Stochastic Shortest Path Interdiction with Path-Only Feedback

Juan S. Borrero
Department of Industrial and Management Systems Engineering, University of South Florida, jsborrero@usf.edu

Denis Sauré, Iván Vidal-Romero
Department of Industrial Engineering, University of Chile, Santiago, Chile, dsaure@dii.uchile.cl, ivan.vidal@ug.uchile.cl

We study a sequential stochastic shortest-path interdiction problem in which an interdictor repeatedly blocks arcs to increase an evader's routing cost. Each period, a cost vector is drawn from a fixed, unknown distribution, and the evader observes the realized costs before routing; the interdictor observes only the evader's chosen path. We derive an instance-dependent logarithmic regret lower bound, computed through a semi-infinite information-allocation problem that admits an exact finite exponential-cone reformulation under regularity conditions. This informational censoring creates observational equivalence: distinct cost distributions can induce identical path observations under some interdiction actions. The resulting *dominated information* property means the actions needed to distinguish confounded environments are inferior in expected value, so pure posterior sampling and value-optimistic rules can fail to learn all relevant alternatives and incur linear regret. We propose Identifying-Exploration Greedy policies, which deliberately sample an identifying action set and achieve logarithmic regret on identifiable instances with a positive optimality gap. Experiments on constructed instances, generated networks, and a border-infiltration case study show when path-only learning succeeds, when it fails, and why deliberate information acquisition is essential for logarithmic regret.

Key words: network interdiction, sequential learning, multi-armed bandits, partial monitoring

1. Introduction

Motivation and Objective. In the security and defense context, interdiction denotes actions that block or degrade an adversary's use of a network, and it is a recurring tool against smuggling, trafficking, and other illicit routing activity that exploits transportation infrastructure (Steinrauf 1991, Gift 2010). These settings routinely pit an interdictor against an adversary who routes through a network, and the adversary (follower/evader) typically holds an informational advantage over the interdictor (leader): evaders, such as smugglers or other illicit actors, operate on the ground and observe the realized, day-to-day conditions of the network, i.e., its realized costs, *before* selecting a route. The interdictor, operating remotely or with limited surveillance capability, observes only the path traversed by the evader, and sees neither the realized costs nor the latent scenario that produced the evader's choice.

Motivated by these settings, in this paper we consider a sequential shortest path interdiction problem where the costs of using the network change over time. The interdictor does not know the underlying distribution generating the costs nor directly observes these costs; he/she only observes the paths selected by the evader (an informational regime we refer to as *path-only feedback*). By learning from the paths used by the evader, the interdictor seeks to sequentially reallocate the interdiction resources to maximize the evader's costs.

Shortest-path interdiction is a particular case of network interdiction, a class of problems studied extensively in the operations research literature; see Smith and Song (2020) for a survey. The original models consider a single-period interaction in which all parameters are deterministic and known to both agents (Fulkerson and Harding 1977, Malik et al. 1989); over the last decades, various extensions have introduced realistic features into these models (Cormican et al. 1998, Bayrak and Bailey 2008, Borrero et al. 2016, Kosmas et al. 2023). In the taxonomy of Smith and Song (2020), our evader is a *wait-and-see* follower, who observes the realization of the network's uncertainty before acting; we study the complementary side of that interaction, namely what an interdictor who does not share this advantage, and observes only the follower's realized route, can learn and exploit over repeated play. This paper studies stochastic sequential settings under path-only feedback and the challenges that this restricted observation structure creates.

Model and Research Question. We formulate this periodic interaction as a multi-armed bandit with combinatorial actions and a bilevel structure. In each period, the leader interacts with an evader in a known graph, whose costs are random variables, independent and identically distributed across periods. We assume that the leader acts first by blocking a subset of arcs. Subsequently, the evader, aware of the realized cost vector for that period, traverses the shortest path between two known nodes in the interdicted graph. The cost incurred by the evader is the leader's per-period payoff. To isolate the informational challenge, we assume the cost distribution has a finite, known support; the true probability mass function remains unknown to the leader.

Path-only feedback raises a basic question: how much must an interdictor explore, and which actions must it play to learn an unknown adversary when it observes only the route taken, and not the costs that produced it? We ask what performance is information-theoretically achievable in this regime; whether standard learning rules attain it; and how the answer changes when the realized costs are also observed.

Main Contributions. Our contributions, summarized in Table 1, are fourfold:

First, we formulate the sequential shortest path interdiction problem with path-only feedback and establish an instance-dependent limit on any policy consistent over the relevant model class. Following Lai and Robbins (1985), we show that limited feedback induces identifiability failures: distinct latent cost distributions can be observationally equivalent under certain interdiction actions.

Table 1 Summary of main results. Regret is measured against the full-information oracle over the planning horizon; rates are stated informally, omitting constants and lower-order factors. Under *path-only feedback* the leader observes only the evader's route, whereas under *full feedback* it also observes the realized costs.

Feedback	Policy / guarantee	Regret	Section
Path-only	Lower bound (any consistent policy)	Instance-dependent logarithmic floor	4
Full	Thompson sampling	Gap-free square-root rate	5
Path-only	Pure Thompson sampling / no diagnostic sampling	Linear	5, 6
Path-only	Identifying-Exploration Greedy	Logarithmic	6

We derive a logarithmic regret lower bound through a semi-infinite Lower Bound Problem (LBP), which prices the least costly asymptotic information allocation that a consistent policy must match against regret-relevant alternatives. For the full-simplex model class, local path-identifiability and a relative Slater condition further yield an exact finite exponential-cone reformulation.

Second, we isolate the central obstruction of the path-only feedback regime: *dominated information*. The interdiction actions required to distinguish between statistically confounded environments may be strictly dominated in expected value. On a three-path instance, pure Thompson sampling (Thompson 1933) incurs linear regret because it never selects the *diagnostic actions*, i.e., the interdiction actions whose feedback can distinguish two different underlying distributions. More generally, every policy that avoids those actions has linear worst-case regret across two observationally confounded environments. This includes pure value-optimistic rules when their selection mechanism never forces diagnostic sampling.

Third, we provide a self-contained full-feedback benchmark that isolates the role of censoring. When the interdictor observes the realized scenario, every period yields a draw from the latent distribution itself, so each observation updates the value of every action regardless of which one was played: the action-dependent information-acquisition problem disappears while the interdiction problem does not. This benchmark has the standard gap-free square-root horizon dependence, and full-feedback Thompson sampling matches it up to constants determined by the payoff geometry. The contrast with the linear regret above pinpoints the cost of path-only feedback.

Fourth, to break the dominated-information trap, we design the *Identifying-Exploration Greedy Policy*. By deliberately forcing exploration on an identifying set of actions (i.e., actions that can distinguish confounded environments), this policy sacrifices short-term value to recover the latent cost distribution. For a fixed identifiable instance with a positive optimality gap, we prove a logarithmic regret guarantee when the exploration multiplier is sufficiently large. The required threshold depends on the unknown instance, so the guarantee is an instance-dependent sufficiency result rather than a uniformly tuned policy.

Main Takeaways. For an interdictor who sees only the evader's route, the analysis yields three guidelines: do not learn solely by blocking whatever currently looks most damaging, since the

actions that reveal how the adversary behaves are precisely the unattractive ones; budget deliberate probes, which cost immediate value but keep the cumulative loss logarithmic on identifiable instances; and treat surveillance that reveals realized costs as purchasable performance, since it restores the standard square-root worst-case loss.

Organization of the Manuscript. Section 2 reviews the relevant literature, and Section 3 formulates the sequential shortest-path interdiction model under path-only feedback. Section 4 derives a lower bound for every consistent policy and introduces the running example on which dominated information is defined. Section 5 studies posterior sampling under path-only feedback: it establishes the full-feedback benchmark, characterizes the censored posterior and its Gibbs implementation, and shows that Thompson sampling can incur regret that grows linearly with the horizon. Section 6 develops an identifying-exploration policy with an instance-dependent logarithmic guarantee, and Section 7 reduces the semi-infinite lower-bound problem to an exact finite exponential-cone program for the full-simplex model class, making its constant computable. Section 8 reports numerical experiments, and Section 9 concludes. All proofs are in the Electronic Companion.

Notation. For $n \in \mathbb{N}$, we write $[n] := \{1, \dots, n\}$. The symbol $\mathbb{1}\{\cdot\}$ denotes the indicator function. We denote by Δ^{m-1} the probability simplex in $\mathbb{R}^m$, where m is the number of cost scenarios. All random elements are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The relative entropy (Kullback–Leibler divergence) between two distributions is denoted $D(\cdot\|\cdot)$, with the conventions $0\log(0/x) = 0$ for $x \geq 0$ and $x\log(x/0) = +\infty$ for $x > 0$. For two distributions μ, ν, we write $\mu \ll \nu$ when μ is absolutely continuous with respect to ν. For a convex set $\mathcal{X}$, $\mathrm{relint}(\mathcal{X})$ denotes its relative interior and $|\mathcal{X}|$ denotes the cardinality of a finite set $\mathcal{X}$. Vectors are column vectors unless stated otherwise, and $\mathbf{1}$ is the all-ones vector of appropriate dimension.

2. Literature Review

Our work builds on, and contributes to, the network interdiction and multi-armed bandit literatures. We position it relative to both.

Network Interdiction. In classical shortest path interdiction (Fulkerson and Harding 1977) and max flow interdiction (McMasters and Mustin 1970, Ghare et al. 1971), a leader aims to maximally disrupt a follower's flow through a network. While the original single-stage deterministic interdiction problems are strongly NP-complete (Ball et al. 1989, Wood 1993), integer programming-based solution approaches are frequently viable in practice (Wood 1993, Israeli and Wood 2002). For a comprehensive survey on deterministic network interdiction, see Smith and Song (2020).

Many works have relaxed the deterministic nature of the classical setup to incorporate various forms of uncertainty. Extensions include settings where the effectiveness of the interdiction effort is uncertain (Cormican et al. 1998, Morton 2010, Kang and Bansal 2023), settings where the network

topology is not known upfront (Hemmecke et al. 2003), and environments where the leader and evader face asymmetric costs (Bayrak and Bailey 2008). More recently, there has been a focus on asymmetric stochastic interdiction where the leader maximizes the expected or conditional value-at-risk of the evader's cost (Nguyen and Smith 2022, 2024), as well as robust optimization approaches to handle the evader's lack of knowledge (Azizi and Seifi 2024).

Sequential Interdiction & Incomplete Information. Our problem is closely tied to multi-period network interdiction, which has traditionally been studied in deterministic settings where agents interact repeatedly over time (Malaviya et al. 2012, Sefair and Smith 2016, Soleimani-Alyar and Ghaffari-Hadigheh 2017). Extending this to incomplete information, Borrero et al. (2016) introduced a sequential shortest-path interdiction setting where the leader must learn deterministic network costs by observing the evader's responses over time, casting repeated interdiction as an online-optimization problem that must balance exploiting the current best action against exploring to improve future decisions—the exploration-exploitation trade-off familiar from the multi-armed bandit literature (Smith and Song 2020). Subsequent research has explored limited forms of feedback (Borrero et al. 2019, Yang et al. 2021), differing cost structures between agents (Borrero et al. 2022), and strategic evaders (Ketkov and Prokopyev 2020). We contribute by giving path-only feedback, the regime in which the interdictor's information is most restricted, a formal regret-theoretic treatment, addressing the need—noted by Smith and Song (2020)—for defense strategies that learn from limited, realized feedback rather than from complete problem knowledge.

Closest to our setting is the recent work of Borrero et al. (2025), who study stochastic sequential shortest-path interdiction. Their model differs from ours in what each agent observes, and it does so on both sides of the interaction. First, their leader receives *semi-bandit* feedback: it observes the path chosen by the evader together with the realized costs of the traversed arcs. Second, their evader never sees cost realizations and routes according to expected costs. We invert both assumptions: our evader observes the realized cost vector before routing, while our leader observes only the resulting path—the *path-only* (or censored) regime. The evader's route is therefore the leader's sole channel of information about the latent costs, which changes both the regret that can be achieved and the structure of the policies that achieve it.

Multi-Armed Bandits & Partial Monitoring. Viewing each interdiction action as an arm, our sequential setting can be cast as a multi-armed bandit problem (Thompson 1933, Robbins 1952). Because actions are subsets of the graph's arcs, our problem connects to the combinatorial bandit literature (Cesa-Bianchi and Lugosi 2012, Gai et al. 2012, Modaresi et al. 2020). Asymptotically efficient bandit policies sample suboptimal arms at a logarithmic rate (Lai and Robbins 1985, Auer et al. 2002). Our full-feedback benchmark follows the corresponding minimax and Thompson

sampling theory; see Bubeck and Cesa-Bianchi (2012) for general bandit regret theory and Agrawal and Goyal (2013) for problem-independent Thompson bounds.

Path-only feedback instead places the problem within partial monitoring, whose general theory classifies worst-case regret through global and local observability (Bartók et al. 2014, Lattimore and Szepesvári 2020). We ask a complementary, instance-dependent question in the Lai–Robbins tradition rather than assigning the entire game to one minimax class.

Which form of observational equivalence holds matters. If distributions with disjoint optimal-action sets are observationally equivalent under *every* action, linear worst-case regret is unavoidable. If they agree only under the actions value-driven policies favor, the separating actions exist but are strictly suboptimal—*dominated information*—so posterior sampling and optimistic rules incur linear regret where forced exploration does not. We adapt Lai and Robbins (1985) and Modaresi et al. (2020) to obtain a Lower Bound Problem and propose a policy for path-feedback setting.

3. Model Formulation

This section formulates the sequential shortest-path interdiction problem under path-only feedback, describing in turn the network, the agents, the timing of the interaction, the information available to the leader, the full-information benchmark, and the performance measure.

Model Primitives. We consider a directed graph $G = (\mathcal{N}, A)$ with node set $\mathcal{N} = [n]$, designated source node 1, and sink node n. Two agents, an *interdictor* (or *leader*) and an *evader*, interact over a finite horizon of T periods. The interdictor's goal is to disrupt the evader's routing by blocking arcs; the evader, in turn, seeks the cheapest available route.

In each period $t \in [T]$, the leader selects an interdiction action B^t from the feasible set

$$\mathcal{B} := \{B \subseteq A : |B| \leq K\},$$

where $K \in \mathbb{N}$ is the *interdiction budget*, the maximum number of arcs the interdictor can block simultaneously. We assume throughout that every feasible interdiction leaves at least one directed path from the source 1 to the sink n, so the evader always has a feasible route.

Stochastic Arc Costs. Arc costs are random. In period t, the cost vector $c^t \in \mathbb{R}_+^{|A|}$ is drawn from a finite, known support $\{c_1, \ldots, c_m\}$, where $m \in \mathbb{N}$ is the number of *cost scenarios* and $c_k \in \mathbb{R}_+^{|A|}$ is the arc-cost vector under scenario k. The vectors $(c^t : t \in [T])$ are independent and identically distributed according to an unknown probability mass function $p = (p_1, \ldots, p_m) \in \mathcal{M}$, where $p_k := \mathbb{P}(c^t = c_k)$. Writing k^t for the realized scenario index, we thus have $c^t = c_{k^t}$. The finite-support assumption is a statistical primitive: the interdictor does not know the scenario probabilities p, but does know that $p \in \mathcal{M}$, where $\mathcal{M} \subseteq \Delta^{m-1}$ is a *model class*, and also knows the set $\{c_1, \ldots, c_m\}$.

