparsity and maximum columnwise ℓ_1 norm.
α, σ	Exposure persistence and known nonlinear propagation link.
η_j^{t+1}, ν_z	Exposure innovation and its conditional sub-Gaussian scale.
$\mathbb{P}_i^\star$, r_i	Known local transition kernel and expected local reward.
R^t, ξ^t, $R_{\max}$	Observed reward, reward noise, and upper bound on expected one-step system reward.
γ, H, μ	Discount factor, rollout horizon, and feasible continuation policy.
$\mathcal{F}_t$, $\mathcal{F}_t^+$	Pre-decision history and post-action history $\mathcal{F}_t^+ = \sigma(\mathcal{F}_t, a^t)$.
Estimation and exploration	
$\mathcal{T}_{\exp}$, $\mathcal{E}_t$	Scheduled exploration rounds and their observed subset $\mathcal{E}_t = \mathcal{T}_{\exp} \cap \{0, \dots, t-1\}$.
$n_{\exp}(t)$	Number of observed exploration rounds, $n_{\exp}(t) = \|\mathcal{E}_t\|$.
$\widetilde{P}_{\cdot j}$, r_{loc}	Burn-in localizer for column j and localization radius.
$\mathcal{W}_j^{\mathrm{loc}}$	Localized estimator domain $\{w \in \mathbb{R}_+^N : \|w - \widetilde{P}_{\cdot j}\|_1 \le r_{\mathrm{loc}}/2\}$.
$\widehat{P}_{\cdot j,t}$, λ_t	Estimated incoming graph column and sparse-regression penalty.
λ_{rid}, $\widehat{V}_{j,t}$	Fisher ridge parameter and plug-in Fisher design matrix.
θ_j^t, ω_j^t	True predictor $\theta_j^t = \alpha z_j^t + \langle P^\star_{\cdot j}, x^t \rangle$ and oracle Fisher weight $\omega_j^t = [\sigma'(\theta_j^t)]^2$.
$V^\star_{j,t}$	Oracle Fisher design $\lambda_{\mathrm{rid}} I + \sum_{\tau \in \mathcal{E}_t} \omega_j^\tau x^\tau (x^\tau)^\top$.
S_j, $\mathcal{K}_j$	True incoming support $S_j = \mathrm{supp}(P^\star_{\cdot j})$ and sparse cone $\mathcal{K}_j = \{v : \|v_{S_j^c}\|_1 \le 3\|v_{S_j}\|_1\}$.
$\Delta_{j,t}$	Column estimation error $\widehat{P}_{\cdot j,t} - P^\star_{\cdot j}$.
$\kappa_{\exp}$, κ_{FI}, κ_{PE}	Exploration-design curvature, local Fisher curvature, and persistent-excitation curvature.
Confidence sets, planning, and regret	
$\mathcal{C}_t$, $\varepsilon_t^{\mathrm{conf}}$	Graph confidence set and its columnwise ℓ_1 radius.
$\|A\|_{1,\infty}$	Columnwise matrix norm $\|A\|_{1,\infty} = \max_{j \in [N]} \|A_{\cdot j}\|_1$.
D_t	Confidence-set diameter $D_t = \sup_{P, \widetilde{P} \in \mathcal{C}_t} \|P - \widetilde{P}\|_{1,\infty}$.
$\mathcal{G}_t$, ρ_t	Finite candidate graph set and its $\|\cdot\|_{1,\infty}$ covering radius.
$Q_{t,H}(a; P)$	Exact expected H-step rollout value under graph P.
$\widehat{Q}_{t,H}^{K_t}(a; P)$, K_t	Monte-Carlo rollout estimate and number of simulated trajectories.
$a_t^\star$	True-graph rollout-oracle action $a_t^\star \in \arg\max_{a \in \mathcal{A}_M} Q_{t,H}(a; P^\star)$.
$\varepsilon_{\mathrm{opt},t}$	Certified residual of approximate budgeted action optimization.
R_H, $b_{t,K}$	Maximum H-step discounted reward $R_H = R_{\max}(1-\gamma^H)/(1-\gamma)$ and Monte-Carlo confidence radius.
$\mathfrak{R}_T^{(H)}$	Truncated rollout regret relative to the true-graph rollout oracle.

particular, the Fisher condition explicitly excludes environments in which every feasible intervention remains permanently saturated.

Assumption 1 (Bounded model and rollout regularity)**.** ***Bounded model.*** *For every round t, source signals satisfy $0 \le x_i^t \le x_{\max}$ and $\|x^t\|_1 \le B_x$. Exposures lie in $\mathcal{Z}_i \subseteq [0, z_{\max}]$. Every $P \in \mathcal{P}$ is nonnegative, column-d-sparse, and satisfies $\max_j \|P_{\cdot j}\|_1 \le P_{\max}$. The expected local rewards obey $0 \le r_i(s_i, a_i, z_i) \le R_{\max,i}$, with $\sum_i R_{\max,i} \le R_{\max}$. The rollout value $Q_{t,H}$ uses expected rewards; observation noise ξ^t is not simulated by the planner.*

Link and exposure innovations. *The known link σ is twice continuously differentiable on the reachable predictor domain and satisfies*

$$|\sigma'(u)| \le L_\sigma, \qquad |\sigma'(u) - \sigma'(v)| \le L_{\sigma'}|u - v|.$$

Conditional on the post-action history $\mathcal{F}_t^+$, η_j^{t+1} is mean-zero and ν_z-sub-Gaussian. The conditional innovation law is graph-independent, can be sampled by the rollout simulator, and admits a common-innovation coupling for rollouts under different candidate graphs. Its support is compatible with $z_j^{t+1} \in \mathcal{Z}_j$ under the additive exposure dynamics Eq. 1.

Local exposure regularity. *For fixed local state s_i and action a_i, the transition kernels and expected rewards satisfy*

$$\|\mathbb{P}_i^\star(\cdot \mid s_i, a_i, z) - \mathbb{P}_i^\star(\cdot \mid s_i, a_i, z')\|_{\mathrm{TV}} \le L_{\mathrm{tr},i}|z - z'|,$$

and

$$|r_i(s_i, a_i, z) - r_i(s_i, a_i, z')| \le L_{r,i}|z - z'|.$$

Coupled rollout stability. *Consider two rollouts that start from the same (s^t, z^t), take the same first action, and subsequently follow the same feasible continuation policy μ under graphs $P, \widetilde{P} \in \mathcal{P}$. Couple their graph-independent exposure innovations identically and their local-state transitions using common randomness or maximal coupling. Define*

$$e_z^h := \mathbb{E}\|z_P^{t+h} - z_{\widetilde{P}}^{t+h}\|_\infty, \qquad e_s^h := \sum_{i=1}^N \Pr(s_{i,P}^{t+h} \ne s_{i,\widetilde{P}}^{t+h}),$$

and $\Delta_P := \|P - \widetilde{P}\|_{1,\infty}$. There exist finite, nonnegative constants $A_{zz}, A_{zs}, A_{sz}, A_{ss}, b_P$, uniform over feasible first actions, candidate graphs, and rollout rounds, such that

$$e_z^{h+1} \le A_{zz}e_z^h + A_{zs}e_s^h + b_P\Delta_P, \tag{13}$$

$$e_s^{h+1} \le A_{sz}e_z^{h+1} + A_{ss}e_s^h. \tag{14}$$

The constants include any differences in continuation actions caused by differences in the coupled states or exposures.

Finally, there exist finite, nonnegative reward-sensitivity constants c_z, c_s such that the coupled expected one-step rewards satisfy

$$\left|\mathbb{E}\left[r(s_P^{t+h}, z_P^{t+h}, a_P^{t+h}) - r(s_{\widetilde{P}}^{t+h}, z_{\widetilde{P}}^{t+h}, a_{\widetilde{P}}^{t+h})\right]\right| \le c_z e_z^h + c_s e_s^h. \tag{15}$$

Assumption 2 (Local sparse identifiability)**.** *After burn-in, each localizer satisfies*

$$\|\widetilde{P}_{\cdot j} - P^\star_{\cdot j}\|_1 \le r_{\mathrm{loc}}/2, \qquad j \in [N],$$

so $P^\star_{\cdot j} \in \mathcal{W}_j^{\mathrm{loc}}$. For every $v \in \mathcal{K}_j$ with $\|v\|_1 \le r_{\mathrm{loc}}$, the nonlinear least-squares loss obeys

$$\mathcal{L}_{j,t}(P^\star_{\cdot j} + v) - \mathcal{L}_{j,t}(P^\star_{\cdot j}) - \left\langle \nabla\mathcal{L}_{j,t}(P^\star_{\cdot j}), v\right\rangle \ge \frac{\kappa_{\mathrm{FI}}}{4}\|v\|_2^2. \tag{16}$$

Exact directed-support recovery additionally requires invertibility of the restricted local Hessian, a local nonlinear irrepresentability condition, and a beta-min condition exceeding the coordinatewise estimation error.