Evader's Response. The evader observes both the interdiction action and the realized cost vector before routing. For an action B, let $\mathcal{P}(B)$ be the set of paths from 1 to n in the interdicted graph $(\mathcal{N}, A \setminus B)$. Under a fixed deterministic tie-breaking rule, let $S(B, c_k)$ denote the selected path, so

$$S(B, c_k) \in \arg\min_{S \in \mathcal{P}(B)} \sum_{a \in S} c_{k,a}.$$

The resulting path is therefore a deterministic function of the interdiction action and the latent cost scenario. We fix one tie-breaking rule throughout; the results hold for any such fixed deterministic rule because the induced map $S(\cdot, \cdot)$ is treated as a model primitive. The total cost incurred by the evader in scenario c_k after action B is $r_{B,k} := \sum_{a \in S(B,c_k)} c_{k,a}$.

Path-Only Feedback. The interdictor aims to maximize the cumulative expected cost incurred by the evader, but its information is limited: at the end of each period t, the leader observes only the path selected by the evader, $S^t := S(B^t, c^t)$, and observes neither the realized cost vector, the scenario index, nor the total path cost. We refer to this informational regime as *path-only feedback*.

Formally, for a candidate distribution $q \in \Delta^{m-1}$ and $B \in \mathcal{B}$, the observable path distribution is

$$\mathbb{P}(S \mid B, q) := \sum_{k=1}^{m} q_k \mathbb{1}\{S(B, c_k) = S\}, \qquad S \in \mathcal{P}(B).$$

We write $P_B(q)$ for this distribution over paths and $P_{B,S}(q) := \mathbb{P}(S \mid B, q)$ for its value at S. Only the paths that some scenario can actually induce carry probability, so we let

$$\mathcal{S}_B := \{S(B, c_k) : k \in [m]\} \subseteq \mathcal{P}(B)$$

denote the *response set* of action B, the set of paths the evader can select when B is played; every sum over path probabilities is indexed by $\mathcal{S}_B$. Distinct latent distributions are therefore observationally equivalent under an action whenever they induce the same distribution over paths.

For later use, define the path-aggregation matrix $M_B \in \{0,1\}^{|\mathcal{S}_B| \times m}$ by $(M_B)_{S,k} := \mathbb{1}\{S(B, c_k) = S\}$. The observation law is therefore the linear image $P_B(q) = M_B q$ of the latent distribution.

Full-Information Benchmark. For a distribution $q \in \Delta^{m-1}$ and an action B, let $r_B := (r_{B,1}, \ldots, r_{B,m})^\top$ denote the payoff vector and $\mu_B(q) := r_B^\top q$ the expected evader cost under q. The full-information value is $V(q) := \max_{B \in \mathcal{B}} \mu_B(q)$, and the corresponding set of full-information optimal interdiction actions is $\mathcal{B}^*(q) := \{B \in \mathcal{B} : \mu_B(q) = V(q)\}$. The single-period gap of action B under q is $\Delta_B(q) := V(q) - \mu_B(q)$, so $\Delta_B(q) = 0$ exactly for actions in $\mathcal{B}^*(q)$. We write $\Delta_{\min}(q) := \min_{B \notin \mathcal{B}^*(q)} \Delta_B(q)$ for the smallest positive gap under q, and, for a set $\mathcal{A} \subseteq \mathcal{B}$ of actions, $\Delta_{\min}^{\mathcal{A}}(q) := \min_{B \in \mathcal{A}} \Delta_B(q)$ for the smallest gap over that set. With the convention $\min \emptyset := 0$, either quantity is positive exactly when its indexing set is nonempty and contains no q-optimal action.

Policies and Regret. Let $H^t := (B^1, S^1, \dots, B^t, S^t)$ denote the observable path-feedback history through period t, with $H^0 := \emptyset$, and let $\mathcal{F}_t := \sigma(H^t)$ be its filtration. An *admissible policy* $\pi = (\pi_t)_{t \geq 1}$ is a sequence of possibly randomized decision rules. In period t, after observing H^{t-1}, rule π_t selects action $B \in \mathcal{B}$ with probability $\pi_t(B \mid H^{t-1})$. The policy's internal randomization is independent of the cost process. For each π, let $\mathbb{P}_p^\pi$ denote the law of the resulting action–scenario–response process and $\mathbb{E}_p^\pi$ expectation under that law. Let $\tau^\pi(B,t) := \sum_{s=1}^t \mathbb{1}\{B^s = B\}$ be the number of times action B has been played through period t. Although the leader observes neither k^t nor the realized costs, regret is evaluated against the full-information benchmark:

$$\mathcal{R}^\pi(T,p) = T \cdot V(p) - \mathbb{E}_p^\pi \left[\sum_{t=1}^T r_{B^t,k^t} \right].$$

Equivalently, since B^t is chosen before c^t is realized, $\mathcal{R}^\pi(T,p) = \sum_{B \in \mathcal{B}} \mathbb{E}_p^\pi[\tau^\pi(B,T)]\Delta_B(p)$. A policy is *consistent under* p if $\mathcal{R}^\pi(T,p) = o(T^\alpha)$ for every $\alpha > 0$. More generally, for a model class $\mathcal{M} \subseteq \Delta^{m-1}$, a policy is *consistent over* $\mathcal{M}$ if it is consistent under every $p \in \mathcal{M}$. When the relevant model class is clear from context, we simply say that the policy is consistent.

4. A Logarithmic Lower Bound on Path-Feedback Regret

This section establishes a fundamental limit on the performance of any policy consistent over a model class $\mathcal{M} \subseteq \Delta^{m-1}$: for a true distribution $p \in \mathcal{M}$, its asymptotic regret is bounded below by an instance-dependent multiple $\kappa^*_{\mathcal{M}}(p)$ of $\log T$. Later, in Section 7, we specialize to the full-simplex model class and show that, under local path-identifiability and a relative Slater condition, the resulting semi-infinite lower-bound problem admits a finite convex reformulation.

The bound rests on the classical change-of-measure argument of Lai and Robbins (1985) and Graves and Lai (1997), but the experiment is different: after action B the leader observes only the evader's selected path, not the realized cost scenario. The relevant observation law is therefore the aggregated path distribution $P_B(q)$, and for two distributions $p, q \in \Delta^{m-1}$ and an action $B \in \mathcal{B}$ the corresponding *path-feedback information number* is

$$I_B(p,q) := D(P_B(p) \| P_B(q)) = \sum_{S \in \mathcal{S}_B} P_{B,S}(p) \log \frac{P_{B,S}(p)}{P_{B,S}(q)}.$$

It depends on p and q only through the observable path laws: the cost scenarios enter only through the paths they induce. It drives both the lower bound below and the dominated-information diagnosis of Section 6.

THEOREM 1 **(Information Floor under Path-Only Feedback)**. *Fix $p, q \in \Delta^{m-1}$ and an admissible policy π. Assume $P_B(p) \ll P_B(q)$ for every $B \in \mathcal{B}$, $\mathcal{B}^*(p) \cap \mathcal{B}^*(q) = \emptyset$, and π is consistent under both p and q. Then*

$$\liminf_{T \to \infty} \frac{\sum_{B \in \mathcal{B}} \mathbb{E}_p^\pi[\tau^\pi(B,T)] I_B(p,q)}{\log T} \geq 1.$$

This theorem is an information-theoretic floor on exploration: to stay consistent against a confounding alternative q with a different optimal action, the policy must let the observed paths accumulate at least $\log T$ units of discriminating information, each play of B contributing $I_B(p,q)$. It does not yet say how that information is most cheaply bought; the next subsection turns it into a regret constant by spending suboptimal plays where they are most informative per unit of regret. Section A of the Electronic Companion proves Theorem 1 through an adaptive relative-entropy identity and a binary change-of-measure argument.

4.1. From Information Accumulation to a Regret Constant

Theorem 1 applies to alternatives q whose optimal actions are disjoint from those under p and for which $P_B(p) \ll P_B(q)$ for every $B \in \mathcal{B}$. To formulate the logarithmic lower-bound problem, we must first distinguish settings with costly exploration from settings with complete observational equivalence. If $I_B(p,q) = 0$ for every $B \in \mathcal{B}$, then p and q induce the same law over every path-feedback history, so no policy can be consistent over a model class containing both (Lemma EC.1 in Section A of the Electronic Companion).

We therefore focus on alternatives in the fixed model class $\mathcal{M}$ that are indistinguishable under actions in $\mathcal{B}^*(p)$ but distinguishable under at least one action outside $\mathcal{B}^*(p)$. Against such alternatives, information is available only through actions with a positive regret gap under p; we call them *costly confounding* alternatives. Excluding also alternatives with infinite information, which reveal a support mismatch and do not determine the finite logarithmic regret constant, we define

$$\begin{aligned}\mathcal{Q}_{\mathcal{M}}(p) := \big\{ q \in \mathcal{M} :\, & \mathcal{B}^*(p) \cap \mathcal{B}^*(q) = \emptyset, \quad \max_{B \in \mathcal{B}^*(p)} \{I_B(p,q)\} = 0, \\ & P_B(p) \ll P_B(q)\ \forall B \in \mathcal{B}, \quad \max_{B \in \mathcal{B}} \{I_B(p,q)\} > 0 \big\},\end{aligned}$$

the running example of Section 4.2 will exhibit a setting with $\mathcal{M} = \Delta^2$ and $\mathcal{Q}_{\mathcal{M}}(p) \neq \emptyset$. Minimizing regret subject to the information requirements that Theorem 1 imposes for each $q \in \mathcal{Q}_{\mathcal{M}}(p)$ therefore yields a valid lower bound on the regret of any consistent policy:

$$\begin{aligned}\kappa^*_{\mathcal{M}}(p) := \inf_{x_B \geq 0,\ B \notin \mathcal{B}^*(p)} \quad & \sum_{B \notin \mathcal{B}^*(p)} x_B \Delta_B(p) \\ \text{s.t.} \quad & \sum_{B \notin \mathcal{B}^*(p)} x_B I_B(p,q) \geq 1, \qquad \forall q \in \mathcal{Q}_{\mathcal{M}}(p).\end{aligned} \tag{1}$$

Here each $x_B \geq 0$ represents the asymptotic number of times action B is played per unit of $\log T$. Thus, if action B is played $x_B \log T$ times, it contributes $x_B \Delta_B(p) \log T$ regret under p and $x_B I_B(p,q) \log T$ path-feedback information against a fixed alternative q. Theorem 1 says that any consistent policy must accumulate at least $\log T$ units of such information against each $q \in \mathcal{Q}_{\mathcal{M}}(p)$.

Since actions in $\mathcal{B}^*(p)$ have zero information against these alternatives, this becomes the displayed constraint over $B \notin \mathcal{B}^*(p)$. The objective then minimizes the corresponding coefficient of $\log T$ in regret. If $\mathcal{Q}_{\mathcal{M}}(p) = \emptyset$, the constraint set is empty and $\kappa^*_{\mathcal{M}}(p) = 0$.

COROLLARY 1 **(Logarithmic Regret Lower Bound)**. *Fix a model class $\mathcal{M} \subseteq \Delta^{m-1}$ and $p \in \mathcal{M}$. If π is consistent over $\mathcal{M}$, then*

$$\liminf_{T\to\infty} \frac{\mathcal{R}^\pi(T,p)}{\log T} \geq \kappa^*_{\mathcal{M}}(p).$$

The extended-value constant $\kappa^*_{\mathcal{M}}(p) \in [0, +\infty]$ is the cheapest asymptotic cost of purchasing the information that Theorem 1 forces: it spends suboptimal plays only where they discriminate the costly confounding alternatives most efficiently, charging each unit of exploration to the regret gap it incurs. It is the path-feedback analogue of the classical Lai–Robbins constant, with the aggregated path information $I_B(p,q)$ in place of the single-arm Kullback–Leibler divergence. The constant is large when costly confounding alternatives are expensive to separate: the informative actions deliver little path-feedback information relative to the regret gaps they incur, so one unit of discriminating information requires a large regret expenditure. Section A of the Electronic Companion translates the information constraints into limiting play counts and minimizes their asymptotic regret.

4.2. A Running Example: The Three-Path Instance

We close the section with an instance that we return to throughout the paper. It is small enough to admit closed-form values and information numbers, yet it already exhibits the obstruction that drives the policy analysis of Sections 5 and 6: the actions that separate two confounding alternatives are exactly the actions that are worst in expected value.

The Instance. Consider the network of Figure 1, with three parallel source–sink paths $\Gamma_1, \Gamma_2, \Gamma_3$, each consisting of a single arc; we therefore identify each path with its arc when writing interdiction actions. There are three cost scenarios, and the triple on each arc of the figure lists its cost under c_1, c_2, and c_3. The interdiction budget is $K = 1$, so the feasible action set is

$$\mathcal{B} = \{B_0, B_1, B_2, B_3\}, \qquad B_0 := \emptyset, \quad B_i := \{\Gamma_i\},\ i = 1,2,3.$$

The evader chooses the available path with minimum realized cost, with ties broken in favor of the lowest-index available path. Table 2 reports, for each action–scenario pair, the path observed by the leader and the cost realized by the evader; the rows of the right block are the payoffs $r_{B_1} = (2,10,1)^\top$, $r_{B_2} = (2,1,10)^\top$, and $r_{B_3} = r_{B_0} = (2,1,1)^\top$. The left block shows that the path observed under B_1 and B_2 is the same in every scenario, while under B_3 and B_0 it varies.

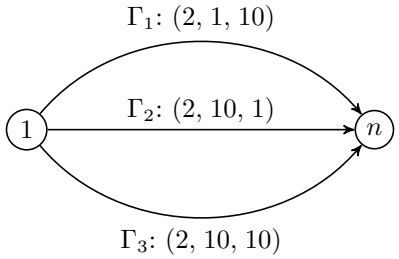

Figure 1 The three-path instance. Each arc is itself a source–sink path Γ_i, labeled with its costs under the scenarios (c_1, c_2, c_3); the interdiction budget is $K = 1$.

Table 2 Evader response in the three-path instance: observed path $S(B, c_k)$ and realized evader cost $r_{B,k}$ for each action–scenario pair.

	path $S(B,c_k)$			cost $r_{B,k}$		
	c_1	c_2	c_3	c_1	c_2	c_3
$B_1 = \{\Gamma_1\}$	Γ_2	Γ_2	Γ_2	2	10	1
$B_2 = \{\Gamma_2\}$	Γ_1	Γ_1	Γ_1	2	1	10
$B_3 = \{\Gamma_3\}$	Γ_1	Γ_1	Γ_2	2	1	1
$B_0 = \emptyset$	Γ_1	Γ_1	Γ_2	2	1	1

To avoid boundary issues in the arguments that follow, take $\mathcal{M} = \Delta^2$, fix $\varepsilon \in (0, 1/4)$, and consider the two interior latent distributions

$$p^\varepsilon := (1/2,\, 1/2 - \varepsilon,\, \varepsilon)\,, \qquad q^\varepsilon := (1/2,\, \varepsilon,\, 1/2 - \varepsilon)\,.$$

Under p^ε, direct substitution gives $\mu_{B_1}(p^\varepsilon) = 6 - 9\varepsilon$, $\mu_{B_2}(p^\varepsilon) = 3/2 + 9\varepsilon$, and $\mu_{B_3}(p^\varepsilon) = \mu_{B_0}(p^\varepsilon) = 3/2$. Since $\varepsilon < 1/4$, the inequality $6 - 9\varepsilon > 3/2 + 9\varepsilon$ holds, so $\mathcal{B}^*(p^\varepsilon) = \{B_1\}$. By symmetry, $\mu_{B_2}(q^\varepsilon) = 6 - 9\varepsilon$, $\mu_{B_1}(q^\varepsilon) = 3/2 + 9\varepsilon$, and $\mu_{B_3}(q^\varepsilon) = \mu_{B_0}(q^\varepsilon) = 3/2$, so $\mathcal{B}^*(q^\varepsilon) = \{B_2\}$. Hence $\mathcal{B}^*(p^\varepsilon) \cap \mathcal{B}^*(q^\varepsilon) = \emptyset$.

Value Ranking: The Diagnostic Actions Are Dominated. For any distribution $q = (q_1, q_2, q_3) \in \Delta^2$, the expected values are $\mu_{B_1}(q) = 2q_1 + 10q_2 + q_3$, $\mu_{B_2}(q) = 2q_1 + q_2 + 10q_3$, and $\mu_{B_3}(q) = \mu_{B_0}(q) = 2q_1 + q_2 + q_3 = 1 + q_1$. Therefore

$$\mu_{B_1}(q) - \mu_{B_3}(q) = 9q_2, \qquad \mu_{B_2}(q) - \mu_{B_3}(q) = 9q_3.$$

Whenever $q_2 > 0$ and $q_3 > 0$, both B_1 and B_2 strictly dominate B_3 and B_0 in expected value.

Dominated information and diagnostic actions. We now ask which actions carry information that separates p^ε from q^ε. Under B_1, the evader always chooses path Γ_2, regardless of the latent scenario, so $P_{B_1}(p^\varepsilon) = P_{B_1}(q^\varepsilon)$ and $I_{B_1}(p^\varepsilon, q^\varepsilon) = 0$. Under B_2, the evader always chooses path Γ_1, so $P_{B_2}(p^\varepsilon) = P_{B_2}(q^\varepsilon)$ and $I_{B_2}(p^\varepsilon, q^\varepsilon) = 0$. In contrast, under B_3, writing the path probabilities in the order (Γ_1, Γ_2) gives

$$P_{B_3}(p^\varepsilon) = (1 - \varepsilon, \varepsilon), \qquad P_{B_3}(q^\varepsilon) = (1/2 + \varepsilon, 1/2 - \varepsilon)\,.$$

Since $\varepsilon \in (0, 1/4)$, these two distributions are different and have the same support, and therefore

$$I_{B_3}(p^\varepsilon, q^\varepsilon) = D\left((1-\varepsilon, \varepsilon) \| (1/2+\varepsilon, 1/2-\varepsilon)\right) > 0.$$

The same calculation holds for B_0. The informative actions are therefore B_3 and B_0, neither of which is optimal under p^ε or under q^ε, and both of which are strictly value-dominated whenever $q_2 > 0$ and $q_3 > 0$. On this instance, informativeness and value point in opposite directions. This motivates the following definition.

Definition 1 (Dominated information). Fix a model class $\mathcal{M}$ and $p \in \mathcal{M}$. We say that the instance exhibits *dominated information* at p if there is an alternative $q \in \mathcal{Q}_{\mathcal{M}}(p)$ such that every action carrying strictly positive information against q is strictly suboptimal under both p and q; that is,

$$I_B(p, q) > 0 \qquad \Longrightarrow \qquad B \notin \mathcal{B}^*(p) \cup \mathcal{B}^*(q), \qquad B \in \mathcal{B}.$$

We call the actions with $I_B(p, q) > 0$ the *diagnostic actions* for the pair (p, q).

The computations above show that the three-path instance exhibits dominated information at p^ε, with q^ε as the confounding alternative and $\{B_0, B_3\}$ as the diagnostic actions. Section 5 shows that this configuration alone is enough to make posterior sampling fail, and Section 6 shows that it defeats value optimism as well before turning it into a policy design.

Remark 1. The three-path instance shows why the definition of $\mathcal{Q}_{\mathcal{M}}(p)$ searches over all feasible actions for distinguishing information. For p^ε and q^ε, the optimal-action sets are disjoint and $I_B(p^\varepsilon, q^\varepsilon) = 0$ for every $B \in \mathcal{B}^*(p^\varepsilon) \cup \mathcal{B}^*(q^\varepsilon)$, whereas $I_{B_3}(p^\varepsilon, q^\varepsilon) > 0$ even though $B_3 \notin \mathcal{B}^*(p^\varepsilon) \cup \mathcal{B}^*(q^\varepsilon)$. Thus the distinguishing action in the definition of $\mathcal{Q}_{\mathcal{M}}(p)$ cannot be restricted to $\mathcal{B}^*(q)$.

5. Posterior Sampling under Path-Only Feedback

The lower-bound results identify how much path-only feedback information a consistent policy must accumulate. We now ask whether the resulting learning difficulty is intrinsic to stochastic interdiction or instead arises from the form of feedback.

We proceed in four steps. We first calibrate against a full-feedback regime in which the leader observes the realized scenario after the evader traverses the interdicted network. This form of feedback removes the action-dependent information-acquisition problem while preserving the interdiction problem itself, and it turns the model into a standard bandit with the usual gap-free $\sqrt{T}$ rate. We then characterize the posterior induced by path-only feedback, and show that although marginal conjugacy fails, latent scenarios can be imputed through a conditionally conjugate Gibbs sampler that makes posterior sampling implementable. This machinery is not enough: on the running example of Section 4.2, even exact path-only posterior sampling incurs linear regret.

Throughout, we use a Bayesian Thompson sampling scheme with a Dirichlet prior on the latent scenario probabilities. Thompson sampling dates back to Thompson (1933) and has been developed as posterior sampling for sequential optimization (see, e.g., Russo and Van Roy 2014). Let $\alpha^0 = (\alpha_1^0, \dots, \alpha_m^0)$ with $\alpha_k^0 > 0$ for every k, and place the prior $p \sim \text{Dirichlet}(\alpha^0)$, writing $\bar{\alpha} := \sum_{k=1}^m \alpha_k^0$ for the total prior concentration. The Bayesian policies of this section therefore use the full simplex Δ^{m-1} as their prior support; a restricted model class $\mathcal{M}$ would instead require a prior supported on $\mathcal{M}$. All proofs of this section's results are deferred to Section B of the Electronic Companion.

5.1. Thompson Sampling under Full Feedback

Under full feedback, the leader observes k^t. Let $H^{\mathrm{F},t} := (B^1, k^1, \dots, B^t, k^t)$ denote the resulting history through period t, and let $N_k(t) := \sum_{s=1}^t \mathbb{1}\{k^s = k\}$ be the number of times scenario k has been observed through period t. Because the scenarios are i.i.d. with mass function p, the likelihood of $H^{\mathrm{F},t}$ is proportional to $\prod_{k=1}^m p_k^{N_k(t)}$. Dirichlet conjugacy therefore gives

$$p \mid H^{\mathrm{F},t} \sim \text{Dirichlet}\big(\alpha_1^0 + N_1(t), \dots, \alpha_m^0 + N_m(t)\big).$$

At the beginning of period t, full-feedback Thompson sampling π^{fTS} draws $\widetilde{p}^t \sim \text{Dirichlet}(\alpha_1^0 + N_1(t-1), \dots, \alpha_m^0 + N_m(t-1))$ and plays

$$B^t \in \arg\max_{B \in \mathcal{B}} \sum_{k=1}^m \widetilde{p}_k^t r_{B,k}.$$

After observing k^t, the policy increments the corresponding scenario count. A full-feedback policy follows the same stochastic-kernel convention as in Section 3, with $H^{\mathrm{F},t-1}$ in place of H^{t-1} and internal randomization independent of the cost process.

Full feedback removes the information-acquisition trade-off created by censoring, but regret remains measured against the full-information oracle $V(p)$. Also, every realized scenario updates the value estimates of all actions, regardless of the action played. The next two results calibrate this regime. Both specialize standard bandit regret theory to the finite-support interdiction model; see Bubeck and Cesa-Bianchi (2012) for the general minimax benchmark and Agrawal and Goyal (2013) for problem-independent bounds for Thompson sampling in classical multi-armed bandits.

Proposition 1 **(Minimax Lower Bound under Full Feedback)**. *There exists a fixed two-scenario full-feedback shortest-path interdiction instance such that, for every full-feedback policy π and every $T \geq 2$,*

$$\sup_{p \in \Delta^1} \mathcal{R}^\pi(T, p) \geq \frac{\sqrt{T}}{8e}.$$

Consequently, no full-feedback policy has gap-free worst-case regret $o(\sqrt{T})$.

The network, scenario costs, and feasible action set in Proposition 1 are fixed across horizons; the supremum is taken separately at each horizon, so, as is common in gap-free minimax constructions, the two latent distributions attaining it depend on T.

On the other hand, the upper bound that matches the rate of Proposition 1 is governed by the directions in which action payoffs differ over the simplex, measured by

$$\operatorname{diam}_{\perp}(r) := \max_{B,B' \in \mathcal{B}} \min_{a \in \mathbb{R}} \|r_B - r_{B'} - a\mathbf{1}\|_2,$$

where the inner minimization removes payoff components that are constant across scenarios, since these have the same value under every distribution in the simplex.

PROPOSITION 2 **(Regret of Full-Feedback Thompson Sampling)**. *There exists a universal constant $C < +\infty$ such that, for every $p \in \Delta^{m-1}$ and $T \geq 2$, full-feedback Thompson sampling satisfies*

$$\mathcal{R}^{\mathrm{fTS}}(T,p) \leq C \operatorname{diam}_{\perp}(r) \left[\sqrt{T} + 1 + \bar{\alpha} \log\left(1 + \frac{T}{\bar{\alpha}}\right)\right].$$

In particular, for fixed prior mass $\bar{\alpha}$, $\sup_{p \in \Delta^{m-1}} \mathcal{R}^{\mathrm{fTS}}(T,p) = O(\operatorname{diam}_{\perp}(r)\sqrt{T})$. If $0 \leq r_{B,k} \leq r_{\max}$ for every scenario and action, this further gives $\sup_{p \in \Delta^{m-1}} \mathcal{R}^{\mathrm{fTS}}(T,p) = O(r_{\max}\sqrt{mT})$.

Together the two propositions show that π^{fTS} is minimax-rate optimal in its horizon dependence for this regime; we make no claim about the leading constant. These results also imply that when the realized scenario is observed, the interdiction problem can be seen as a standard bandit and nothing pathological occurs. Therefore, whatever goes wrong under path-only feedback is a consequence of censoring, not of the interdiction structure.

5.2. The Censored Posterior

When extending Thompson sampling to path-only feedback, we must account for the fact that the observed path may be generated by more than one latent scenario. After playing B^t and observing S^t, the leader can only infer that the realized scenario belongs to the nonempty *compatible set*

$$\mathcal{Z}^t := \{k \in [m] : S(B^t, c_k) = S^t\}.$$

Equivalently, using the aggregation matrix M_{B^t}, we have $\mathcal{Z}^t = \{k \in [m] : (M_{B^t})_{S^t,k} = 1\}$. The likelihood contribution of the observed path in period t is therefore

$$\mathbb{P}(S^t \mid B^t, p) = \sum_{k \in \mathcal{Z}^t} p_k.$$

The following result characterizes the resulting posterior; Section B of the Electronic Companion derives it directly from the path-feedback likelihood.

PROPOSITION 3 **(Censored Posterior)**. *Fix an admissible policy and a possible realization h^t of the path-feedback history H^t. The posterior density of p conditional on $H^t = h^t$ is proportional to*

$$f_t(p) \propto \prod_{k=1}^{m} p_k^{\alpha_k^0 - 1} \prod_{s=1}^{t} \left(\sum_{k \in \mathcal{Z}^s} p_k \right), \qquad p \in \Delta^{m-1}. \tag{2}$$

When every observed path identifies its latent scenario, each compatible set is a singleton and the posterior remains Dirichlet. In general, however, it is not, because each observation contributes a sum of scenario probabilities rather than a single scenario count.

5.3. Gibbs Imputation and an Implementable Sampler

Since the censored posterior in (2) is not Dirichlet, sampling from it requires more than a conjugate update. We recover conditional conjugacy by introducing latent scenario assignments Z^s, one for each $s \in [t]$, where Z^s denotes the unobserved scenario index; necessarily $Z^s \in \mathcal{Z}^s$, or equivalently $S^s = S(B^s, c_{Z^s})$. This is the usual device for incomplete observations (Dempster et al. 1977); the resulting conditional sampler is a partially collapsed Gibbs construction (Geyer 1992). The latent distribution p is common to all periods and therefore carries no time index; its posterior law depends on t through the path-feedback history H^t.

PROPOSITION 4 **(Gibbs Conditionals for the Censored Posterior)**. *Fix $t \in [T]$. Given latent assignments $Z := (Z^s : s \in [t])$, define $n_k(Z) := \sum_{s=1}^{t} \mathbb{1}\{Z^s = k\}$. Then*

$$p \mid Z, H^t \sim \text{Dirichlet}\big(\alpha_1^0 + n_1(Z), \ldots, \alpha_m^0 + n_m(Z)\big),$$

and, for every $s \in [t]$ and $k \in [m]$, conditional on p,

$$\mathbb{P}(Z^s = k \mid p, H^t) = \frac{p_k \mathbb{1}\{k \in \mathcal{Z}^s\}}{\sum_{j \in \mathcal{Z}^s} p_j}.$$

Both conditionals are standard draws, so augmenting the state with Z makes posterior sampling implementable. Integrating p out of the second conditional yields a collapsed form that depends only on the remaining assignments, and sampling from it repeatedly defines a *sweep*: one pass in which every Z^s, $s \in [t-1]$, is resampled in turn, followed by a Dirichlet draw of p given the updated assignments. Section B.1 of the Electronic Companion states the collapsed conditional, gives the resulting sampler as Algorithm 2, and derives both conditional laws.

We reserve π^{TS} for the ideal path-only Thompson policy that draws exactly from the censored posterior before every decision and maximizes the sampled expected payoff. Writing $R := (R_t : t \in [T])$ for the number of sweeps performed before the draw in each period, we denote the finite-sweep implementation by $\pi^{\text{cTS}}(R)$. Under full feedback no imputation is needed, since k^t is observed, and the Dirichlet update of Section 5.1 is recovered.

REMARK 2. Run to stationarity before every Thompson draw, the Gibbs chain returns an exact sample from (2), and $\pi^{\mathrm{cTS}}(R)$ coincides in law with π^{TS}; the usual irreducibility and aperiodicity are understood on the compatible support of the augmented posterior; with a finite sweep count the draw is approximate. Section B.1 of the Electronic Companion collects implementation details.

5.4. The Limits of Posterior Sampling

Neither π^{TS} nor $\pi^{\mathrm{cTS}}(R)$ can create information that the played actions never reveal. In this sense, the following observation can be thought as the Bayesian counterpart of the observational equivalence obstruction in the lower-bound analysis.

REMARK 3. Under censored feedback, the posterior can only learn the components of p that are identifiable through the played actions and their compatible sets. In the path-only case, if two distributions p and q satisfy $M_B p = M_B q$ for every $B \in \mathcal{B}$, then every path-only feedback history has the same likelihood under p and q. Hence no Bayesian procedure, including Thompson sampling, can distinguish between them from path-only observations alone.

Any failure that persists under exact posterior sampling is therefore informational rather than computational. We now show that the failure is not confined to the fully unidentifiable case of Remark 3. On the running example of Section 4.2 the latent distribution is identifiable from the full action set, yet the ideal path-only Thompson policy π^{TS} has linear worst-case regret. The reason is dominated information (Definition 1): the diagnostic actions B_0, B_3 are strictly dominated in value, so they are never optimal for a nondegenerate draw and are never played.

PROPOSITION 5 **(Linear Regret of Path-Only Thompson Sampling)**. *In the three-path instance of Section 4.2, consider pure path-only Thompson sampling with a Dirichlet prior satisfying $\alpha_k^0 > 0$ for all k. Let $\widetilde{p}$ denote a generic draw from this prior. Suppose the Thompson sampler uses exact posterior samples under the path-only posterior and plays $B^t \in \arg\max_{B\in\mathcal{B}} \mu_B(\widetilde{p}^t)$ with deterministic tie-breaking. Then, under $p^\varepsilon = (1/2, 1/2-\varepsilon, \varepsilon)$ with $\varepsilon \in (0, 1/4)$, the expected regret is linear:*

$$\mathcal{R}^{\mathrm{TS}}(T, p^\varepsilon) = \left(\frac{9}{2} - 18\varepsilon\right) \mathbb{P}(\widetilde{p}_3 > \widetilde{p}_2)\, T.$$

The probability in this expression is strictly positive. In particular, the worst-case regret of pure path-only Thompson sampling over this instance family is $\Theta(T)$.