Assumption 3 (Reachable sparse Fisher excitation)**.** *On an explicit exploration round, the exploration policy may randomize after observing $\mathcal{F}_t$. Uniformly over target arms $j \in [N]$ and directions $v \in \mathcal{K}_j$,*

$$\mathbb{E}\left[\omega_j^t\langle v, x^t\rangle^2 \mid \mathcal{F}_t\right] \ge \kappa_{\mathrm{exp}}\|v\|_2^2. \tag{17}$$

The weighted design process is conditionally bounded and satisfies the sparse-cone martingale covering condition used in Appendix C. In addition, informative reachable actions satisfy $|\sigma'(\theta_j^t)| \ge m_\sigma > 0$.

Assumption 4 (Controlled planning approximation)**.** *At each exploitation round, the candidate graph set satisfies*

$$\sup_{P\in\mathcal{C}_t}\inf_{\widetilde{P}\in\mathcal{G}_t}\|P-\widetilde{P}\|_{1,\infty}\le\rho_t. \tag{18}$$

The action-search routine returns $a^t\in\mathcal{A}_M$ such that

$$\max_{a\in\mathcal{A}_M}\max_{P\in\mathcal{G}_t}\widehat{Q}^{K_t}_{t,H}(a;P)-\max_{P\in\mathcal{G}_t}\widehat{Q}^{K_t}_{t,H}(a^t;P)\le\varepsilon_{\mathrm{opt},t}. \tag{19}$$

The residual is measured relative to the full feasible action class; in the practical beam-search implementation it represents the certified or assumed gap between the returned action and the exhaustive Monte-Carlo objective.

C Proofs of the Main Theoretical Results

This section proves the theoretical results in Section 4. The proof follows the same sequence as the main analysis. We first control the adaptive nonlinear score and establish sparse graph recovery. We then show that Fisher-aware exploration supplies the restricted curvature required by the estimator. The remaining subsections translate graph error into stochastic rollout error and combine statistical and computational approximation terms into the final regret bounds.

Throughout the proofs, let $m_t:=n_{\mathrm{exp}}(t)\vee 1$. For target arm j, define

$$\theta_j^\tau:=\alpha z_j^\tau+\langle P^\star_{\cdot j},x^\tau\rangle,\qquad z_j^{\tau+1}=\sigma(\theta_j^\tau)+\eta_j^{\tau+1}. \tag{20}$$

The nonlinear least-squares loss based on explicit exploration transitions is

$$\mathcal{L}_{j,t}(w):=\frac{1}{m_t}\sum_{\tau\in\mathcal{E}_t}\left[z_j^{\tau+1}-\sigma\big(\alpha z_j^\tau+\langle w,x^\tau\rangle\big)\right]^2. \tag{21}$$

The corresponding localized sparse estimator is

$$\widehat{P}_{\cdot j,t}\in\arg\min_{w\in\mathcal{W}_j^{\mathrm{loc}}}\left\{\mathcal{L}_{j,t}(w)+\lambda_t\|w\|_1\right\}. \tag{22}$$

C.1 Proof of Adaptive Score Concentration

The action at round τ may depend on all previous observations. However, the source signal x^τ and the predictor θ_j^τ are determined before the next innovation $\eta_j^{\tau+1}$ is observed. This predictability allows us to treat the score as a martingale sum.

Lemma 1 (Adaptive score concentration)**.** Under Assumption 1, there exists a universal constant $c_0>0$ such that, with probability at least $1-\delta$, simultaneously for all $j\in[N]$ and all post-burn-in rounds $t\in[T]$,

$$\left\|\nabla\mathcal{L}_{j,t}(P^\star_{\cdot j})\right\|_\infty\le c_0\nu_z L_\sigma B_x\sqrt{\frac{\log(2N^2T/\delta)}{n_{\mathrm{exp}}(t)\vee 1}}. \tag{23}$$

Proof. Fix a source coordinate $i\in[N]$ and target arm $j\in[N]$. Define the calendar-time increment

$$Z_{ij}^\tau:=1\{\tau\in\mathcal{T}_{\mathrm{exp}}\}\eta_j^{\tau+1}\sigma'(\theta_j^\tau)x_i^\tau. \tag{24}$$

Because the exploration schedule is predictable, the indicator $1\{\tau\in\mathcal{T}_{\mathrm{exp}}\}$ is known before $\eta_j^{\tau+1}$ is observed. Moreover, x^τ and θ_j^τ are $\mathcal{F}_\tau^+$-measurable. Hence

$$\mathbb{E}[Z_{ij}^\tau\mid\mathcal{F}_\tau^+]=1\{\tau\in\mathcal{T}_{\mathrm{exp}}\}\sigma'(\theta_j^\tau)x_i^\tau\mathbb{E}[\eta_j^{\tau+1}\mid\mathcal{F}_\tau^+]=0. \tag{25}$$

Thus, $\{Z_{ij}^\tau\}_{\tau\ge0}$ is a martingale-difference sequence with respect to the post-action filtration.

Conditional sub-Gaussianity of the innovation gives, for every $\lambda\in\mathbb{R}$,

$$\begin{aligned}\mathbb{E}\big[\exp(\lambda Z_{ij}^\tau)\mid\mathcal{F}_\tau^+\big]&\le\exp\left(\frac{\lambda^2\nu_z^2}{2}1\{\tau\in\mathcal{T}_{\mathrm{exp}}\}[\sigma'(\theta_j^\tau)]^2(x_i^\tau)^2\right)\\&\le\exp\left(\frac{\lambda^2\nu_z^2L_\sigma^2B_x^2}{2}1\{\tau\in\mathcal{T}_{\mathrm{exp}}\}\right),\end{aligned} \tag{26}$$

where we used $|\sigma'(\theta_j^\tau)| \le L_\sigma$ and $|x_i^\tau| \le \|x^\tau\|_1 \le B_x$.

For any fixed t, iterating the conditional moment-generating-function bound over $\tau < t$ gives

$$\mathbb{E}\left[\exp\left(\lambda \sum_{\tau\in\mathcal{E}_t} \eta_j^{\tau+1}\sigma'(\theta_j^\tau)x_i^\tau\right)\right] \le \exp\left(\frac{\lambda^2 n_{\exp}(t)\nu_z^2 L_\sigma^2 B_x^2}{2}\right). \tag{27}$$

Consequently, for every $u > 0$,

$$\Pr\left(\left|\sum_{\tau\in\mathcal{E}_t} \eta_j^{\tau+1}\sigma'(\theta_j^\tau)x_i^\tau\right| \ge u\right) \le 2\exp\left(-\frac{u^2}{2n_{\exp}(t)\nu_z^2 L_\sigma^2 B_x^2}\right) \tag{28}$$

whenever $n_{\exp}(t) > 0$. If the exploration count is predictable but history dependent, the same bound follows by conditioning on each possible value of $n_{\exp}(t)$ and union bounding over at most T possible counts; this additional logarithmic factor is already absorbed by $\log(2N^2T/\delta)$.

We next compute the score. Differentiating Eq. 21 yields

$$\begin{aligned}\nabla\mathcal{L}_{j,t}(w) = -\frac{2}{m_t}\sum_{\tau\in\mathcal{E}_t} &\left[z_j^{\tau+1} - \sigma\big(\alpha z_j^\tau + \langle w, x^\tau\rangle\big)\right] \\ &\times \sigma'\big(\alpha z_j^\tau + \langle w, x^\tau\rangle\big)\, x^\tau.\end{aligned} \tag{29}$$

Evaluating this expression at $w = P_{\cdot j}^\star$ and using Eq. 20 gives

$$[\nabla\mathcal{L}_{j,t}(P_{\cdot j}^\star)]_i = -\frac{2}{m_t}\sum_{\tau\in\mathcal{E}_t} \eta_j^{\tau+1}\sigma'(\theta_j^\tau)x_i^\tau. \tag{30}$$

Set $u = \nu_z L_\sigma B_x\sqrt{2n_{\exp}(t)\log(2N^2T/\delta)}$. Substitution into Eq. 28, followed by division by m_t, shows that the ith score coordinate is bounded at the claimed order. A union bound over $i, j \in [N]$ and $t \in [T]$ proves Eq. 23. When $n_{\exp}(t) = 0$, the sum in Eq. 21 is empty and the gradient is zero by convention. □

C.2 Proof of the Basic Inequality and Sparse Cone

For a feasible displacement v, define the local Bregman divergence

$$\mathcal{D}_{j,t}(v) := \mathcal{L}_{j,t}(P_{\cdot j}^\star + v) - \mathcal{L}_{j,t}(P_{\cdot j}^\star) - \left\langle \nabla\mathcal{L}_{j,t}(P_{\cdot j}^\star), v\right\rangle. \tag{31}$$

The local-identifiability assumption requires $\mathcal{D}_{j,t}(v) \ge 0$ over the feasible localization region and provides the stronger quadratic lower bound once v belongs to the sparse cone.