Section B.2 of the Electronic Companion proves the result by showing that the informative actions are never selected, so the posterior remains at its prior and a suboptimal action is played with constant probability. Section 6 shows that the same mechanism defeats value optimism, and then turns the diagnosis into a policy design.

6. Dominated Information and Identifying Exploration

Section 5 showed that pure path-only Thompson sampling can incur linear regret on the three-path instance, even with exact posterior sampling. This section turns that negative example into a policy design. The root cause is dominated information (Definition 1): the actions that distinguish two confounded environments can be strictly dominated in expected value, so a policy that only optimizes sampled or optimistic values may never collect the observations that separate them.

We proceed in four steps. First, we show that linear regret holds for *every* policy that avoids the diagnostic actions, not only for posterior sampling. Second, we show that policies based on value-optimism (particularly, UCB policies) can also fail to play diagnostic actions and thus incur linear regret. Third, we develop an identifiability notion to formalize what it means to gather *enough* information in our setting. Finally, we give a simple identifying-exploration policy that attains a rate-optimal regret by exploiting the fact that a consistent policy may play a suboptimal action $O(\log T)$ times without breaking the overall $\log T$ rate.

6.1. The Necessity of Playing Diagnostic Actions

The linear-regret result of Section 5.4 is not due to posterior sampling, but to never playing the diagnostic pair $\{B_0, B_3\}$, which we formalize next.

PROPOSITION 6 **(Linear Regret without Diagnostic Actions)**. *Consider the three-path instance and the two interior environments $p^\varepsilon = (1/2, 1/2-\varepsilon, \varepsilon)$ and $q^\varepsilon = (1/2, \varepsilon, 1/2-\varepsilon)$, with $\varepsilon \in (0, 1/4)$. Let π be any admissible policy that, with probability one under both environments, never plays B_0 or B_3. Then*

$$\max\{\mathcal{R}^\pi(T, p^\varepsilon), \mathcal{R}^\pi(T, q^\varepsilon)\} \geq \left(\frac{9}{4} - 9\varepsilon\right) T.$$

Thus, such a policy has linear worst-case regret.

Next, we use this observation to prove that policies that are optimistic with respect to uncertainty, such as UCB-type policies, face the same issue as Thompson sampling under path-only feedback.

6.2. The Price of Optimism in the Face of Uncertainty

Let $\widehat{P}_B(t-1)$ be the vector of empirical path frequencies of action B, and let $\mathrm{rad}_B(t) \geq 0$ be a confidence radius. For each t, define the confidence set

$$\Theta_t := \big\{q \in \mathcal{M} : \|M_B q - \widehat{P}_B(t-1)\|_\infty \leq \mathrm{rad}_B(t) \ \text{ for every sampled } B\big\}.$$

Unsampled actions impose no constraint. Whenever Θ_t is nonempty, the simple Path-Only Value-UCB policy π^{UCB} plays $B^t \in \arg\max_{B \in \mathcal{B}} \sup_{q \in \Theta_t} r_B^\top q$, using a fixed tie-breaking rule. On histories

for which Θ_t is empty, the policy uses any fixed measurable fallback rule; this contingency does not arise under the hypothesis of Proposition 7. For this rule, define the initial optimistic set by

$$\mathcal{B}^0_{\mathrm{UCB}} := \arg\max_{B\in\mathcal{B}} \sup_{q\in\mathcal{M}} r_B^\top q.$$

The set $\mathcal{B}^0_{\mathrm{UCB}}$ depends on the model class, but not on p. Finally, we say an action B is *scenario-blind over* $\mathcal{M}$ if there is a path S_B such that $P_{B,S_B}(q)=1$ for every $q\in\mathcal{M}$; equivalently, every scenario that receives positive mass under some law in $\mathcal{M}$ induces the same path S_B whenever B is played.

PROPOSITION 7 **(Absorption at Scenario-Blind Actions)**. *If every action in $\mathcal{B}^0_{\mathrm{UCB}}$ is scenario-blind over $\mathcal{M}$, then, under every $p\in\mathcal{M}$, $B^t\in\mathcal{B}^0_{\mathrm{UCB}}$ for all $t\in[T]$ with probability one. In particular, if $\mathcal{B}^0_{\mathrm{UCB}}$ is a singleton with unique element $\widehat{B}$, then $\mathcal{R}^{\mathrm{UCB}}(T,p)=T\Delta_{\widehat{B}}(p)$.*

The conclusion is independent of the particular nonnegative confidence radii: along the realized histories, observations from scenario-blind actions leave every distribution in $\mathcal{M}$ feasible.

On the running example with $\mathcal{M}=\Delta^2$, direct computation gives $\mathcal{B}^0_{\mathrm{UCB}}=\{B_1,B_2\}$: the initial optimistic values of these actions are 10, whereas those of B_0 and B_3 are 2. Both B_1 and B_2 are scenario-blind, as shown in Table 2. Proposition 7 therefore implies that Value-UCB never plays B_0 or B_3. Applying Proposition 6 gives, for every $\varepsilon\in(0,1/4)$,

$$\max\{\mathcal{R}^{\mathrm{UCB}}(T,p^\varepsilon),\mathcal{R}^{\mathrm{UCB}}(T,q^\varepsilon)\}\geq\left(\frac{9}{4}-9\varepsilon\right)T.$$

Thus, simple Value-UCB has linear worst-case regret on the three-path instance.

6.3. Identifying-Exploration Design

The negative examples above show that under path-only feedback the leader must 'buy' information from actions that are not value-maximizing. We present a simple sufficient design that does so, fixing a set of identifying actions, sampling each logarithmically often, and then playing greedily against a least-squares estimate of the latent distribution.

Let $\mathcal{M}\subseteq\Delta^{m-1}$ be the model class of Section 3. Throughout this section, we assume that any restrictions beyond the simplex are linear equalities, so $\mathcal{M}$ is the intersection of an affine subspace with Δ^{m-1}; namely, a polytope. Define

$$\mathcal{V}_{\mathcal{M}} := \mathrm{span}\{p'-p'' : p',p''\in\mathcal{M}\}.$$

Note that $\mathcal{V}_{\mathcal{M}}$ records how two candidate distributions in $\mathcal{M}$ can differ. Because both distributions sum to one, every $v\in\mathcal{V}_{\mathcal{M}}$ satisfies $\mathbf{1}^\top v=0$. When $\mathcal{M}=\Delta^{m-1}$, there are no additional restrictions and $\mathcal{V}_{\mathcal{M}}=\{v\in\mathbb{R}^m:\mathbf{1}^\top v=0\}$. Prior information narrows this space. For example, if the probability of scenario 1 is known, then every $v\in\mathcal{V}_{\mathcal{M}}$ also satisfies $v_1=0$.

We say that $\mathcal{E} \subseteq \mathcal{B}$ is *identifying over* $\mathcal{M}$ if, for all $p, q \in \mathcal{M}$,

$$M_B q = M_B p \quad \forall B \in \mathcal{E} \qquad \Longrightarrow \qquad q = p.$$

Repeated plays of $\mathcal{E}$ let the leader estimate the path laws $(M_B p)_{B\in\mathcal{E}}$; the set is identifying when these determine p uniquely within $\mathcal{M}$. Note that, in the running example, no single action is identifying over Δ^2 because c_1 and c_2 induce the same path under every action. If $q_1 = 1/2$ is known, however, $\mathcal{E} = \{B_3\}$ identifies the restricted class: its Γ_2 probability is q_3, and normalization gives $q_2 = 1/2 - q_3$.

Next, we provide a more useful characterization of identifiability. For any $\mathcal{E} \subseteq \mathcal{B}$, define the stacked observation matrix

$$M_{\mathcal{E}} := \begin{bmatrix} (M_B)_{B\in\mathcal{E}} \\ \mathbf{1}^\top \end{bmatrix}$$

where the upper block stacks the matrices M_B vertically.

Lemma 1 **(Identifying-Subspace Equivalence)**. *Suppose that $\mathcal{M}$ is nonempty and convex. A subset $\mathcal{E} \subseteq \mathcal{B}$ is identifying over $\mathcal{M}$ if and only if the restriction of $M_{\mathcal{E}}$ to $\mathcal{V}_{\mathcal{M}}$ is injective; that is,*

$$\ker(M_{\mathcal{E}}) \cap \mathcal{V}_{\mathcal{M}} = \{0\}.$$

When $\mathcal{M} = \Delta^{m-1}$, these conditions are equivalent to $M_{\mathcal{E}}$ having full column rank m.

Over the full simplex, the subspace condition of Lemma 1 admits a left-inverse certificate that can be more useful computationally.

Corollary 2 **(Left-Inverse Certificate for Identifying Sets)**. *Let $\mathbb{I}_m$ denote the $m \times m$ identity matrix. A subset $\mathcal{E} \subseteq \mathcal{B}$ is identifying over Δ^{m-1} if and only if there exist matrices $W_B \in \mathbb{R}^{m\times|\mathcal{S}_B|}$ for $B \in \mathcal{E}$ and a vector $w_0 \in \mathbb{R}^m$ such that*

$$\sum_{B\in\mathcal{E}} W_B M_B + w_0 \mathbf{1}^\top = \mathbb{I}_m.$$

Remark 4. An identifying set over Δ^{m-1} exists if and only if the full action set is identifying: $M_{\mathcal{B}}$ has column rank m, equivalently the incidence vectors of all path partitions, together with $\mathbf{1}^\top$, span $\mathbb{R}^m$. Pairwise scenario separation is necessary but not sufficient, because the path-frequency equations induced by different actions may still be linearly dependent. If the resulting rank deficiency yields two observationally equivalent distributions with disjoint optimal sets, Lemma EC.1 rules out consistency over any model class containing both distributions.

6.4. Identifying-Exploration Greedy Policy

Fix an identifying set $\mathcal{E} \subseteq \mathcal{B}$. For $B \in \mathcal{E}$ and $S \in \mathcal{S}_B$, let $N^{\mathrm{fe}}_{B,S}(t)$ be the number of forced-exploration periods up to time t in which action B was played and path S was observed. Let

$$\tau^{\mathrm{fe}}(B,t) := \sum_{S \in \mathcal{S}_B} N^{\mathrm{fe}}_{B,S}(t).$$

When $\tau^{\mathrm{fe}}(B,t) > 0$, define $\widehat{P}^{\mathrm{fe}}_{B,S}(t) := N^{\mathrm{fe}}_{B,S}(t)/\tau^{\mathrm{fe}}(B,t)$. The least-squares estimator is

$$\widehat{p}^{\,t} \in \arg\min_{q \in \mathcal{M}} \sum_{B \in \mathcal{E}:\tau^{\mathrm{fe}}(B,t)>0} \left\| \widehat{P}^{\mathrm{fe}}_{B}(t) - M_B q \right\|_2^2.$$

Any minimizer can be used; restricting it to $\mathcal{M}$ rather than to Δ^{m-1} is what lets the estimator exploit known or linked scenario probabilities. For a multiplier $\beta > 0$, define the logarithmic exploration target $g_t := \max\{1, \lceil \beta \log t \rceil\}$. Fix an arbitrary deterministic ordering of the actions in $\mathcal{E}$; this ordering is used only to choose among under-explored actions. The proposed identifying-exploration greedy policy π^{IE} is described in Algorithm 1.

Policy π^{IE} guarantees logarithmic regret, as stated below in Theorem 2. It is an instance-dependent logarithmic guarantee for a fixed $p \in \mathcal{M}$ with positive gap $\Delta_{\min}(p)$, provided that a fixed set $\mathcal{E}$ identifying over $\mathcal{M}$ is sampled with a sufficiently large multiplier β. Three instance quantities govern the required multiplier: the payoff radius

$$\rho_*(p) := \max_{B \notin \mathcal{B}^*(p)} \min_{B^* \in \mathcal{B}^*(p)} \|r_{B^*} - r_B\|_2,$$

which is positive and finite because $\mathcal{B}$ is finite and $\Delta_{\min}(p) > 0$, the design conditioning

$$\underline{\sigma}_{\mathcal{E}} := \inf\{\|M_{\mathcal{E}} v\|_2 : v \in \mathcal{V}_{\mathcal{M}},\ \|v\|_2 = 1\} > 0,$$

whose positivity follows from Lemma 1 and compactness of the unit sphere in $\mathcal{V}_{\mathcal{M}}$, and the largest response-set size $s_{\max} := \max_{B \in \mathcal{E}} |\mathcal{S}_B|$.

THEOREM 2 **(Logarithmic Regret of Identifying-Exploration Greedy)**. *Fix $p \in \mathcal{M}$ with $\dim(\mathcal{V}_{\mathcal{M}}) \geq 1$, $\Delta_{\min}(p) > 0$, and let $\emptyset \neq \mathcal{E} \subseteq \mathcal{B}$ be identifying over $\mathcal{M}$. If*

$$\beta > \frac{2 s_{\max} |\mathcal{E}| \rho_*(p)^2}{\underline{\sigma}_{\mathcal{E}}^2 \Delta_{\min}(p)^2}, \qquad \textit{then} \qquad \mathcal{R}^{\mathrm{IE}}(T,p) \leq g_T \sum_{B \in \mathcal{E}} \Delta_B(p) + C(p,\mathcal{E},\beta),$$

with $C(p,\mathcal{E},\beta) < +\infty$ independent of T. In particular, since $g_T = O(\log T)$, $\mathcal{R}^{\mathrm{IE}}(T,p) = O(\log T)$.

REMARK 5. The theorem is stated for a fixed set $\mathcal{E}$ identifying over $\mathcal{M}$. Section D of the Electronic Companion gives an exact mixed-integer linear programming (MILP) formulation for selecting a low-cost identifying set. The full-simplex formulation returns a set identifying over Δ^{m-1} and hence over any $\mathcal{M} \subseteq \Delta^{m-1}$; the direction-space implementation described there can exploit restrictions defining $\mathcal{M}$. Alternatively, one may take all actions whenever $\mathcal{B}$ is identifying over $\mathcal{M}$, or use another experiment-design procedure.

Algorithm 1 Identifying-Exploration Greedy Policy

Input: horizon T, actions $\mathcal{B}$, model class $\mathcal{M}$, matrices $(M_B)_{B\in\mathcal{B}}$, payoffs $(r_B)_{B\in\mathcal{B}}$, identifying set $\mathcal{E}$ with a fixed ordering, multiplier $\beta>0$; $N^{\mathrm{fe}}_{B,S}(0)\leftarrow 0$ for all $B\in\mathcal{E}$, $S\in\mathcal{S}_B$.
for $t=1,\dots,T$ **do**
  $g_t\leftarrow\max\{1,\lceil\beta\log t\rceil\}$.
  if $\tau^{\mathrm{fe}}(B,t-1)<g_t$ for some $B\in\mathcal{E}$ **then**
    Play the first under-explored action B^t in the fixed ordering, observe path S^t, and increment $N^{\mathrm{fe}}_{B^t,S^t}$. (*forced exploration*)
  else
    Play $B^t\in\arg\max_{B\in\mathcal{B}} r_B^\top\widehat{p}^{\,t-1}$ at the least-squares estimate $\widehat{p}^{\,t-1}$, and observe the path. (*greedy*)
  end if
end for

Counters not named in a step are carried forward unchanged; only forced-exploration plays update them.

In summary, path-only feedback changes the nature of exploration. Information can be hidden in actions that are dominated in value, so pure posterior sampling and pure value optimism may fail to separate confounding environments. Forcing logarithmic sampling of an identifying action set restores learnability under a positive optimality gap and identifiability over the model class.

7. Computing the Lower-Bound Constant

The lower-bound problem (1) is semi-infinite: it has one information constraint for every alternative in $\mathcal{Q}_{\mathcal{M}}(p)$, and these generally form a continuum. This section reduces it to a finite convex program, making $\kappa^*(p)$ computable on a fixed instance and the calibration of Section 8.3 possible.

Throughout we specialize to the full-simplex model class $\mathcal{M}=\Delta^{m-1}$ and abbreviate $\mathcal{Q}(p):=\mathcal{Q}_{\Delta^{m-1}}(p)$ and $\kappa^*(p):=\kappa^*_{\Delta^{m-1}}(p)$.

The reduction rests on one structural feature of path-only feedback: the observation law is a *linear* image of the latent distribution. Recall that the path-aggregation matrix M_B satisfies $P_B(q)=M_Bq$, and that $\mu_B(q)=r_B^\top q$. The conditions defining $\mathcal{Q}(p)$ and $I_B(p,q)$ therefore depend on the alternative q only through finitely many linear maps, which is what allows a continuum of constraints to collapse into finitely many.

Throughout this section we assume w.l.o.g. that $\mathcal{B}\setminus\mathcal{B}^*(p)\neq\emptyset$,[1] and that decision-changing alternatives be separable from p by at least one action. This rules out the complete-observational-equivalence case identified in the previous section.

Assumption 1 **(Local Path-Identifiability).** *There is no $q\in\Delta^{m-1}$ such that $\mathcal{B}^*(p)\cap\mathcal{B}^*(q)=\emptyset$ and $I_B(p,q)=0$ for every $B\in\mathcal{B}$.*

The reduction proceeds in three moves, developed in full in Section E of the Electronic Companion. First, an alternative changes the optimal action exactly when a single *challenger* $B'\notin\mathcal{B}^*(p)$

[1] Otherwise every feasible action is optimal under p, no alternative can have an optimal set disjoint from $\mathcal{B}^*(p)$, and $\mathcal{Q}(p)=\emptyset$ with $\kappa^*(p)=0$.

strictly beats every p-optimal action under q, so $\mathcal{Q}(p)$ splits into at most $|\mathcal{B} \setminus \mathcal{B}^*(p)|$ convex *challenger cells* $\mathcal{Q}_{B'}(p)$, each cut out by finitely many affine constraints; under Assumption 1 these cells cover $\mathcal{Q}(p)$ exactly, and the semi-infinite constraint of (1) reduces to one worst-case constraint per nonempty cell, indexed by $\mathcal{C}(p) := \{B' \in \mathcal{B} \setminus \mathcal{B}^*(p) : \mathcal{Q}_{B'}(p) \neq \emptyset\}$. Second, replacing each cell by its closure leaves that worst-case value unchanged and turns the inner infimum into a minimum over a compact set;[2] the resulting constraints are concave in the exploration weights x, hence convex. Third, convex duality replaces each inner minimization by the existence of a dual certificate obeying finitely many linear inequalities, provided the inner problem is regular. The relevant constraint qualification, stated as Assumption EC.1 in Section E of the Electronic Companion, requires a relative-interior point of the p-optimal observation equalities that strictly satisfies the challenger inequalities and lies in every divergence term's effective domain. The same section gives a cell-detection linear program that produces candidate points, with relative-interior membership checked separately. Strong duality then yields an attained certificate for every cell and every x.

The only nonlinearity left in a certificate is a perspective term, which the exponential cone $\mathcal{K}_{\exp} := \mathrm{cl}\{(z_1, z_2, z_3) : z_2 > 0,\ z_2 \exp(z_1/z_2) \leq z_3\}$ represents exactly through auxiliary variables. Assembling the three moves yields the reformulation.

THEOREM 3 **(Finite Exponential-Cone Reformulation)**. *Under Assumption 1 and the relative Slater condition (Assumption EC.1), $\kappa^*(p)$ is the optimal value of the finite exponential-cone program*

$$
\begin{aligned}
\min \quad & \sum_{B \notin \mathcal{B}^*(p)} x_B \Delta_B(p) && (3)\\
s.t. \quad & \sum_{B \notin \mathcal{B}^*(p)} \sum_{\substack{S \in \mathcal{S}_B: \\ P_{B,S}(p) > 0}} P_{B,S}(p)(u_{B',B,S} + x_B) + \sum_{B^* \in \mathcal{B}^*(p)} \sum_{S \in \mathcal{S}_{B^*}} P_{B^*,S}(p) \eta_{B',B^*,S} \geq 1, \quad \forall B' \in \mathcal{C}(p),\\
& \sum_{B \notin \mathcal{B}^*(p)} \lambda_{B',B,S(B,c_k)} + \sum_{B^* \in \mathcal{B}^*(p)} \Big(\eta_{B',B^*,S(B^*,c_k)} + \nu_{B',B^*} (r_{B'} - r_{B^*})_k \Big) \leq 0, \quad \forall k \in [m],\ \forall B' \in \mathcal{C}(p),\\
& (u_{B',B,S}, x_B, \lambda_{B',B,S}) \in \mathcal{K}_{\exp}, \quad \forall B' \in \mathcal{C}(p),\ \forall B \notin \mathcal{B}^*(p),\ \forall S \in \mathcal{S}_B \text{ with } P_{B,S}(p) > 0,\\
& x_B \geq 0, \quad \lambda_{B',B,S} \geq 0, \quad \nu_{B',B^*} \geq 0, \quad \eta_{B',B^*,S} \in \mathbb{R},
\end{aligned}
$$

where the last line ranges over all $B' \in \mathcal{C}(p)$, $B \notin \mathcal{B}^*(p)$, $B^* \in \mathcal{B}^*(p)$, and over $S \in \mathcal{S}_B$ in the λ variables and $S \in \mathcal{S}_{B^*}$ in the η variables.

The program is convex: the objective and the dual-feasibility constraints are linear, and each exponential-cone constraint is convex. Its size, however, is governed by the combinatorial action

[2] The two-path instance of Example EC.1 shows that this step is not vacuous: its least informative costly confounding alternative is approached only at the boundary of the cell.

space—one exploration variable per suboptimal action and one challenger block per cell—and $|\mathcal{B}| = \sum_{j=0}^{K} \binom{|A|}{j}$ grows exponentially in the budget K, so the representation is exact and finite for a fixed network but is not by itself a scalable algorithm; the cell-detection linear program prunes the candidate challengers to $\mathcal{C}(p)$, which can be much smaller. Finally, $\kappa^*(p)$ does not enter the guarantees of Section 6, whose policies attain the logarithmic rate but not the optimal constant, because the multiplier β of Theorem 2 need only exceed a conservative threshold; the optimal allocation x^* supplies the value of β for which the achieved constant equals $\kappa^*(p)$, as Section 8.3 shows. A policy that made this choice from data would have to solve (1) repeatedly at a running estimate of p, which is practical for a finite convex program and not for a semi-infinite one.

8. Numerical Experiments

Table 3 specifies five experiments. E1–E3 are mechanism studies: restricted, purpose-built action classes and long horizons isolate a single effect. E4 tests the policies at realistic scale under the full all-arc class $\mathcal{B} = \{B \subseteq A : |B| \leq K\}$, on generated random graphs and on the Arizona–Mexico infiltration network of Unsal (2010). E5 returns to that topology under a checkpoint-closure design in which diagnostic information is dominated.

8.1. Experimental Setup

Policies. We compare *Full-feedback TS*, a Thompson-sampling benchmark that observes the latent scenario after each play; *Path-only TS*, which samples from the posterior induced by path-only observations; *UCB*, the path-only Value-UCB rule of Algorithm 3 in Section F.1 of the Electronic Companion; and *IE-Greedy*, the identifying-exploration greedy policy of Section 6.4. Both TS policies use the symmetric prior $\alpha^0 = (1, \ldots, 1)$. The all-arc benchmarks add *IE-Greedy Relaxed*, an estimator variant specified in Section F.2 of the Electronic Companion; it tracks IE-Greedy throughout, so we do not discuss it separately. Scenario streams are shared across policies within each replication, and boldface marks the minimum computed before rounding, so a displayed tie may contain a single bold entry.

Performance measure. All curves report mean cumulative pseudo-regret, $\widehat{\mathcal{R}}(T) = n_{\text{rep}}^{-1} \sum_{j=1}^{n_{\text{rep}}} \sum_{t=1}^{T} \Delta_{B^{t,(j)}}(p)$, where $B^{t,(j)}$ is the action chosen in period t of replication j. Benchmark tables report pseudo-regret at the terminal horizon and mean wall-clock time in seconds.

Implementation. All algorithms were coded in Julia (Bezanson et al. 2017), using Gurobi for the all-arc graph oracle and the Value-UCB linear programs and MOSEK for the lower-bound, conic calibration, and identifying-set routines. Section F.3 of the Electronic Companion lists package versions, hardware, and the timing protocol.[3]

[3] Reported times are descriptive rather than controlled microbenchmarks.

Table 3 Design of the five experiments. $\mathcal{M}$ is the model class used by Value-UCB and IE-Greedy, m the number of latent scenarios, and n_{rep} the replications. Unless a caption says otherwise, every figure and table uses these settings.

	Network	Action class $\mathcal{B}$	Model class $\mathcal{M}$	m	T	n_{rep}	β
E1 (§8.2)	Three parallel paths	$\{B_0, B_1, B_2, B_3\}$	$\{(\frac{1}{2}, \frac{1}{2}-\eta, \eta) : \eta \in [0, \frac{1}{2}]\}$	3	50,000	300	20
E2 (§8.3)	Three parallel paths	$\{B_0, B_1, B_2, B_3\}$	$\{q : q_3 = q_4\}$	4	50,000	300	37.079*
E3 (§8.4)	2×30 layered DAGs	purpose-built	full simplex	—	20,000	30/network	20
E4 (§8.5)	ER(20, 0.5)	$\lvert B\rvert \le K,\ K \in \{1,2,3,4\}$	full simplex	10	1,000	30×30	20
	Arizona–Mexico	$\lvert B\rvert \le K,\ K \in \{1,2,3,4\}$	full simplex	3	5,000	30×30	20
E5 (§8.6)	Arizona–Mexico	12 entrances, $\lvert B\rvert \le 11$	full simplex Δ^2	3	50,000	30×13	230

*Oracle multiplier $x^*_{B_1}$ from the lower-bound solution, used for calibration only.

8.2. Dominated Information on the Three-Path Instance

Design. We use the three-path instance of Section 4.2 with $\varepsilon = 0.1$, so that $p^\varepsilon = (1/2, 1/2 - \varepsilon, \varepsilon)$, action B_1 is uniquely optimal, and its value gap against B_2 is 2.7. Neither B_1 nor B_2 separates p^ε from the confounding alternative $q^\varepsilon = (1/2, \varepsilon, 1/2 - \varepsilon)$; only the dominated actions B_3 and B_0 do, and we take B_3 as the diagnostic action. On this model class the Value-UCB initial index vector over (B_1, B_2, B_3, B_0) is $(6, 6, 3/2, 3/2)$, and the B_1–B_2 tie is broken uniformly at random.

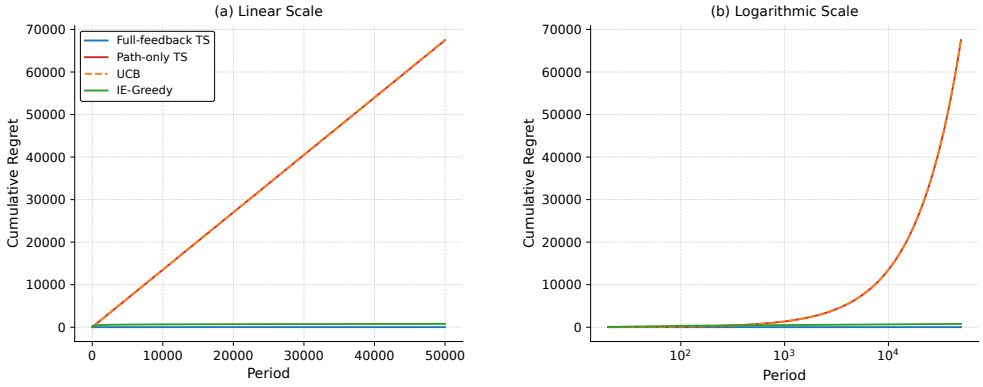

Figure 2 E1: mean cumulative pseudo-regret on the three-path instance, on linear (left) and logarithmic (right) period scales.

Results. In Figure 2, Path-only TS and UCB are nearly indistinguishable and linear, at 1.3497 and 1.3505 terminal pseudo-regret per period, exactly as Proposition 5 and the absorption argument of Proposition 7 predict for this specification. Full-feedback TS identifies the optimal action early and ends at 14.4; IE-Greedy forces the diagnostic action B_3 and ends at 781.2. Identifying exploration therefore avoids linear regret, but its finite-horizon cost depends on a multiplier β that is not obvious ex ante: Section F.4 of the Electronic Companion varies β on this instance and finds the best final pseudo-regret at $\beta = 5$, with both smaller and larger multipliers substantially worse.

8.3. Calibration against the Lower Bound

Design. The three-path instance has an infinite lower-bound constant over the full simplex, so a finite calibration requires the four-scenario cost support of Example EC.2 on the same topology and action set. Solving (1) gives $\kappa^*_{\mathcal{M}}(p) = 15.017$ and $x^*_{B_1} = 37.079$. For this calibration only, IE-Greedy is given the oracle multiplier $\beta = x^*_{B_1}$ and the identifying set $\{B_1, B_3\}$; since B_3 is optimal and $\Delta_{B_1}(p) = 0.405$, its forced-exploration cost is calibrated to $\kappa^*_{\mathcal{M}}(p) \log T$.

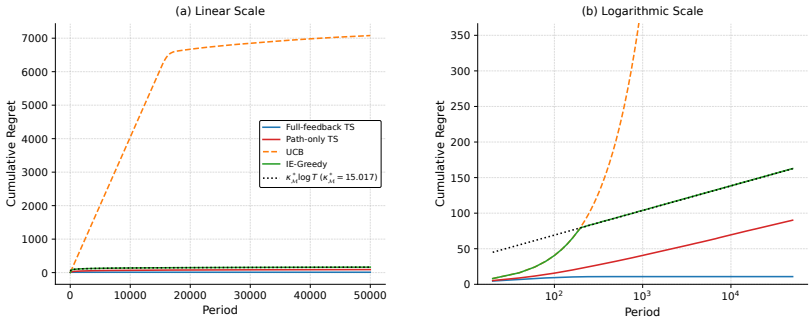

Figure 3 E2: mean cumulative pseudo-regret on the four-scenario calibration instance. Left, full regret scale; right, logarithmic period axis on a lower-bound-scale vertical range.

Results. In Figure 3, oracle-calibrated IE-Greedy tracks $\kappa^*_{\mathcal{M}}(p) \log T$ almost exactly: at the final horizon $\widehat{\mathcal{R}}(T)/(\kappa^*_{\mathcal{M}}(p) \log T) = 1.0020$, with fitted logarithmic coefficient 15.040. In other words, IE-Greedy achieves the best possible asymptotic performance on this instance. We note, however, that this calibrates the lower-bound allocation rather than establishing adaptive achievability, since the β multiplier uses the true solution. Path-only TS has lower finite-horizon regret here, ending at 90.13, because the model-class restriction makes path observations partially informative; UCB is far more conservative and ends at 7,078.9, visible only on the full-scale panel.

8.4. Network Families

Design. We next test whether these mechanisms persist beyond hand-built examples. The instances in the dominated-information family are built so that the value-maximizing actions do not reveal the confounding scenario, while a strictly suboptimal action does. The self-revealing family instead filters random layered directed acyclic graphs until every feasible action reveals the latent scenario through the observed path. For each network, IE-Greedy's identifying set is selected by MILP (EC.11) (Section D of the Electronic Companion) and verified by Lemma 1.