Lemma 2 (Basic inequality and cone condition)**.** Let $\Delta_{j,t} := \widehat{P}_{\cdot j,t} - P_{\cdot j}^\star$. If $\lambda_t \ge 2\left\|\nabla\mathcal{L}_{j,t}(P_{\cdot j}^\star)\right\|_\infty$, then

$$\frac{\kappa_{\mathrm{FI}}}{4}\|\Delta_{j,t}\|_2^2 + \frac{\lambda_t}{2}\|(\Delta_{j,t})_{S_j^c}\|_1 \le \frac{3\lambda_t}{2}\|(\Delta_{j,t})_{S_j}\|_1. \tag{32}$$

In particular,

$$\|(\Delta_{j,t})_{S_j^c}\|_1 \le 3\|(\Delta_{j,t})_{S_j}\|_1, \qquad \Delta_{j,t} \in \mathcal{K}_j. \tag{33}$$

Proof. Fix j, t and abbreviate

$$w^\star := P_{\cdot j}^\star, \qquad \widehat{w} := \widehat{P}_{\cdot j,t}, \qquad \Delta := \widehat{w} - w^\star.$$

The localization condition ensures that $w^\star$ is feasible for Eq. 22. Therefore, optimality of $\widehat{w}$ gives

$$\mathcal{L}_{j,t}(\widehat{w}) + \lambda_t\|\widehat{w}\|_1 \le \mathcal{L}_{j,t}(w^\star) + \lambda_t\|w^\star\|_1. \tag{34}$$

Adding and subtracting the first-order term at $w^\star$ yields

$$\mathcal{D}_{j,t}(\Delta) \le -\left\langle \nabla\mathcal{L}_{j,t}(w^\star), \Delta\right\rangle + \lambda_t\left(\|w^\star\|_1 - \|\widehat{w}\|_1\right). \tag{35}$$

By Hölder's inequality,

$$-\langle \nabla \mathcal{L}_{j,t}(w^\star), \Delta \rangle \le \|\nabla \mathcal{L}_{j,t}(w^\star)\|_\infty \|\Delta\|_1 \le \frac{\lambda_t}{2}\left(\|\Delta_{S_j}\|_1 + \|\Delta_{S_j^c}\|_1\right). \tag{36}$$

Because $w^\star_{S_j^c} = 0$,

$$\|w^\star\|_1 - \|\widehat{w}\|_1 = \|w^\star_{S_j}\|_1 - \|w^\star_{S_j} + \Delta_{S_j}\|_1 - \|\Delta_{S_j^c}\|_1 \le \|\Delta_{S_j}\|_1 - \|\Delta_{S_j^c}\|_1, \tag{37}$$

where the final inequality follows from the reverse triangle inequality.

Substituting Eq. 36 and Eq. 37 into Eq. 35 gives

$$\mathcal{D}_{j,t}(\Delta) \le \frac{3\lambda_t}{2}\|\Delta_{S_j}\|_1 - \frac{\lambda_t}{2}\|\Delta_{S_j^c}\|_1. \tag{38}$$

The local Bregman divergence is nonnegative over the feasible region. Therefore,

$$0 \le \frac{3\lambda_t}{2}\|\Delta_{S_j}\|_1 - \frac{\lambda_t}{2}\|\Delta_{S_j^c}\|_1,$$

which proves Eq. 33.

It remains to invoke local curvature. Both $\widehat{w}$ and $w^\star$ belong to $\mathcal{W}_j^{\mathrm{loc}}$, and hence

$$\|\Delta\|_1 \le \|\widehat{w} - \widetilde{P}_{\cdot j}\|_1 + \|w^\star - \widetilde{P}_{\cdot j}\|_1 \le r_{\mathrm{loc}}.$$

Together with $\Delta \in \mathcal{K}_j$, this permits applying Eq. 16:

$$\mathcal{D}_{j,t}(\Delta) \ge \frac{\kappa_{\mathrm{FI}}}{4}\|\Delta\|_2^2.$$

Combining this lower bound with Eq. 38 proves Eq. 32. □

C.3 Proof of Theorem 1

Proof. Work on the uniform event of Lemma 1 and choose

$$\lambda_t = 2c_0\nu_z L_\sigma B_x \sqrt{\frac{\log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}}. \tag{39}$$

Fix j, t. By Lemma 2, $\Delta_{j,t} \in \mathcal{K}_j$. Since $|S_j| \le d$, Cauchy–Schwarz gives

$$\|(\Delta_{j,t})_{S_j}\|_1 \le \sqrt{|S_j|}\|(\Delta_{j,t})_{S_j}\|_2 \le \sqrt{d}\,\|\Delta_{j,t}\|_2. \tag{40}$$

Dropping the nonnegative off-support term from Eq. 32 and applying Eq. 40 gives

$$\frac{\kappa_{\mathrm{FI}}}{4}\|\Delta_{j,t}\|_2^2 \le \frac{3\lambda_t\sqrt{d}}{2}\|\Delta_{j,t}\|_2.$$

If $\Delta_{j,t} = 0$, the desired bounds are immediate. Otherwise, dividing by $\|\Delta_{j,t}\|_2$ yields

$$\|\Delta_{j,t}\|_2 \le \frac{6\lambda_t\sqrt{d}}{\kappa_{\mathrm{FI}}}. \tag{41}$$

Substituting Eq. 39 proves

$$\|\Delta_{j,t}\|_2 \le \frac{C_1\nu_z L_\sigma B_x}{\kappa_{\mathrm{FI}}}\sqrt{\frac{d\log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}}$$

for a universal constant C_1.

For the ℓ_1 error, the cone condition gives

$$\begin{aligned}\|\Delta_{j,t}\|_1 &= \|(\Delta_{j,t})_{S_j}\|_1 + \|(\Delta_{j,t})_{S_j^c}\|_1 \\ &\le 4\|(\Delta_{j,t})_{S_j}\|_1 \le 4\sqrt{d}\,\|\Delta_{j,t}\|_2. \end{aligned} \tag{42}$$

Combining Eq. 42 with Eq. 41 proves

$$\|\Delta_{j,t}\|_1 \le \frac{C_2 d \nu_z L_\sigma B_x}{\kappa_{\mathrm{FI}}} \sqrt{\frac{\log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}}.$$

The probability statement is simultaneous over all j and t because the score event in Lemma 1 is uniform. □

C.4 Proof of Confidence-Set Validity

Proposition 1 (Uniform confidence-set validity)**.** *Suppose the conditions of Theorem 1 hold and choose*

$$\varepsilon_t^{\mathrm{conf}} = c_{\mathrm{conf}} \frac{d \nu_z L_\sigma B_x}{\kappa_{\mathrm{FI}}} \sqrt{\frac{\log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}} \tag{43}$$

with $c_{\mathrm{conf}} \ge C_2$. Then, with probability at least $1-\delta$,

$$P^\star \in \mathcal{C}_t, \qquad \mathrm{diam}_{1,\infty}(\mathcal{C}_t) \le 2\varepsilon_t^{\mathrm{conf}}, \qquad \forall t \in [T]. \tag{44}$$

Proof. On the event of Theorem 1,

$$\|P^\star_{\cdot j} - \widehat{P}_{\cdot j,t}\|_1 \le \varepsilon_t^{\mathrm{conf}}$$

for every j and t. Since $P^\star \in \mathcal{P}$, this is precisely the membership condition defining $\mathcal{C}_t$ in Eq. 5; hence $P^\star \in \mathcal{C}_t$.