Results. In the dominated-information family, Path-only TS and UCB remain linear, ending near 27,000, while IE-Greedy ends at 716.5 and Full-feedback TS at 16.3. The issue of dominated information is therefore not exclusive to the three-path example. In the self-revealing family, Full-feedback TS and Path-only TS remain close, at 12.81 and 12.75; UCB learns more conservatively

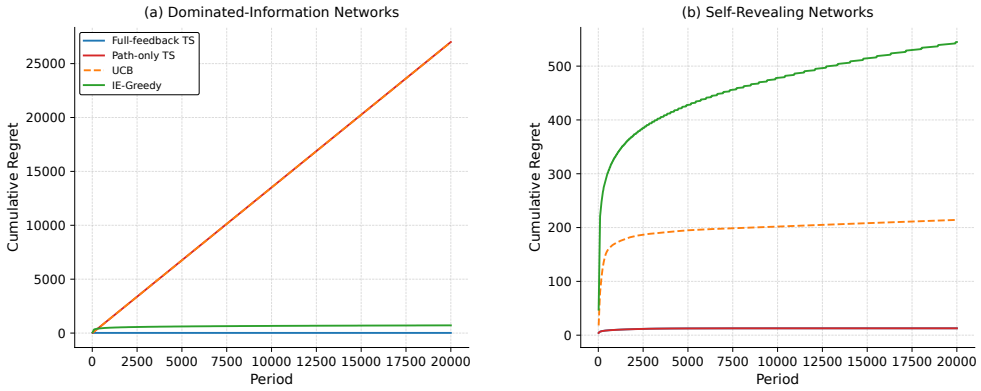

Figure 4 E3: mean cumulative pseudo-regret across generated layered networks, averaged over replications and then over networks. Left, dominated-information family; right, self-revealing family.

and ends at 214.2, still below the forced-exploration cost of IE-Greedy, which ends at 544.8. The contrast between the two panels reveals an important tradeoff centered on dominated information: when there is none, as in Figure 4(b), TS and UCB outperform IE-Greedy; however, these policies cannot deal with settings where information is dominated, as shown in Figure 4(a).

8.5. All-Arc Benchmarks at Scale

Design. We use two classes of generated graphs and one real network. The generated classes are layered graphs in the sense of Ryzhov and Powell (2011), reported in Section F.7 of the Electronic Companion, and directed Erdős–Rényi random graphs (Erdős and Rényi 1959) with a source–sink backbone added to guarantee feasibility. Each arc receives a baseline cost drawn uniformly from $[1, 10]$, rescaled per scenario by an independent factor drawn uniformly from $[0.35, 2.20]$; the true scenario law is drawn from a symmetric Dirichlet distribution, and we sample 30 accepted nontrivial instances per row, discarding candidates that fail feasibility or identification checks or whose scenario-dependent oracle actions do not create a nontrivial learning problem. The policies know the scenario support but not the true law or the realized scenario stream.

The real network is the Arizona–Mexico infiltration network of Unsal (2010) in Figure 5, an instance of the illicit-network interdiction problems that motivate this line of work (Smith and Song 2020). It models unauthorized crossings between ports of entry along a stretch of the U.S.–Mexico border: 38 path-relevant nodes and 109 directed arcs run from a Mexico-side entry point to a Phoenix destination, so the interdictor allocates limited detection resources against an evader with superior on-the-ground knowledge of realized crossing conditions. We convert the three arc-level detection-probability vectors in the appendix of Unsal (2010) into a latent cost support through $c_{k,a} = -\log(1 - \varphi_{k,a})$.[4]

[4] The original description lists 39 vertices, but one is isolated and cannot lie on a source–sink path. Interpreting these columns as latent cost scenarios is our modeling choice rather than a feature of the original study.

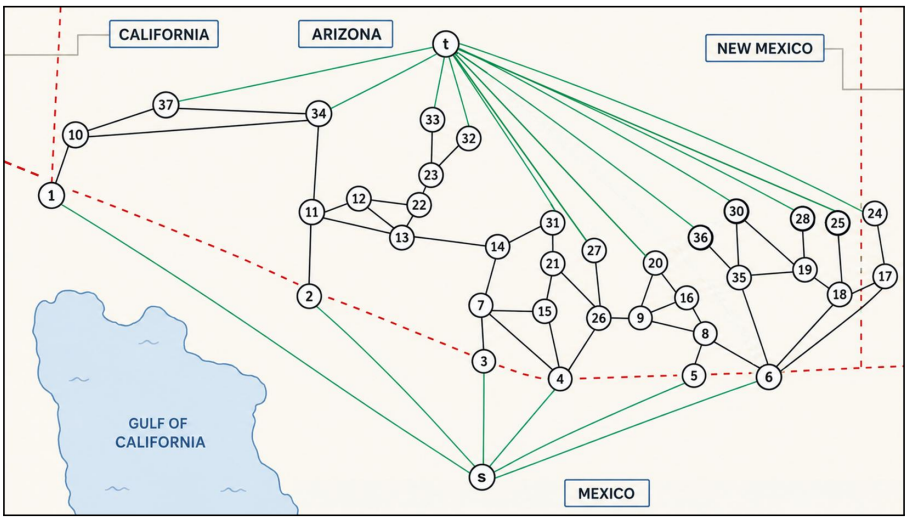

Figure 5 Infiltration network near the Arizona–Mexico border, extracted from Unsal (2010). Dashed lines are state and national borders; green edges have zero detection probability, black edges positive.

Protocol. Unlike the mechanism experiments, these runs use the full interdiction class over all interdictable arcs: a layered graph with $L = 3$ intermediate layers and width $w = 5$ has 60 interdictable arcs, so its $K = 3$ class contains 36,051 feasible sets. All rows enumerate the action class completely and precompute the exact payoff and observation table for every action–scenario pair.[5] In Table 4, regret cells report the mean final cumulative pseudo-regret on the first line and (SE; max) on the second, computed across accepted instances after averaging over replications within each instance; the action column reports mean $|\mathcal{B}|$ and Rej. the sampled candidates rejected before acceptance. On the Arizona–Mexico network the topology and the three source-derived scenario vectors are fixed while the scenario law varies, so Table 5 instead averages 30 streams within each of 30 independently drawn laws and reports standard errors across the 30 law-level means.

Table 4 E4: all-arc random graph benchmark. Rows vary the budget K; columns report terminal pseudo-regret and policy running time.

Nodes	Arc prob.	K	mean $\lvert\mathcal{B}\rvert$	Rej.	Mean regret (SE; max)					Policy time (s)				
					Full-feedback TS	Path-only TS	UCB	IE-Greedy	IE-Greedy Relaxed	Full-feedback TS	Path-only TS	UCB	IE-Greedy	IE-Greedy Relaxed
20	0.5	1	108	123	**13.5** (2.3; 42.4)	37.5 (9.5; 193.2)	113.8 (26.0; 455.8)	287.1 (37.9; 770.6)	287.7 (37.9; 770.6)	**0.0088**	1.489	0.1154	1.873	0.0134
20	0.5	2	5.7k	11	**13.8** (2.4; 45.8)	40.0 (8.3; 201.0)	577.7 (133.7; 2,572)	484.9 (116.6; 3,560)	484.9 (116.6; 3,560)	0.1752	2.626	0.5507	2.036	**0.1662**
20	0.5	3	197.2k	4	**19.3** (2.3; 53.6)	65.7 (14.4; 398.4)	674.3 (147.9; 3,600)	937.8 (98.5; 2,487)	938.3 (98.5; 2,487)	**5.797**	8.078	7.098	7.818	5.843
20	0.5	4	5.08M	13	**30.2** (3.7; 64.7)	105.0 (14.9; 290.9)	863.8 (204.6; 4,890)	1,495 (132.5; 3,495)	1,497 (132.4; 3,495)	**194.5**	194.9	201.1	197.9	198.6

Results. Three patterns emerge. First, Full-feedback TS is the best informational reference in these runs, since it observes the realized scenario. Second, Path-only TS stays close to it in many all-arc graphs: in random graphs with $K = 4$ it ends at 105.0 versus 30.2, on the Arizona–Mexico network the two remain close at all four budgets, and Section F.7 of the Electronic Companion

[5] The same code uses a deterministic most-vital-arcs oracle when enumeration is unavailable, but that route is not used in the benchmarks reported here.

Table 5 E4: all-arc benchmark on the Arizona–Mexico network. Rows vary the budget K; columns follow Table 4.

K	$\lvert\mathcal{B}\rvert$	Mean regret (SE)					Policy time (s)				
		Full-feedback TS	Path-only TS	UCB	IE-Greedy	IE-Greedy Relaxed	Full-feedback TS	Path-only TS	UCB	IE-Greedy	IE-Greedy Relaxed
1	110	0.37 (0.090)	**0.37** (0.089)	2.29 (0.48)	3.43 (0.60)	3.43 (0.60)	0.049	1.11	0.293	0.118	**0.046**
2	6.00k	**0.20** (0.045)	0.51 (0.15)	1.66 (0.69)	12.6 (0.52)	12.6 (0.52)	**0.118**	0.723	0.699	0.159	0.133
3	215.9k	**0.67** (0.17)	1.14 (0.18)	19.0 (6.24)	26.2 (0.83)	26.2 (0.83)	**3.58**	4.45	4.77	3.64	3.59
4	5.78M	**0.26** (0.076)	0.59 (0.20)	5.21 (2.89)	41.6 (2.33)	41.6 (2.33)	157.6	**152.8**	158.5	153.8	153.7

reproduces the pattern on layered graphs and Section F.6 plots the corresponding time paths. Third, explicit identifying exploration is expensive at these short horizons, where IE-Greedy's forced diagnostic samples often dominate its regret. UCB is the most instance dependent: it remains with one optimistic action for long stretches, which is consistent with the nearly deterministic regret paths and small standard errors of several UCB rows, and on the Arizona–Mexico network it incurs zero pseudo-regret for 25 of the 30 laws at $K = 2$ and 26 at $K = 4$, while a heavier nonzero-law tail at $K = 3$ raises its mean regret to 19.0.

Interpretation. Read the running times together with the action-space column. In the random $K = 4$ block the mean action space exceeds five million interdiction sets, yet every simulation finishes in roughly three minutes per instance — not because the action class is restricted, but because the finite-support representation permits exact reuse. Once the induced path and payoff are tabulated for each action–scenario pair, many interdiction sets share the same payoff vector across the 10 scenarios, and UCB solves one confidence problem per distinct vector while retaining all feasible actions and the same tie-breaking rule. Its running time is therefore governed by the number of distinct payoff and observation classes rather than the raw action count.

8.6. Checkpoint Closures Induce Dominated Information

The all-arc benchmark uses small budgets over all arcs and leaves many routing alternatives open. We now ask whether the same Arizona–Mexico topology can support dominated information when the action class lets the interdictor pin the evader near the destination. This experiment replaces the scenario construction and the action class, and its results are not pooled with those of E4.

Design. Let $c^{\text{base}}(a) = -\log(1 - \varphi_{1,a})$. Scenario 1 uses c^{base}; Scenario 2 multiplies the costs of arcs $(17, 24)$ and $(35, 36)$ by 15 and 0.2, respectively, and Scenario 3 swaps these two factors.[6] The feasible actions are all subsets of the 12 arcs entering the destination with cardinality at most 11, giving exactly 4,095 actions and excluding total closure; the optimal pinning actions close 11 of the 12 entrances, so this experiment represents a near-total checkpoint closure rather than a low-budget interdiction. The law grid $p^{\varepsilon} = (1/2, 1/2 - \varepsilon, \varepsilon)$, $\varepsilon \in \{0.01, \ldots, 0.13\}$, was fixed before any policy was

[6] These perturbations are designed rather than measured.

simulated and contains all positive multiples of 0.01 within the structurally certified regime. For each grid point the swapped law $q^\varepsilon = (1/2, \varepsilon, 1/2 - \varepsilon)$ is confounding: the p^ε- and q^ε-optimal actions are disjoint and uninformative for separating the pair, while every informative action is strictly suboptimal under both laws. IE-Greedy's exact selector returns a common rank-two set of two actions, and $\beta = 230$ is the smallest multiple of 10 strictly above the largest IE-Greedy threshold of Theorem 2 on the p-grid.[7]

Exact Value-UCB prediction. Let $B^\dagger$ denote the closure leaving only the entrance from node 24 open. At period 1 the confidence set is Δ^2, so the initial index of B is $\max_{k\in[3]} r_{B,k}$, and across all 4,095 actions $B^\dagger$ uniquely maximizes it, at 4.887 against 1.103 for the runner-up. Under every constructed scenario $B^\dagger$ induces the same path through nodes 6, 17, and 24, so it is scenario-blind and Proposition 7 applies unchanged to the numerical specification: Value-UCB selects $B^\dagger$ in every period. That closure is optimal under every p^ε on the grid, whereas under q^ε the closure leaving only the entrance from node 32 open is optimal and $\Delta_{B^\dagger}(q^\varepsilon) > 0$. Hence $\mathcal{R}^{\mathrm{UCB}}(T, p^\varepsilon) = 0$ and $\mathcal{R}^{\mathrm{UCB}}(T, q^\varepsilon) = T\Delta_{B^\dagger}(q^\varepsilon) > 0$: Value-UCB is absorbed from the first decision.

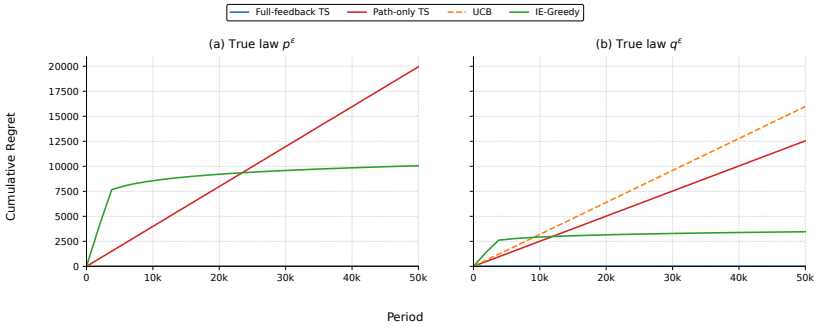

Figure 6 E5: mean cumulative pseudo-regret in the checkpoint-closure experiment, averaged over streams and then over the 13-law grid, under the p^ε laws (left) and the q^ε laws (right).

Results. Figure 6 and Table 6 report both law families. Under p^ε, Path-only TS remains linear, reaching 19,943.8 at $T = 50{,}000$. UCB remains at zero up to floating-point tolerance because it repeats $B^\dagger$ without identifying the latent law, while IE-Greedy pays a large front-loaded exploration cost and then flattens, rising only from 9,204.2 at $T = 20{,}000$ to 10,056.7 at $T = 50{,}000$, by which point it is below Path-only TS. The paired q-side runs expose the selection effect. UCB repeats the same scenario-blind closure, now suboptimal, and grows linearly to 15,976.1, exceeding the 12,543.9 of Path-only TS; the apparent UCB success under p thus reverses when q is the data-generating

[7] Two elements are post hoc. After observing that UCB incurred zero regret under p^ε, we added a paired q-side run: for each law and replication the q stream swaps scenario labels 2 and 3 in the corresponding p stream, with seeds, implementations, and $\beta = 230$ transferred without recalibration. Since the transferred multiplier fails the q-side threshold at all 13 grid points, the q block is a fixed-parameter comparison rather than a theorem-certified one, and we read it as a mechanism diagnostic rather than an independent validation. We also later extended both runs to $T = 50{,}000$, reusing the original seeds and reproducing the archived trajectories through $T = 20{,}000$ exactly.

Table 6 E5: checkpoint-closure benchmark. Left and right blocks report regret (SE) under the paired p^ε and q^ε laws at $T \in \{5{,}000, 20{,}000, 50{,}000\}$; parentheses give the Monte Carlo standard error conditional on the law grid.

	True law p^ε: Regret (SE)			True law q^ε: Regret (SE)		
Policy	$T = 5{,}000$	$T = 20{,}000$	$T = 50{,}000$	$T = 5{,}000$	$T = 20{,}000$	$T = 50{,}000$
Full-feedback TS	2.33 (0.14)	2.33 (0.14)	2.33 (0.14)	**12.6** (0.63)	**15.4** (1.19)	**15.7** (1.25)
Path-only TS	1,994.6 (2.43)	7,977.3 (5.10)	19,943.8 (7.94)	1,254.2 (0.52)	5,018.4 (1.11)	12,543.9 (1.75)
UCB	**0.00** (0.00)	**0.00** (0.00)	**0.00** (0.00)	1,597.6 (0.00)	6,390.4 (0.00)	15,976.1 (0.00)
IE-Greedy	7,915.3 (0.00)	9,204.2 (0.00)	10,056.7 (0.00)	2,704.2 (0.18)	3,151.4 (2.75)	3,458.6 (8.04)

law, although the policy and its parameters are unchanged. Full-feedback TS plateaus in both families, and IE-Greedy reaches 3,458.6, below both path-only policies at the extended horizon.

Interpretation. The two Arizona–Mexico experiments isolate the role of the action class on a common topology. Under low-budget all-arc interdiction, valuable actions provide enough feedback for Path-only TS to stay close to full feedback. When checkpoint closures let the interdictor pin the evader, informative actions become value-dominated and posterior sampling can remain trapped. Dominated information is therefore compatible with this geometry under a near-total closure design; we do not claim it is present in measured border operations.

9. Concluding Remarks

Path-only feedback turns stochastic shortest-path interdiction from a combinatorial bandit into a partial-monitoring problem. Because the evader's route aggregates cost scenarios into a single observation, distinct environments can be observationally equivalent under some actions while the actions separating them are strictly value-dominated. This *dominated information* makes pure posterior sampling and value-optimistic rules incur linear regret on the constructed instances, whereas full feedback removes the obstruction and restores the gap-free square-root minimax rate.

The instance-dependent Lower Bound Problem gives the exploration frequencies any consistent policy must sustain, and for the full-simplex model class it reduces to an exact finite exponential-cone program. Making the regret-optimal allocation solvable by off-the-shelf conic solvers speaks to an oft-noted bottleneck for adaptive defense strategies: their computational practicality at scale (Smith and Song 2020). The Identifying-Exploration policy forces logarithmic sampling of an identifying set to restore the right regret order on identifiable instances with a positive gap. Its conservatism is that it learns the entire latent distribution rather than only the relevant confounding alternatives, so a gap between our necessary and sufficient conditions remains.

For an interdictor who observes only routes, blocking whatever currently looks most damaging can fail to learn at all, because the blockades that reveal how the adversary responds are the ones that look least attractive. Whether this matters is a property of the network, not of the policy,

so the recommendation is to budget probes known to be locally wasteful. The analysis also prices surveillance: observing realized costs converts the problem into a standard learning problem, and the gap between the two regimes is the quantity against which additional sensing should be weighed. Two directions seem immediate: a fully adaptive policy attaining $\kappa^*_{\mathcal{M}}(p)$ without identifying the entire latent distribution, and a strategic evader who routes so as to conceal the latent costs.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K, eds., *Advances in Neural Information Processing Systems*, volume 24, 2312–2320 (Curran Associates, Inc.).

Agrawal S, Goyal N (2013) Further optimal regret bounds for Thompson sampling. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, 99–107 (PMLR).

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2-3):235–256.

Azizi E, Seifi A (2024) Shortest path network interdiction with incomplete information: a robust optimization approach. *Annals of Operations Research* 355(2):727–759.

Ball M, Golden B, Vohra R (1989) Finding the most vital arcs in a network. *Operations Research Letters* 8(2):73–76.

Bartók G, Foster DP, Pál D, Rakhlin A, Szepesvári C (2014) Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research* 39(4):967–997.

Bayrak H, Bailey MD (2008) Shortest path network interdiction with asymmetric information. *Networks* 52(3):133–140.

Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: A fresh approach to numerical computing. *SIAM review* 59(1):65–98.

Borrero JS, Prokopyev OA, Sauré D (2016) Sequential shortest path interdiction with incomplete information. *Decision Analysis* 13(1):68–98.

Borrero JS, Prokopyev OA, Sauré D (2019) Sequential interdiction with incomplete information and learning. *Operations Research* 67(1):72–89.

Borrero JS, Prokopyev OA, Sauré D (2022) Learning in sequential bilevel linear programming. *INFORMS Journal on Optimization* 4(2):174–199.

Borrero JS, Sauré D, Trigo N (2025) Optimal sequential stochastic shortest path interdiction. *European Journal of Operational Research* 326(3):641–655, URL `http://dx.doi.org/10.1016/j.ejor.2025.04.009`.

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5(1):1–122.

Cesa-Bianchi N, Lugosi G (2012) Combinatorial bandits. *Journal of Computer and System Sciences* 78(5):1404–1422.

Cormican K, Morton D, Wood R (1998) Stochastic network interdiction. *Operations Research* 46(2):184–197.

Cover TM, Thomas JA (2006) *Elements of Information Theory* (Hoboken, NJ: Wiley), 2nd edition.

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. Servedio RA, Zhang T, eds., *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, 355–366 (Omnipress).

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1):1–22.

Erdős P, Rényi A (1959) On random graphs I. *Publicationes Mathematicae Debrecen* 6:290–297.

Fulkerson D, Harding G (1977) Maximizing the minimum source-sink path subject to a budget constraint. *Mathematical Programming* 13(1):116–118.

Gai Y, Krishnamachari B, Jain R (2012) Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking* 20(5):1466–1478.

Garivier A, Ménard P, Stoltz G (2019) Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research* 44(2):377–399, URL `http://dx.doi.org/10.1287/moor.2017.0928`.

Geyer CJ (1992) Practical Markov chain Monte Carlo. *Statistical Science* 7(4):473–483.

Ghare P, Montgomery D, Turner W (1971) Optimal interdiction policy for a flow network. *Naval Research Logistics Quarterly* 18(1):37–45.

Gift PD (2010) *Planning for an adaptive evader with application to drug interdiction operations*. Master's thesis, Naval Postgraduate School, Monterey, California.

Graves TL, Lai TL (1997) Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM Journal on Control and Optimization* 35(3):715–743.

Hemmecke R, Schultz R, Woodruff DL (2003) Interdicting stochastic networks with binary interdiction effort. Woodruff DL, ed., *Network Interdiction and Stochastic Integer Programming*, 69–84 (Boston, MA: Springer US).

Israeli E, Wood R (2002) Shortest-path network interdiction. *Networks* 40(2):97–111.

Kang S, Bansal M (2023) Distributionally risk-receptive and risk-averse network interdiction problems with general ambiguity set. *Networks* 81(1):3–22.

Ketkov SS, Prokopyev OA (2020) On greedy and strategic evaders in sequential interdiction settings with incomplete information. *Omega* 92:102161.

Kosmas D, Sharkey TC, Mitchell JE, Maass KL, Martin L (2023) Interdicting restructuring networks with applications in illicit trafficking. *European Journal of Operational Research* 308(2):832–851.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4–22.

Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press).

Malaviya A, Rainwater C, Sharkey T (2012) Multi-period network interdiction problems with applications to city-level drug enforcement. *IIE Transactions* 44(5):368–380.

Malik K, Mittal A, Gupta S (1989) The k-most vital arcs in the shortest path problem. *Operations Research Letters* 8(4):223–227.

McMasters A, Mustin T (1970) Optimal interdiction of a supply network. *Naval Research Logistics Quarterly* 17(3):261–268.

Modaresi S, Sauré D, Vielma JP (2020) Learning in combinatorial optimization: What and how to explore. *Operations Research* 68(5):1585–1604.

Morton DP (2010) Stochastic network interdiction. Cochran JJ, Cox LA, Keskinocak P, Kharoufeh JP, Smith JC, eds., *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Inc.).

Nguyen DH, Smith JC (2022) Network interdiction with asymmetric cost uncertainty. *European Journal of Operational Research* 297(1):239–251.

Nguyen DH, Smith JC (2024) Asymmetric stochastic shortest-path interdiction under conditional value-at-risk. *IISE Transactions* 56(4):398–410.

Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

Rockafellar RT (1970) *Convex Analysis* (Princeton University Press).

Rusmevichientong P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Mathematics of Operations Research* 35(2):395–411.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4):1221–1243.

Ryzhov I, Powell W (2011) Information collection on a graph. *Operations Research* 59(1):188–201.

Sefair JA, Smith JC (2016) Dynamic shortest-path interdiction. *Networks* 68(4):315–330.

Smith JC, Song Y (2020) A survey of network interdiction models and algorithms. *European Journal of Operational Research* 283(3):797–811.

Soleimani-Alyar M, Ghaffari-Hadigheh A (2017) Solving multi-period interdiction via generalized Benders decomposition. *Acta Mathematicae Applicatae Sinica, English Series* 33:633–644.

Steinrauf R (1991) *Network Interdiction Models*. Master's thesis, Naval Postgraduate School, Monterey, California.

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Unsal O (2010) *Two-person zero-sum network-interdiction game with multiple inspector types*. Master's thesis, Naval Postgraduate School.

Wood RK (1993) Deterministic network interdiction. *Mathematical and Computer Modelling* 17(2):1–18.

Yang J, Borrero JS, Prokopyev OA, Sauré D (2021) Sequential shortest path interdiction with incomplete information and limited feedback. *Decision Analysis* 18(3):218–244.

Electronic Companion: Proofs and Computational Details

Appendix A: Proofs for the Path-Feedback Information Lower Bound

This appendix establishes the observational-equivalence obstruction, the expected path-feedback information lower bound, and the resulting regret bound. Throughout we use the notation of Sections 3 and 4.

One construction deserves emphasis, since the change-of-measure argument rests on it. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ fixed in the Introduction carries two independent i.i.d. uniform sequences: one generates the cost scenarios under a specified latent distribution, the other implements the policy's internal randomization. Feeding these sequences to the kernels π_t and to the scenario probabilities p induces a law $\mathbb{P}_p^\pi$ on the canonical space of actions, scenarios and responses, with expectation $\mathbb{E}_p^\pi$. Thus $\mathbb{P}_p^\pi$ and $\mathbb{P}_q^\pi$ are distinct induced measures on the *same* measurable space, not two names for $\mathbb{P}$. Restricted to a σ-field $\mathcal{G}$ we write $D_{\mathcal{G}}(\mathbb{P}_p^\pi \| \mathbb{P}_q^\pi)$ and $\mathbb{P}_p^\pi \ll \mathbb{P}_q^\pi$ on $\mathcal{G}$; for a realization h^T of H^T we abbreviate $\mathbb{P}_p^\pi(h^T) := \mathbb{P}_p^\pi(H^T = h^T)$.

Because the cost vectors are i.i.d. and exogenous and the two primitive sequences are independent, c^t is independent of the past and of the randomization selecting B^t; conditional on (H^{t-1}, B^t), the observed path S^t has law $P_{B^t}(p)$ under $\mathbb{P}_p^\pi$. Every history factorization below uses this.

Proof of Theorem 1. The theorem asserts that the information accumulated against q must grow at least like $\log T$. The proof reaches this in two steps that meet at the same quantity, the divergence between the two experiments: Step 1 shows it equals the play-count-weighted sum of information numbers, and Step 2 bounds it below by $(1 - o(1)) \log T$ using consistency alone. Reading the two together gives the bound.

Step 1 (KL Divergence under Adaptive Sampling). Under path-only feedback the experiment is a finite-armed bandit whose arms are the interdiction actions, and the two environments differ only through the path law $P_B(\cdot)$. The identity below is therefore the divergence decomposition for adaptively sampled bandits (Lattimore and Szepesvári 2020, Lemma 15.1), with $I_B(p, q)$ as the per-play divergence of arm B; we derive it because the derivation is what ties it to path-only feedback. Let $\pi(B \mid h)$ denote the probability that π selects B after history h, and for a realization h^T of H^T let h^{t-1} be its prefix through period $t-1$. The two history laws factor as

$$\mathbb{P}_p^\pi(h^T) = \prod_{t=1}^{T} \pi(B^t \mid h^{t-1}) P_{B^t, S^t}(p), \qquad \mathbb{P}_q^\pi(h^T) = \prod_{t=1}^{T} \pi(B^t \mid h^{t-1}) P_{B^t, S^t}(q).$$

The common policy factors and $P_B(p) \ll P_B(q)$ make $\mathbb{P}_p^\pi(h^T) > 0$ imply $\mathbb{P}_q^\pi(h^T) > 0$ for every h^T; hence $\mathbb{P}_p^\pi \ll \mathbb{P}_q^\pi$ on $\mathcal{F}_T = \sigma(H^T)$. Finiteness of $\mathcal{S}_B$ then also gives $I_B(p,q) < +\infty$ for every B. Averaging the log-likelihood ratio under p then gives

$$\begin{aligned} D_{\mathcal{F}_T}(\mathbb{P}_p^\pi \| \mathbb{P}_q^\pi) &\overset{(a)}{=} \sum_{h^T : \mathbb{P}_p^\pi(h^T)>0} \mathbb{P}_p^\pi(h^T) \log \frac{\mathbb{P}_p^\pi(h^T)}{\mathbb{P}_q^\pi(h^T)} \\ &\overset{(b)}{=} \sum_{t=1}^{T} \mathbb{E}_p^\pi \left[\log \frac{P_{B^t,S^t}(p)}{P_{B^t,S^t}(q)} \right] \\ &\overset{(c)}{=} \sum_{t=1}^{T} \mathbb{E}_p^\pi \left[\mathbb{E}_p^\pi \left[\log \frac{P_{B^t,S^t}(p)}{P_{B^t,S^t}(q)} \,\middle|\, H^{t-1}, B^t \right] \right] \\ &\overset{(d)}{=} \sum_{t=1}^{T} \mathbb{E}_p^\pi \left[\sum_{S \in \mathcal{S}_{B^t}} P_{B^t,S}(p) \log \frac{P_{B^t,S}(p)}{P_{B^t,S}(q)} \right] \\ &\overset{(e)}{=} \sum_{t=1}^{T} \mathbb{E}_p^\pi [I_{B^t}(p,q)] \\ &\overset{(f)}{=} \sum_{B \in \mathcal{B}} \mathbb{E}_p^\pi [\tau^\pi(B,T)] I_B(p,q). \end{aligned} \tag{EC.1}$$

In (EC.1), (a) is the definition of KL divergence for the finite history laws, the sum ranging over the realizations h^T of H^T with positive probability under p; the absolute-continuity relation established above ensures $\mathbb{P}_q^\pi(h^T) > 0$ whenever $\mathbb{P}_p^\pi(h^T) > 0$. Equality (b) substitutes the two history factorizations, cancels their common policy factors, takes the logarithm of the resulting product and uses linearity of expectation. Equality (c) is the law of total expectation, conditioning on (H^{t-1}, B^t); conditional on these variables the selected action is fixed and, under p, S^t has distribution $P_{B^t}(p)$, so evaluating that conditional expectation over $S \in \mathcal{S}_{B^t}$ gives (d). Equality (e) is the definition of $I_{B^t}(p,q)$, and (f) groups periods according to the selected action, using $\tau^\pi(B,T) = \sum_{t=1}^{T} \mathbb{1}\{B^t = B\}$.