Now take arbitrary $P, \widetilde{P} \in \mathcal{C}_t$. For every target column j,

$$\|P_{\cdot j} - \widetilde{P}_{\cdot j}\|_1 \le \|P_{\cdot j} - \widehat{P}_{\cdot j,t}\|_1 + \|\widetilde{P}_{\cdot j} - \widehat{P}_{\cdot j,t}\|_1 \le 2\varepsilon_t^{\mathrm{conf}}. \tag{45}$$

Taking the maximum over columns and then the supremum over $P, \widetilde{P} \in \mathcal{C}_t$ proves the diameter bound. □

C.5 Proof of Matrix-Level Graph Recovery

Corollary 2 (Matrix-level graph recovery)**.** *On the event of Theorem 1,*

$$\|\widehat{P}_t - P^\star\|_{1,\infty} \le \varepsilon_t^{\mathrm{conf}}, \tag{46}$$

$$\|\widehat{P}_t - P^\star\|_F \le \frac{C_1 \nu_z L_\sigma B_x}{\kappa_{\mathrm{FI}}} \sqrt{\frac{N d \log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}}. \tag{47}$$

Proof. The first result follows by maximizing the uniform columnwise ℓ_1 bound:

$$\|\widehat{P}_t - P^\star\|_{1,\infty} = \max_{j \in [N]} \|\widehat{P}_{\cdot j,t} - P^\star_{\cdot j}\|_1.$$

For the Frobenius norm,

$$\begin{aligned}\|\widehat{P}_t - P^\star\|_F^2 &= \sum_{j=1}^{N} \|\widehat{P}_{\cdot j,t} - P^\star_{\cdot j}\|_2^2 \\ &\le N \left[\frac{C_1 \nu_z L_\sigma B_x}{\kappa_{\mathrm{FI}}} \sqrt{\frac{d \log(2N^2T/\delta)}{n_{\exp}(t) \vee 1}} \right]^2. \end{aligned} \tag{48}$$

Taking square roots proves Eq. 47. □

C.6 Proof of Directed Support Recovery

Proposition 2 (Directed support recovery)**.** *Suppose the restricted local Hessian is invertible and the local nonlinear irrepresentability condition holds with a strictly positive margin. Then there exists a constant $C_{\mathrm{supp}} > 0$ such that*

$$\|\widehat{P}_{S_j,j,t} - P^\star_{S_j,j}\|_\infty \le C_{\mathrm{supp}}\lambda_t \tag{49}$$

and $\widehat{P}_{ij,t} = 0$ for $i \notin S_j$. If, in addition,

$$\min_{i\in S_j} P^\star_{ij} > 2C_{\mathrm{supp}}\lambda_t, \tag{50}$$

then thresholding $\widehat{P}_{\cdot j,t}$ at $C_{\mathrm{supp}}\lambda_t$ recovers S_j.

Proof. Fix a target arm j. Consider the restricted optimization problem in which coordinates in S_j^c are fixed to zero:

$$\widehat{w}_{S_j} \in \arg\min_{w_{S_j}} \left\{ \mathcal{L}_{j,t}(w_{S_j}, 0) + \lambda_t \|w_{S_j}\|_1 \right\}, \tag{51}$$

subject to the inherited localization and nonnegativity constraints.

Let $\widehat{u}_{S_j}$ be an ℓ_1 subgradient on the support. The restricted KKT equation can be written as

$$\nabla_{S_j}\mathcal{L}_{j,t}(\widehat{w}) + \lambda_t \widehat{u}_{S_j} + n_{S_j}(\widehat{w}) = 0, \tag{52}$$

where $n_{S_j}(\widehat{w})$ denotes the normal-cone contribution from the localization and nonnegativity constraints. Under the beta-min condition and for sufficiently large $n_{\exp}(t)$, the true support coordinates remain in the interior, so this normal-cone term vanishes on S_j.

Expanding the gradient around $w^\star = P^\star_{\cdot j}$ gives

$$\nabla_{S_j}\mathcal{L}_{j,t}(\widehat{w}) = \nabla_{S_j}\mathcal{L}_{j,t}(w^\star) + H_{j,t,S_jS_j}(\widehat{w}_{S_j} - w^\star_{S_j}) + \mathcal{R}_{j,t}, \tag{53}$$

where H_{j,t,S_jS_j} is the restricted local Hessian and $\mathcal{R}_{j,t}$ is the nonlinear Taylor remainder. Substituting Eq. 53 into Eq. 52 gives

$$\widehat{w}_{S_j} - w^\star_{S_j} = -H^{-1}_{j,t,S_jS_j}\left[\nabla_{S_j}\mathcal{L}_{j,t}(w^\star) + \lambda_t\widehat{u}_{S_j} + \mathcal{R}_{j,t}\right]. \tag{54}$$

The score event, bounded inverse restricted Hessian, and local remainder condition imply

$$\|\widehat{w}_{S_j} - w^\star_{S_j}\|_\infty \le C_{\mathrm{supp}}\lambda_t.$$

The local irrepresentability condition controls the off-support KKT quantities. Specifically, it guarantees that the dual witness constructed from the restricted solution satisfies

$$\left\|\nabla_{S_j^c}\mathcal{L}_{j,t}(\widehat{w})\right\|_\infty < \lambda_t.$$

Thus the restricted solution, extended by zeros on S_j^c, satisfies the KKT conditions of the full problem and has no false-positive coordinates.

Finally, for every $i \in S_j$, Eq. 49 and Eq. 50 imply

$$\widehat{P}_{ij,t} \ge P^\star_{ij} - C_{\mathrm{supp}}\lambda_t > C_{\mathrm{supp}}\lambda_t.$$

Every off-support coordinate is zero. Thresholding at $C_{\mathrm{supp}}\lambda_t$ therefore exactly recovers the directed incoming support. □

C.7 Proof of Theorem 2

Recall the oracle derivative-weighted design

$$V^\star_{j,t} = \lambda_{\mathrm{rid}} I + \sum_{\tau\in\mathcal{E}_t} \omega^\tau_j x^\tau (x^\tau)^\top, \qquad \omega^\tau_j = [\sigma'(\theta^\tau_j)]^2.$$

Proof. Fix $j \in [N]$ and a direction $v \in \mathcal{K}_j$ with $\|v\|_2 = 1$. Define

$$Y_\tau(v) := 1\{\tau \in \mathcal{T}_{\exp}\}\omega_j^\tau \langle v, x^\tau\rangle^2.$$

Unlike the score proof, the relevant conditional mean is taken with respect to the pre-decision filtration $\mathcal{F}_\tau$, because the exploration policy may randomize its action after observing $\mathcal{F}_\tau$. Assumption 3 gives

$$\mathbb{E}[Y_\tau(v) \mid \mathcal{F}_\tau] \geq 1\{\tau \in \mathcal{T}_{\exp}\}\kappa_{\exp}\|v\|_2^2 = 1\{\tau \in \mathcal{T}_{\exp}\}\kappa_{\exp}. \tag{55}$$

For $v \in \mathcal{K}_j$,

$$\begin{aligned}\|v\|_1 &= \|v_{S_j}\|_1 + \|v_{S_j^c}\|_1 \leq 4\|v_{S_j}\|_1 \\ &\leq 4\sqrt{|S_j|}\|v\|_2 \leq 4\sqrt{d}.\end{aligned} \tag{56}$$

Consequently,

$$0 \leq Y_\tau(v) \leq L_\sigma^2\langle v, x^\tau\rangle^2 \leq L_\sigma^2\|v\|_1^2\|x^\tau\|_\infty^2 \leq 16L_\sigma^2 B_x^2 d. \tag{57}$$

Let

$$\overline{Y}_\tau(v) := Y_\tau(v) - \mathbb{E}[Y_\tau(v) \mid \mathcal{F}_\tau].$$

The sequence $\{\overline{Y}_\tau(v)\}$ is a bounded martingale-difference sequence. By the sparse-cone martingale covering condition in Assumption 3, the associated Freedman bound extends uniformly over

$$\mathbb{S}_j := \mathcal{K}_j \cap \{v : \|v\|_2 = 1\}.$$

More precisely, for $n_{\exp}(t) \geq Cd\log(2NT/\delta)$, with probability at least $1-\delta$,

$$\sup_{v\in\mathbb{S}_j}\left|\frac{1}{n_{\exp}(t)}\sum_{\tau\in\mathcal{E}_t}\overline{Y}_\tau(v)\right| \leq \frac{\kappa_{\exp}}{2}, \tag{58}$$

simultaneously over targets and post-burn-in rounds. The effective entropy is of order $d\log N$ because the cone concentrates its ℓ_1 mass on at most d support coordinates; the precise covering statement is included in Assumption 3.

Combining Eq. 55 and Eq. 58 gives

$$\inf_{v\in\mathbb{S}_j}\frac{1}{n_{\exp}(t)}\sum_{\tau\in\mathcal{E}_t}\omega_j^\tau\langle v, x^\tau\rangle^2 \geq \kappa_{\exp} - \frac{\kappa_{\exp}}{2} = \frac{\kappa_{\exp}}{2}. \tag{59}$$

Because the ridge term is positive semidefinite,

$$v^\top V_{j,t}^\star v \geq \sum_{\tau\in\mathcal{E}_t}\omega_j^\tau\langle v, x^\tau\rangle^2.$$

Therefore,

$$\inf_{v\in\mathbb{S}_j}\frac{v^\top V_{j,t}^\star v}{n_{\exp}(t)} \geq \frac{\kappa_{\exp}}{2}.$$

Rescaling an arbitrary nonzero $v \in \mathcal{K}_j$ proves Eq. 9.

Finally, the nonlinear local-curvature assumption controls the difference between the empirical Fisher quadratic form and the Bregman divergence of the nonlinear least-squares loss. In the localized region, the second-order remainder is dominated by a fixed fraction of the Fisher term. Consequently, the local RSC constant can be chosen so that

$$\kappa_{\mathrm{FI}} \asymp \kappa_{\exp},$$

which establishes the final claim of Theorem 2. □

C.8 Proof of Theorem 3

We compare two stochastic rollouts initialized from the same observed state–exposure pair (s^t, z^t) and taking the same first action $a \in \mathcal{A}_M$. One rollout evolves under graph P, and the other under graph $\widetilde{P}$. Thereafter, both rollouts use the same feasible continuation policy μ.

The graph-independent exposure innovations are coupled identically. Conditional on the resulting exposures and actions, the local-state transitions are coupled using common randomness or maximal coupling. Let

$$e_z^h := \mathbb{E}\Big[\|z_P^{t+h} - z_{\widetilde{P}}^{t+h}\|_\infty\Big], \qquad e_s^h := \sum_{i=1}^{N} \Pr\Big(s_{i,P}^{t+h} \neq s_{i,\widetilde{P}}^{t+h}\Big). \tag{60}$$

Because the two rollouts start from the same observed condition,

$$e_z^0 = e_s^0 = 0. \tag{61}$$

Write $\Delta_P := \|P - \widetilde{P}\|_{1,\infty}$. Under Assumption 1, the coupled dynamics admit nonnegative constants $A_{zz}, A_{zs}, A_{sz}, A_{ss}$ and b_P such that

$$e_z^{h+1} \le A_{zz} e_z^h + A_{zs} e_s^h + b_P \Delta_P, \tag{62}$$

$$e_s^{h+1} \le A_{sz} e_z^{h+1} + A_{ss} e_s^h. \tag{63}$$

The constants account for persistence, nonlinear-link sensitivity, changes in source signals induced by coupled state trajectories, and changes in continuation actions under μ.

For the basic state-dependent source model in Section 2, one may take

$$A_{zz} = L_\sigma \alpha, \qquad A_{zs} = L_\sigma P_{\max} L_x, \qquad b_P = L_\sigma B_x,$$

with A_{sz} and A_{ss} determined by the exposure sensitivity of the local transition kernels and the coupling stability of μ.

Lemma 3 (Joint rollout coupling recursion)**.** Let

$$e^h := \begin{pmatrix} e_z^h \\ e_s^h \end{pmatrix}, \qquad B := \begin{pmatrix} A_{zz} & A_{zs} \\ A_{sz}A_{zz} & A_{ss} + A_{sz}A_{zs} \end{pmatrix}, \qquad c := \begin{pmatrix} b_P \\ A_{sz} b_P \end{pmatrix}.$$

Then

$$e^{h+1} \preceq B e^h + c\Delta_P, \tag{64}$$

where $\preceq$ denotes coordinatewise inequality. Consequently,

$$e^h \preceq \left(\sum_{\ell=0}^{h-1} B^\ell\right) c\Delta_P. \tag{65}$$

Proof. Equation Eq. 62 gives the first coordinate of Eq. 64. Substituting it into Eq. 63 yields

$$\begin{aligned} e_s^{h+1} &\le A_{sz}\left(A_{zz}e_z^h + A_{zs}e_s^h + b_P\Delta_P\right) + A_{ss}e_s^h \\ &= A_{sz}A_{zz}e_z^h + \left(A_{ss} + A_{sz}A_{zs}\right)e_s^h + A_{sz}b_P\Delta_P, \end{aligned}$$

which is the second coordinate.

We now prove Eq. 65 by induction. For $h = 0$, the right-hand side is the empty sum and equals zero, matching Eq. 61. If the claim holds at step h, then

$$\begin{aligned} e^{h+1} &\preceq B e^h + c\Delta_P \\ &\preceq B\left(\sum_{\ell=0}^{h-1} B^\ell\right) c\Delta_P + c\Delta_P \\ &= \left(\sum_{\ell=0}^{h} B^\ell\right) c\Delta_P. \end{aligned}$$

This completes the induction. □

To control rewards, let $c_z, c_s \geq 0$ be constants such that the coupled one-step expected rewards satisfy

$$\left|\mathbb{E}[R_P^{t+h} - R_{\widetilde{P}}^{t+h}]\right| \leq c_z e_z^h + c_s e_s^h. \tag{66}$$

Under additive rewards that are Lipschitz in exposure and bounded across local states, one may take

$$c_z = \sum_{i=1}^{N} L_{r,i}, \quad c_s = R_{\text{span}},$$

where R_{span} is an upper bound on the one-step reward change caused by coupled state and continuation-action disagreement. Define $c_r := \begin{pmatrix} c_z \\ c_s \end{pmatrix}$.

Proof of Theorem 3. By the definition of rollout value,

$$\left|Q_{t,H}(a; P) - Q_{t,H}(a; \widetilde{P})\right| \leq \sum_{h=0}^{H-1} \gamma^h \left|\mathbb{E}[R_P^{t+h} - R_{\widetilde{P}}^{t+h}]\right|. \tag{67}$$

Using Eq. 66 and Lemma 3,

$$\left|\mathbb{E}[R_P^{t+h} - R_{\widetilde{P}}^{t+h}]\right| \leq c_r^\top e^h \leq c_r^\top \left(\sum_{\ell=0}^{h-1} B^\ell\right) c\Delta_P. \tag{68}$$

Substituting Eq. 68 into Eq. 67 gives

$$\left|Q_{t,H}(a; P) - Q_{t,H}(a; \widetilde{P})\right| \leq C_H \Delta_P, \tag{69}$$

where

$$C_H := \sum_{h=0}^{H-1} \gamma^h c_r^\top \left(\sum_{\ell=0}^{h-1} B^\ell\right) c. \tag{70}$$

Since $H < \infty$ and all entries of B, c, and c_r are finite, $C_H < \infty$. Recalling that $\Delta_P = \|P - \widetilde{P}\|_{1,\infty}$ proves Eq. 10.

Suppose additionally that the coupled recursion is contractive: $\rho(B) < 1$, where $\rho(B)$ is the spectral radius of B. Because B is nonnegative,

$$\sum_{\ell=0}^{\infty} B^\ell = (I - B)^{-1}$$

exists and is entrywise nonnegative. Therefore,

$$C_H \leq \sum_{h=0}^{H-1} \gamma^h c_r^\top (I - B)^{-1} c \leq \frac{c_r^\top (I - B)^{-1} c}{1 - \gamma}. \tag{71}$$

The right-hand side is independent of H, proving the contractive-dynamics claim. □

Corollary 3 (Explicit contractive rollout constant)**.** *Under $\rho(B) < 1$, Theorem 3 holds uniformly over finite rollout horizons with*

$$C_\infty := \frac{c_r^\top (I - B)^{-1} c}{1 - \gamma}. \tag{72}$$

Proof. This is exactly Eq. 71. □

C.9 Proof of Graph-Net Approximation

The confidence set is generally infinite, while the practical planner evaluates only a finite candidate set $\mathcal{G}_t \subseteq \mathcal{C}_t$. Recall the covering condition

$$\sup_{P \in \mathcal{C}_t} \inf_{\widetilde{P} \in \mathcal{G}_t} \|P - \widetilde{P}\|_{1,\infty} \leq \rho_t.$$

Lemma 4 (Graph-net approximation)**.** For every action $a \in \mathcal{A}_M$,

$$0 \leq \sup_{P \in \mathcal{C}_t} Q_{t,H}(a; P) - \max_{\widetilde{P} \in \mathcal{G}_t} Q_{t,H}(a; \widetilde{P}) \leq C_H \rho_t. \tag{73}$$

Proof. The lower bound follows immediately from $\mathcal{G}_t \subseteq \mathcal{C}_t$.

For the upper bound, fix an arbitrary $\epsilon > 0$. By the definition of supremum, there exists $P_{a,\epsilon} \in \mathcal{C}_t$ such that

$$Q_{t,H}(a; P_{a,\epsilon}) \geq \sup_{P \in \mathcal{C}_t} Q_{t,H}(a; P) - \epsilon. \tag{74}$$

By the covering property, there exists $\widetilde{P}_{a,\epsilon} \in \mathcal{G}_t$ satisfying $\|P_{a,\epsilon} - \widetilde{P}_{a,\epsilon}\|_{1,\infty} \leq \rho_t$. Theorem 3 gives

$$Q_{t,H}(a; P_{a,\epsilon}) \leq Q_{t,H}(a; \widetilde{P}_{a,\epsilon}) + C_H \rho_t.$$

Combining this inequality with Eq. 74,

$$\begin{aligned} \sup_{P \in \mathcal{C}_t} Q_{t,H}(a; P) &\leq Q_{t,H}(a; \widetilde{P}_{a,\epsilon}) + C_H \rho_t + \epsilon \\ &\leq \max_{\widetilde{P} \in \mathcal{G}_t} Q_{t,H}(a; \widetilde{P}) + C_H \rho_t + \epsilon. \end{aligned}$$

Letting $\epsilon \downarrow 0$ proves the upper bound. □

C.10 Proof of Uniform Monte-Carlo Concentration

For graph P and action a, let $G_{t,H}^{(k)}(a; P)$ denote the discounted return of rollout k:

$$G_{t,H}^{(k)}(a; P) := \sum_{h=0}^{H-1} \gamma^h r\left(s_{P,k}^{t+h}, z_{P,k}^{t+h}, a_{P,k}^{t+h}\right). \tag{75}$$

By boundedness of expected one-step rewards,

$$0 \leq G_{t,H}^{(k)}(a; P) \leq R_H := R_{\max} \frac{1 - \gamma^H}{1 - \gamma}. \tag{76}$$

The Monte-Carlo estimator is

$$\widehat{Q}_{t,H}^{K_t}(a; P) := \frac{1}{K_t} \sum_{k=1}^{K_t} G_{t,H}^{(k)}(a; P). \tag{77}$$

Lemma 5 (Uniform Monte-Carlo rollout concentration)**.** Condition on $\mathcal{F}_t$ and let $\mathcal{I}_t$ be a finite collection of graph–action pairs evaluated by the theoretical planner, with $L_t := |\mathcal{I}_t|$. Then, with conditional probability at least $1 - \delta/T$,

$$\left|\widehat{Q}_{t,H}^{K_t}(a; P) - Q_{t,H}(a; P)\right| \leq b_{t,K} := R_H \sqrt{\frac{\log(2TL_t/\delta)}{2K_t}} \tag{78}$$

simultaneously for every $(a, P) \in \mathcal{I}_t$.

Proof. Fix $(a, P) \in \mathcal{I}_t$. Conditional on $\mathcal{F}_t$, the simulated returns $\{G_{t,H}^{(k)}(a; P)\}_{k=1}^{K_t}$ are independent, have common mean $Q_{t,H}(a; P)$, and lie in $[0, R_H]$. Hoeffding's inequality gives, for every $u > 0$,

$$\Pr\left(\left|\widehat{Q}_{t,H}^{K_t}(a; P) - Q_{t,H}(a; P)\right| > u \;\middle|\; \mathcal{F}_t\right) \leq 2 \exp\left(-\frac{2K_t u^2}{R_H^2}\right). \tag{79}$$

Set $u = R_H \sqrt{\frac{\log(2TL_t/\delta)}{2K_t}}$. Then the right-hand side of Eq. 79 is $\delta/(TL_t)$. A union bound over the L_t pairs proves Eq. 78. Removing the conditioning gives the same unconditional probability statement. □

For the exhaustive theoretical planner, one may take

$$\mathcal{I}_t = \mathcal{A}_M \times \mathcal{G}_t, \qquad L_t = |\mathcal{A}_M||\mathcal{G}_t|.$$

For a practical beam-search planner, $\mathcal{I}_t$ contains only the evaluated pairs. In that case, $\varepsilon_{\mathrm{opt},t}$ must include both the beam-search omission gap and the residual optimization error relative to the full feasible action class.

C.11 Proof of Theorem 4

For compactness, define the exact optimistic value

$$U_t(a) := \sup_{P \in \mathcal{C}_t} Q_{t,H}(a; P), \tag{80}$$

its finite-net counterpart

$$U_t^{\mathrm{net}}(a) := \max_{P \in \mathcal{G}_t} Q_{t,H}(a; P), \tag{81}$$

and its Monte-Carlo approximation

$$\widehat{U}_t(a) := \max_{P \in \mathcal{G}_t} \widehat{Q}_{t,H}^{K_t}(a; P). \tag{82}$$

Proof. Work on an exploitation round t and on the event that $P^\star \in \mathcal{C}_t$ and Lemma 5 holds for all graph–action pairs required below.

Because the true graph belongs to the confidence set,

$$Q_{t,H}(a_t^\star; P^\star) \le U_t(a_t^\star). \tag{83}$$

By Lemma 4,

$$U_t(a_t^\star) \le U_t^{\mathrm{net}}(a_t^\star) + C_H \rho_t. \tag{84}$$

Uniform Monte-Carlo concentration implies

$$U_t^{\mathrm{net}}(a_t^\star) \le \widehat{U}_t(a_t^\star) + b_{t,K}. \tag{85}$$

The action-search guarantee Eq. 19 gives

$$\widehat{U}_t(a_t^\star) \le \max_{a \in \mathcal{A}_M} \widehat{U}_t(a) \le \widehat{U}_t(a^t) + \varepsilon_{\mathrm{opt},t}. \tag{86}$$

Applying the Monte-Carlo event to the selected action,

$$\widehat{U}_t(a^t) \le U_t^{\mathrm{net}}(a^t) + b_{t,K} \le U_t(a^t) + b_{t,K}, \tag{87}$$

where the final inequality uses $\mathcal{G}_t \subseteq \mathcal{C}_t$.

It remains to compare $U_t(a^t)$ with the value of the selected action under the true graph. For any $P \in \mathcal{C}_t$, Theorem 3 gives

$$\begin{aligned} Q_{t,H}(a^t; P) &\le Q_{t,H}(a^t; P^\star) + C_H \|P - P^\star\|_{1,\infty} \\ &\le Q_{t,H}(a^t; P^\star) + C_H \operatorname{diam}_{1,\infty}(\mathcal{C}_t), \end{aligned} \tag{88}$$

where the second inequality uses $P, P^\star \in \mathcal{C}_t$. Taking the supremum over P gives

$$U_t(a^t) \le Q_{t,H}(a^t; P^\star) + C_H \operatorname{diam}_{1,\infty}(\mathcal{C}_t). \tag{89}$$

Combining Eq. 83–Eq. 89 yields

$$\begin{aligned} Q_{t,H}(a_t^\star; P^\star) - Q_{t,H}(a^t; P^\star) \le C_H \operatorname{diam}_{1,\infty}(\mathcal{C}_t) + C_H \rho_t \\ + 2b_{t,K} + \varepsilon_{\mathrm{opt},t}, \end{aligned}$$

which is Eq. 11. □

C.12 Proof of Theorem 5

Proof. Let $\mathcal{T}_{\mathrm{opt}} := [T] \setminus \mathcal{T}_{\mathrm{exp}}$ be the exploitation rounds. Since each rollout value lies in $[0, R_H]$, the instantaneous truncated rollout regret is at most R_H. Hence

$$\mathfrak{R}_T^{(H)} \le R_H |\mathcal{T}_{\mathrm{exp}}| + \sum_{t \in \mathcal{T}_{\mathrm{opt}}} \left[Q_{t,H}(a_t^\star; P^\star) - Q_{t,H}(a^t; P^\star) \right]. \tag{90}$$

Consider the joint event on which the following hold uniformly:

1. the adaptive graph-recovery bound in Theorem 1;
2. confidence-set validity in Proposition 1;
3. Fisher restricted curvature in Theorem 2;
4. all roundwise Monte-Carlo events in Lemma 5.

Allocating fixed fractions of δ to these events and applying a union bound gives joint probability at least $1 - \delta$, after changing universal constants and logarithmic factors.

On this event, Proposition 1 gives $\mathrm{diam}_{1,\infty}(\mathcal{C}_t) \le 2\varepsilon_t^{\mathrm{conf}}$. Theorems 1 and 2 imply

$$\varepsilon_t^{\mathrm{conf}} \le C_{\mathrm{conf}} \frac{d \nu_z L_\sigma B_x}{\kappa_{\mathrm{exp}}} \sqrt{\frac{\log(2N^2 T/\delta)}{n_{\mathrm{exp}}(t) \vee 1}}. \tag{91}$$

Since $n_{\mathrm{exp}}(t) \asymp t^{2/3}$ after burn-in,

$$\varepsilon_t^{\mathrm{conf}} = \widetilde{\mathcal{O}}\left(\frac{d \nu_z L_\sigma B_x}{\kappa_{\mathrm{exp}}} t^{-1/3} \right). \tag{92}$$

For every exploitation round, Theorem 4 gives

$$\begin{aligned} &Q_{t,H}(a_t^\star; P^\star) - Q_{t,H}(a^t; P^\star) \\ &\le 2C_H \varepsilon_t^{\mathrm{conf}} + C_H \rho_t + 2b_{t,K} + \varepsilon_{\mathrm{opt},t}. \end{aligned} \tag{93}$$

By assumption, $\rho_t \le \widetilde{\mathcal{O}}(c_\rho t^{-1/3}), \varepsilon_{\mathrm{opt},t} \le \widetilde{\mathcal{O}}(c_{\mathrm{opt}} t^{-1/3})$. Moreover, $K_t = \widetilde{\Omega}(c_K t^{2/3})$ implies

$$b_{t,K} = R_H \sqrt{\frac{\log(2TL_t/\delta)}{2K_t}} \le \widetilde{\mathcal{O}}\left(\frac{R_H}{\sqrt{c_K}} t^{-1/3} \right). \tag{94}$$

Combining these rates with Eq. 92 gives

$$\begin{aligned} &Q_{t,H}(a_t^\star; P^\star) - Q_{t,H}(a^t; P^\star) \\ &\le \widetilde{\mathcal{O}}\left(\left[\frac{C_H d \nu_z L_\sigma B_x}{\kappa_{\mathrm{exp}}} + C_H c_\rho + \frac{R_H}{\sqrt{c_K}} + c_{\mathrm{opt}} \right] t^{-1/3} \right). \end{aligned} \tag{95}$$

The integral comparison gives

$$\sum_{t=1}^{T} t^{-1/3} \le 1 + \int_1^T x^{-1/3}\, dx \le \frac{3}{2} T^{2/3}. \tag{96}$$

Furthermore, $|\mathcal{T}_{\mathrm{exp}}| = n_{\mathrm{exp}}(T) = \Theta(T^{2/3})$. Substituting Eq. 95 and Eq. 96 into Eq. 90 proves

$$\mathfrak{R}_T^{(H)} \le \widetilde{\mathcal{O}}\left(\left[R_H + \frac{C_H d \nu_z L_\sigma B_x}{\kappa_{\mathrm{exp}}} + C_H c_\rho + \frac{R_H}{\sqrt{c_K}} + c_{\mathrm{opt}} \right] T^{2/3} \right).$$

This is Eq. 12. □

C.13 Proof of Corollary 1

Proof. Under persistent Fisher excitation, all post-burn-in reward-seeking transitions may be included in the estimator. Let $n_{\mathrm{all}}(t)$ denote the number of post-burn-in transitions available by round t. Since $n_{\mathrm{all}}(t) = \Theta(t)$, the graph-recovery argument gives

$$\varepsilon_{t,\mathrm{PE}}^{\mathrm{conf}} \le \widetilde{\mathcal{O}}\left(\frac{d\nu_z L_\sigma B_x}{\kappa_{\mathrm{PE}}} t^{-1/2} \right). \tag{97}$$

Consequently, $\mathrm{diam}_{1,\infty}(\mathcal{C}_t) \le 2\varepsilon_{t,\mathrm{PE}}^{\mathrm{conf}}$. The computational assumptions give $\rho_t = \widetilde{\mathcal{O}}(c_\rho t^{-1/2})$, $\varepsilon_{\mathrm{opt},t} = \widetilde{\mathcal{O}}(c_{\mathrm{opt}} t^{-1/2})$. Since $K_t = \widetilde{\Omega}(c_K t)$, $b_{t,K} = \widetilde{\mathcal{O}}\left(\frac{R_H}{\sqrt{c_K}} t^{-1/2} \right)$. The per-round regret on exploitation rounds is therefore bounded by

$$\widetilde{\mathcal{O}}\left(\left[\frac{C_H d\nu_z L_\sigma B_x}{\kappa_{\mathrm{PE}}} + C_H c_\rho + \frac{R_H}{\sqrt{c_K}} + c_{\mathrm{opt}} \right] t^{-1/2} \right). \tag{98}$$

By assumption, explicit exploration occurs on at most $O(\sqrt{T})$ rounds, and therefore contributes at most $O(R_H\sqrt{T})$. Finally,

$$\sum_{t=1}^{T} t^{-1/2} \le 1 + \int_1^T x^{-1/2}\, dx \le 2\sqrt{T}. \tag{99}$$

Summing Eq. 98, adding the exploration cost, and absorbing logarithmic factors proves

$$\mathfrak{R}_T^{(H)} \le \widetilde{\mathcal{O}}\left(\left[R_H + \frac{C_H d\nu_z L_\sigma B_x}{\kappa_{\mathrm{PE}}} + C_H c_\rho + \frac{R_H}{\sqrt{c_K}} + c_{\mathrm{opt}} \right] \sqrt{T} \right),$$

which proves. □

C.14 Proof of the Information–Regret Tradeoff

The scheduled rate can also be understood by optimizing the number of rounds devoted to guaranteed information acquisition.

Suppose that by time T, only $n_{\mathrm{inf}}(T)$ transitions are guaranteed to provide sparse derivative-weighted Fisher curvature. The columnwise graph confidence radius then scales as $\varepsilon_T^{\mathrm{conf}} = \widetilde{\mathcal{O}}\left(\frac{d\nu_z L_\sigma B_x}{\kappa_{\mathrm{exp}} \sqrt{n_{\mathrm{inf}}(T)}} \right)$. If this uncertainty persists over the decision horizon, rollout stability contributes at most $\widetilde{\mathcal{O}}\left(\frac{C_H d\nu_z L_\sigma B_x T}{\kappa_{\mathrm{exp}} \sqrt{n_{\mathrm{inf}}(T)}} \right)$ to cumulative regret. Meanwhile, the informative interventions themselves contribute at most $R_H n_{\mathrm{inf}}(T)$. Ignoring graph-net, Monte-Carlo, and action-search terms, the leading tradeoff is therefore

$$\Phi(n) := R_H n + \frac{C_H d\nu_z L_\sigma B_x T}{\kappa_{\mathrm{exp}} \sqrt{n}}. \tag{100}$$

Treating $n > 0$ as continuous, differentiate:

$$\Phi'(n) = R_H - \frac{1}{2} \frac{C_H d\nu_z L_\sigma B_x T}{\kappa_{\mathrm{exp}} n^{3/2}}.$$

Setting $\Phi'(n) = 0$ gives

$$n^{3/2} = \frac{C_H d\nu_z L_\sigma B_x}{2R_H \kappa_{\mathrm{exp}}} T. \tag{101}$$

Hence $n_{\mathrm{inf}}(T) \asymp \left(\frac{C_H d\nu_z L_\sigma B_x}{R_H \kappa_{\mathrm{exp}}} \right)^{2/3} T^{2/3}$, up to universal constants and logarithmic factors. Substituting this value into Eq. 100 yields $\Phi(n_{\mathrm{inf}}(T)) = \widetilde{\mathcal{O}}(T^{2/3})$, which explains the scheduled-exploration exponent in Theorem 5.

Consequences of the bounds. The graph-recovery contribution grows linearly with the maximum incoming sparsity d in columnwise ℓ_1 error and inversely with the reachable Fisher curvature. Thus, saturation affects performance through $\kappa_{\exp}$ or κ_{PE}, rather than through source-signal variation alone. The scheduled theorem applies to the version of N-GDOI with $n_{\exp}(t) \asymp t^{2/3}$; adaptive threshold-triggered exploration is a practical variant and is not claimed to instantiate the same schedule exactly. Finally, ρ_t, $b_{t,K}$, and $\varepsilon_{\mathrm{opt},t}$ must decay at the statistical rate. If any of these terms remains bounded away from zero on a linear number of rounds, its cumulative contribution is linear in T.

D Additional Experimental Details and Results

This section provides implementation details and additional results supporting the experiments in Section 5. We first specify the environment-generation procedure, baseline adaptations, and hyperparameter protocol. We then examine the temporal relationship between graph recovery and decision quality, diagnose the Fisher-information mechanism underlying N-GDOI, and evaluate robustness to alternative graph structures and exposure noise, alongside computational runtime details.

D.1 Environment Construction and Evaluation Protocol

Unless otherwise stated, all methods are evaluated on identical realizations of the propagation graph, local transition randomness, exposure innovations, and reward noise. We use common random numbers whenever feasible to reduce between-method variance. Hyperparameters are selected on a disjoint set of validation seeds and then fixed across all reported test seeds.

For each environment seed, we generate a directed four-community stochastic block model with no self-loops, maximum in-degree $d = 4$, and nonzero edge weights independently sampled from $[0.08, 0.28]$. The source-signal, exposure, local-transition, and reward functions are shared across methods. The sigmoid setting is calibrated so that both derivative-informative and low-derivative operating regions are reachable under feasible interventions.

D.2 Baseline Implementations and Hyperparameters

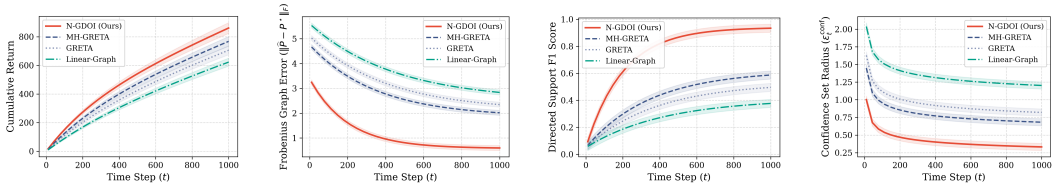
Myopic selects the feasible batch maximizing immediate expected reward under the observed current state. UCWhittle (Wang et al., 2023) and GINO-Q (Chen et al., 2026) are applied to the same local state and reward spaces while treating arms as conditionally independent; they therefore do not model graph-mediated spillovers.

Linear-Graph is a misspecified-estimation baseline that replaces the nonlinear exposure dynamics with a linearized model ($z^{t+1} - \alpha z^t = P^\top x^t$) while utilizing the common nonlinear rollout planner. It allows us to evaluate the performance loss caused by linear model misspecification independently of the planning architecture.

For graph-recovery comparisons, we evaluate only methods equipped with an explicit online graph estimator: Linear-Graph, an online estimated-graph adaptation of GRETA (Herlihy & Dickerson, 2023) (since the original method assumes a known graph), its multi-hop extension MH-GRETA, and N-GDOI. Independent-arm baselines and standard policy baselines do not maintain a propagation-graph estimate and are omitted from graph-recovery tables. All graph-aware methods receive the same observed state-exposure trajectories and use only information available before the corresponding decision round. N-GDOI uses a burn-in length of $t_{\mathrm{burn}} = 60$, rollout horizon $H = 5$, rollout count $K_t = 16$, and beam width $B_{\mathrm{beam}} = 4$.

D.3 Learning Curves for Return and Graph Recovery

Figure 3 shows the evolution of decision performance and graph recovery in the default partially saturated sigmoid environment. N-GDOI initially incurs an exploration cost during burn-in and scheduled Fisher-exploration rounds. As its graph confidence set contracts, however, its graph error decreases rapidly and its post-burn-in return surpasses the baselines. This temporal pattern is consistent with the theoretical sequence of graph recovery followed by graph-optimistic planning.



(a) Cumulative return. (b) Frobenius error. (c) Directed F1 score. (d) Confidence radius.

Figure 3: **Temporal evolution of control and graph learning.** Curves show continuous evolution over time steps under the default partially saturated sigmoid environment ($N = 30$, $M = 5$, $\alpha = 0.4$). N-GDOI successfully reduces graph uncertainty, improves structural recovery, and outperforms baselines in cumulative return. Solid lines and shaded regions represent means and 95% confidence intervals over matched seeds.

D.4 Robustness to Graph Topology and Exposure Noise

To evaluate the structural and algorithmic robustness of N-GDOI beyond the default network configuration, Table 5 extends our evaluation to diverse graph topologies and varying levels of exposure innovation noise. Specifically, we consider three representative directed graph classes:

- **Stochastic Block Model (SBM):** Tests performance under modular community structures characterized by dense intra-community connectivity and sparse inter-community interactions.
- **Directed Erdős–Rényi (ER):** Evaluates random sparse connectivity without localized community assumptions, challenging the estimator's ability to handle unstructured sparsity.
- **Directed Ring Graphs:** Represents localized, chain-like propagation dynamics where spillovers travel sequentially across immediate neighborhood chains.

Across all three topologies, N-GDOI consistently maintains its advantage in both post-burn-in return and directed support F1 recovery. This confirms that N-GDOI's Fisher-aware exploration mechanism is robust and not overfitted to specific network geometries.

Furthermore, we assess sensitivity to observation uncertainty by scaling the exposure innovation noise variance $\nu_z \in \{0.01, 0.03, 0.06\}$. As noise increases, the precision of online sparse regression naturally degrades across all methods. However, N-GDOI exhibits graceful degradation and superior resilience. Unlike baseline methods that rely on passive data collection or unguided trajectories—which easily fail under high observation noise—N-GDOI actively targets informative state-action regions via Fisher curvature maximization. This ensures that the collected samples remain informative despite noise contamination, thereby preserving high graph-recovery fidelity and robust rollout planning performance.

Table 5: **Robustness to graph topology and exposure noise.** Each entry reports post-burn-in return / directed F1. Results are mean ± standard error over matched seeds under varying network structures and innovation noise levels ν_z.

	Graph topology			**Exposure noise scale**		
Method	SBM	Directed ER	Directed ring	$\nu_z = 0.01$	$\nu_z = 0.03$	$\nu_z = 0.06$
MH-GRETA	$445.6 \pm 11.2/.62$	$432.1 \pm 12.5/.59$	$410.4 \pm 10.8/.52$	$462.5 \pm 9.4/.64$	$445.6 \pm 11.2/.62$	$415.2 \pm 13.1/.56$
Linear-Graph	$391.0 \pm 11.8/.41$	$382.4 \pm 10.5/.38$	$365.2 \pm 11.1/.35$	$410.8 \pm 8.9/.48$	$391.0 \pm 11.8/.41$	$358.4 \pm 12.4/.32$
N-GDOI	$\mathbf{492.2 \pm 8.4/.94}$	$\mathbf{481.5 \pm 9.1/.91}$	$\mathbf{458.7 \pm 8.8/.87}$	$\mathbf{515.3 \pm 7.1/.97}$	$\mathbf{492.2 \pm 8.4/.96}$	$\mathbf{452.6 \pm 9.5/.85}$

D.5 Empirical Truncated Rollout Regret

The theoretical guarantee in Section 4 concerns truncated rollout regret relative to the true-graph H-step rollout oracle, rather than realized one-step reward. Because the synthetic environment provides access to the true propagation graph $P^\star$, local transition kernels, and exposure innovation model, we can estimate this benchmark directly for all evaluated methods.

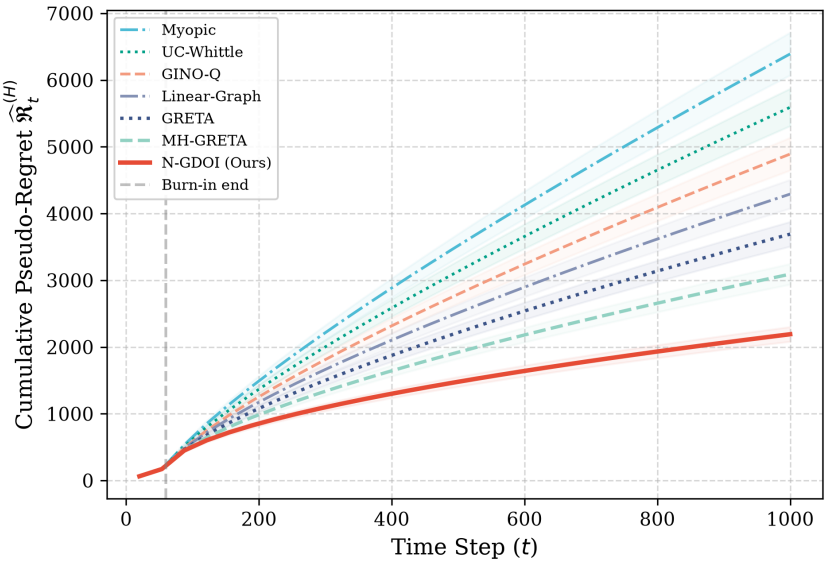


Figure 4: **Empirical truncated rollout pseudo-regret.** Cumulative H-step rollout pseudo-regret relative to a high-precision true-graph oracle under the default partially saturated sigmoid environment. N-GDOI achieves the slowest regret growth among all independent-arm and network-aware baselines.

At each decision round t, we construct an oracle action

$$\widehat{a}_t^\star \in \arg\max_{a \in \mathcal{A}_M} \widehat{Q}_{t,H}^{K_{\text{oracle}}}(a; P^\star), \tag{102}$$

where K_{oracle} is substantially larger than the rollout budget used by the tested methods. We then estimate the value of both $\widehat{a}_t^\star$ and the action a^t selected by each method (including MYOPIC, UC-WHITTLE, GINO-Q, LINEAR-GRAPH, GRETA, MH-GRETA, and N-GDOI) using an independent set of K_{eval} common-random-number rollouts under the true graph. The resulting empirical pseudo-regret is

$$\widehat{\mathfrak{R}}_T^{(H)} := \sum_{t=0}^{T-1} \left[\widehat{Q}_{t,H}^{K_{\text{eval}}}(\widehat{a}_t^\star; P^\star) - \widehat{Q}_{t,H}^{K_{\text{eval}}}(a^t; P^\star) \right]. \tag{103}$$

This quantity provides a high-precision Monte-Carlo approximation to the theoretical benchmark in Eq. 7.

Figure 4 reports the cumulative empirical truncated rollout pseudo-regret across all compared methods in the default partially saturated sigmoid environment. N-GDOI incurs an initial information-acquisition cost during the burn-in phase and scheduled exploration rounds, but its regret subsequently grows at a slower rate as graph uncertainty contracts. In contrast, independent-arm baselines (MYOPIC, UC-WHITTLE, GINO-Q) and network-aware or misspecified baselines exhibit consistently steeper regret trajectories due to their inability to effectively resolve nonlinear network spillovers or avoid saturated regions. These results empirically support the theoretical decomposition and demonstrate the superior decision efficiency of N-GDOI.