Step 2 (Binary Change of Measure). We bound the divergence below through a single event whose probability consistency pins down in opposite directions under p and under q, using the Bernoulli divergence between those two probabilities. Recall the gaps $\Delta_{\min}(p)$ and $\Delta_{\min}^{\mathcal{B}^*(p)}(q)$ of Section 3. Both are strictly positive here, which is what licenses the two divisions below. Since $\mathcal{B}^*(p) \cap \mathcal{B}^*(q) = \emptyset$ and $\mathcal{B}^*(q) \neq \emptyset$, the set $\mathcal{B} \setminus \mathcal{B}^*(p)$ contains $\mathcal{B}^*(q)$ and is therefore nonempty, so $\Delta_{\min}(p) > 0$; and every $B \in \mathcal{B}^*(p)$ lies outside $\mathcal{B}^*(q)$, hence is q-suboptimal, so $\Delta_{\min}^{\mathcal{B}^*(p)}(q) > 0$. In particular neither quantity is the empty minimum covered by the convention $\min \emptyset := 0$. Under p, every period outside $\mathcal{B}^*(p)$ incurs regret at least $\Delta_{\min}(p)$. Hence

$$\mathcal{R}^\pi(T,p) \geq \Delta_{\min}(p)\, \mathbb{E}_p^\pi \left[\sum_{B \notin \mathcal{B}^*(p)} \tau^\pi(B,T) \right].$$

Fix $\gamma \in (0,1)$ and define the event $G_T^\gamma := \left\{ \sum_{B \in \mathcal{B}^*(p)} \tau^\pi(B,T) \geq \gamma T \right\}$. Markov's inequality gives the first inequality below, while the preceding regret bound gives the second:

$$\mathbb{P}_p^\pi((G_T^\gamma)^c) = \mathbb{P}_p^\pi \left(\sum_{B \notin \mathcal{B}^*(p)} \tau^\pi(B,T) > (1-\gamma)T \right) \leq \frac{\mathbb{E}_p^\pi \left[\sum_{B \notin \mathcal{B}^*(p)} \tau^\pi(B,T) \right]}{(1-\gamma)T} \leq \frac{\mathcal{R}^\pi(T,p)}{(1-\gamma)T\Delta_{\min}(p)}.$$

Since π is consistent under p, $\mathcal{R}^\pi(T,p) = o(T)$. The right-hand side therefore converges to zero, so $\mathbb{P}_p^\pi(G_T^\gamma) \to 1$. Under q, every action in $\mathcal{B}^*(p)$ is suboptimal and incurs regret at least $\Delta_{\min}^{\mathcal{B}^*(p)}(q)$. Thus

$$\mathcal{R}^\pi(T,q) \geq \Delta_{\min}^{\mathcal{B}^*(p)}(q)\, \mathbb{E}_q^\pi \left[\sum_{B \in \mathcal{B}^*(p)} \tau^\pi(B,T) \right].$$

Applying the same two inequalities under q gives

$$\mathbb{P}_q^\pi(G_T^\gamma) = \mathbb{P}_q^\pi \left(\sum_{B \in \mathcal{B}^*(p)} \tau^\pi(B,T) \geq \gamma T \right) \leq \frac{\mathbb{E}_q^\pi \left[\sum_{B \in \mathcal{B}^*(p)} \tau^\pi(B,T) \right]}{\gamma T} \leq \frac{\mathcal{R}^\pi(T,q)}{\gamma T \Delta_{\min}^{\mathcal{B}^*(p)}(q)}.$$

Since π is also consistent under q, for every $\varepsilon > 0$ we have $\mathcal{R}^\pi(T,q) = o(T^\varepsilon)$. Dividing this by the factor $\gamma T \Delta_{\min}^{\mathcal{B}^*(p)}(q)$ of the preceding bound, whose only T-dependence is the factor T, gives $\mathbb{P}_q^\pi(G_T^\gamma) = o(T^{\varepsilon-1})$. Because the action counts are determined by the action coordinates of H^T, the event G_T^γ belongs to $\mathcal{F}_T = \sigma(H^T)$. For $u, v \in [0,1]$, let $d(u\|v)$ denote the KL divergence between Bernoulli laws with success probabilities u and v, under the standing extended-value conventions. Partitioning the history space over this event gives

$$\begin{aligned} D_{\mathcal{F}_T}(\mathbb{P}_p^\pi \| \mathbb{P}_q^\pi) &\overset{(a)}{=} \sum_{h^T \in G_T^\gamma} \mathbb{P}_p^\pi(h^T) \log \frac{\mathbb{P}_p^\pi(h^T)}{\mathbb{P}_q^\pi(h^T)} + \sum_{h^T \in (G_T^\gamma)^c} \mathbb{P}_p^\pi(h^T) \log \frac{\mathbb{P}_p^\pi(h^T)}{\mathbb{P}_q^\pi(h^T)} \\ &\overset{(b)}{\geq} \mathbb{P}_p^\pi(G_T^\gamma) \log \frac{\mathbb{P}_p^\pi(G_T^\gamma)}{\mathbb{P}_q^\pi(G_T^\gamma)} + \mathbb{P}_p^\pi((G_T^\gamma)^c) \log \frac{\mathbb{P}_p^\pi((G_T^\gamma)^c)}{\mathbb{P}_q^\pi((G_T^\gamma)^c)} \\ &\overset{(c)}{=} d\big(\mathbb{P}_p^\pi(G_T^\gamma) \,\big\|\, \mathbb{P}_q^\pi(G_T^\gamma)\big). \end{aligned}$$

In this display, (a) partitions the defining finite-history sum for KL divergence. Histories with zero probability under p contribute zero under the standing convention, while the absolute-continuity relation from Step 1 ensures that $\mathbb{P}_q^\pi(h^T) > 0$ whenever $\mathbb{P}_p^\pi(h^T) > 0$. Inequality (b) applies the log-sum inequality (Cover and Thomas 2006, Theorem 2.7.1) separately on G_T^γ and its complement. Equality (c) recognizes the resulting expression as the Bernoulli KL divergence. For all $u, v \in [0,1]$, it satisfies $d(u\|v) \geq u \log(1/v) - \log 2$ (Garivier et al. 2019, Equation (11)). To obtain the logarithmic lower bound, fix $\eta, \varepsilon \in (0,1)$. Since $\mathbb{P}_p^\pi(G_T^\gamma) \to 1$ and, as established above, $\mathbb{P}_q^\pi(G_T^\gamma) = o(T^{\varepsilon-1})$, there exists T_0 such that, for every $T \geq T_0$, $\mathbb{P}_p^\pi(G_T^\gamma) \geq 1-\eta > 0$ and $\mathbb{P}_q^\pi(G_T^\gamma) \leq T^{\varepsilon-1}$. By absolute continuity

from Step 1, the first bound implies $\mathbb{P}_q^\pi(G_T^\gamma) > 0$; hence the second bound gives $-\log \mathbb{P}_q^\pi(G_T^\gamma) \geq (1-\varepsilon)\log T$. Substituting these bounds into the Bernoulli-KL inequality gives

$$d\big(\mathbb{P}_p^\pi(G_T^\gamma) \,\big\|\, \mathbb{P}_q^\pi(G_T^\gamma)\big) \geq \mathbb{P}_p^\pi(G_T^\gamma) \log \frac{1}{\mathbb{P}_q^\pi(G_T^\gamma)} - \log 2 \geq (1-\eta)(1-\varepsilon)\log T - \log 2.$$

Dividing by $\log T$, taking the limit inferior, and then letting $\eta \downarrow 0$ and $\varepsilon \downarrow 0$, gives

$$\liminf_{T\to\infty} \frac{D_{\mathcal{F}_T}(\mathbb{P}_p^\pi \| \mathbb{P}_q^\pi)}{\log T} \geq 1.$$

Substituting the identity (EC.1) of Step 1, which expresses $D_{\mathcal{F}_T}(\mathbb{P}_p^\pi \| \mathbb{P}_q^\pi)$ as $\sum_{B\in\mathcal{B}} \mathbb{E}_p^\pi[\tau^\pi(B,T)] I_B(p,q)$, into the displayed limit gives

$$\liminf_{T\to\infty} \frac{\sum_{B\in\mathcal{B}} \mathbb{E}_p^\pi[\tau^\pi(B,T)] I_B(p,q)}{\log T} \geq 1. \qquad \blacksquare$$

LEMMA EC.1 **(Observational-Equivalence Obstruction)**. *Suppose that $P_B(p) = P_B(q)$ for every $B \in \mathcal{B}$ and $\mathcal{B}^*(p) \cap \mathcal{B}^*(q) = \emptyset$. Then every admissible policy induces the same law over observable histories under p and q. Moreover, for every horizon T,*

$$\max\{\mathcal{R}^\pi(T,p), \mathcal{R}^\pi(T,q)\} \geq C_{p,q} T$$

for some constant $C_{p,q} > 0$ independent of T. Consequently, no policy can be consistent over any model class containing both p and q.

Proof of Lemma EC.1. The proof has two parts. The first shows that the two environments induce the same law over observable histories; the second turns that into a linear regret bound. Let $\pi(B \mid h)$ denote the probability that policy π selects action $B \in \mathcal{B}$ after observable history h. We prove equality of the history laws by induction on t. The base case follows from $\mathbb{P}_p^\pi(H^0 = \emptyset) = 1 = \mathbb{P}_q^\pi(H^0 = \emptyset)$. Now fix $t \geq 1$ and suppose that $\mathbb{P}_p^\pi(H^{t-1} = h^{t-1}) = \mathbb{P}_q^\pi(H^{t-1} = h^{t-1})$ for every possible history h^{t-1}. Then, for every $B \in \mathcal{B}$ and $S \in \mathcal{P}(B)$,

$$\begin{aligned} \mathbb{P}_p^\pi(H^{t-1} = h^{t-1}, B^t = B, S^t = S) &\overset{(a)}{=} \mathbb{P}_p^\pi(H^{t-1} = h^{t-1}) \cdot P_{B,S}(p) \cdot \pi(B \mid h^{t-1}) \\ &\overset{(b)}{=} \mathbb{P}_q^\pi(H^{t-1} = h^{t-1}) \cdot P_{B,S}(q) \cdot \pi(B \mid h^{t-1}) \\ &\overset{(c)}{=} \mathbb{P}_q^\pi(H^{t-1} = h^{t-1}, B^t = B, S^t = S). \end{aligned} \tag{EC.2}$$

In (EC.2), equalities (a) and (c) factor the joint probability under p and under q; (b) applies the induction hypothesis and $P_B(p) = P_B(q)$, the policy factor being common to both. Since $H^t = (H^{t-1}, B^t, S^t)$, this covers every value of H^t, so every admissible policy induces the same law over observable histories under p and q. The action-count vector is a deterministic function of H^T, so it too has the same law under both environments.

We now turn to regret. Since $\mathcal{B}^*(q)$ is nonempty and disjoint from $\mathcal{B}^*(p)$, the set $\mathcal{B} \setminus \mathcal{B}^*(p)$ is nonempty. Every action in this set has a strictly positive gap under p. Likewise, $\mathcal{B}^*(p)$ is nonempty, and every action in $\mathcal{B}^*(p)$ has a strictly positive gap under q. Because $\mathcal{B}$ is finite, the gaps $\Delta_{\min}(p)$ and $\Delta_{\min}^{\mathcal{B}^*(p)}(q)$ of Section 3 are well-defined and strictly positive. Using the action-count representation of regret gives

$$\mathcal{R}^\pi(T,p) = \sum_{B \notin \mathcal{B}^*(p)} \mathbb{E}_p^\pi[\tau^\pi(B,T)]\Delta_B(p) \geq \sum_{B \notin \mathcal{B}^*(p)} \mathbb{E}_p^\pi[\tau^\pi(B,T)]\Delta_{\min}(p),$$
$$\mathcal{R}^\pi(T,q) \geq \sum_{B \in \mathcal{B}^*(p)} \mathbb{E}_q^\pi[\tau^\pi(B,T)]\Delta_B(q) \geq \sum_{B \in \mathcal{B}^*(p)} \mathbb{E}_q^\pi[\tau^\pi(B,T)]\Delta_{\min}^{\mathcal{B}^*(p)}(q) = \Delta_{\min}^{\mathcal{B}^*(p)}(q) \sum_{B \in \mathcal{B}^*(p)} \mathbb{E}_p^\pi[\tau^\pi(B,T)].$$

Both lines bound each gap below by the corresponding minimum gap, both positive as just shown. The second line first discards the nonnegative terms with $B \notin \mathcal{B}^*(p)$, and the last equality uses the equality of the action-count laws. Writing $C_{p,q} := \frac{1}{2}\min\{\Delta_{\min}(p), \Delta_{\min}^{\mathcal{B}^*(p)}(q)\} > 0$, we therefore obtain

$$\begin{aligned}
\max\{\mathcal{R}^\pi(T,p), \mathcal{R}^\pi(T,q)\} &\geq \frac{1}{2}\big\{\mathcal{R}^\pi(T,p) + \mathcal{R}^\pi(T,q)\big\} \\
&\geq C_{p,q}\left\{\sum_{B \notin \mathcal{B}^*(p)} \mathbb{E}_p^\pi[\tau^\pi(B,T)] + \sum_{B \in \mathcal{B}^*(p)} \mathbb{E}_p^\pi[\tau^\pi(B,T)]\right\} \\
&= C_{p,q}T.
\end{aligned}$$

The first step bounds the maximum by the average; the second substitutes the two bounds above, whose coefficients are both at least $C_{p,q}$. The final equality holds because $\mathcal{B}^*(p)$ and $\mathcal{B} \setminus \mathcal{B}^*(p)$ partition the feasible action set and exactly one action is selected in each period. This is the stated bound. If a policy were consistent over a model class containing both p and q, then, taking $\alpha = 1$ in the definition of consistency, both regrets would be $o(T)$. This contradicts the linear lower bound above. ∎

Proof of Corollary 1. The argument extracts from any bounded-regret subsequence a limiting play-count allocation and uses the information floor to show that this allocation is feasible for the lower-bound problem. We distinguish according to whether $\mathcal{Q}_{\mathcal{M}}(p)$ is empty. If $\mathcal{Q}_{\mathcal{M}}(p) = \emptyset$, the lower-bound problem (1) has no information constraints. The zero allocation is therefore feasible and has objective value zero, so $\kappa^*_{\mathcal{M}}(p) = 0$. Since regret is nonnegative, the result follows.

Suppose now that $\mathcal{Q}_{\mathcal{M}}(p) \neq \emptyset$. Then $\mathcal{B} \setminus \mathcal{B}^*(p)$ is nonempty: for every $q \in \mathcal{Q}_{\mathcal{M}}(p)$, the optimal set $\mathcal{B}^*(q)$ is nonempty and disjoint from $\mathcal{B}^*(p)$. Since optimal actions have zero gap,

$$\frac{\mathcal{R}^\pi(T,p)}{\log T} = \sum_{B \notin \mathcal{B}^*(p)} \frac{\mathbb{E}_p^\pi[\tau^\pi(B,T)]}{\log T}\Delta_B(p).$$

Take a subsequence $(T_n)_{n \geq 1}$ along which the left-hand side converges to its limit inferior. If this limit inferior is infinite, there is nothing to prove. Otherwise, boundedness of the regret ratio and

positivity of every suboptimal gap imply that each coordinate sequence $\mathbb{E}_p^\pi[\tau^\pi(B,T_n)]/\log T_n$ is bounded. Because $\mathcal{B}\setminus\mathcal{B}^*(p)$ is finite, we may pass to a further subsequence, relabeled $(T_n)_{n\geq 1}$, along which all coordinates converge. For each $B\notin\mathcal{B}^*(p)$, write $x_B := \lim_{n\to\infty}\mathbb{E}_p^\pi[\tau^\pi(B,T_n)]/\log T_n \in [0,+\infty)$.

Fix any $q\in\mathcal{Q}_{\mathcal{M}}(p)$. By definition, we have $\mathcal{B}^*(p)\cap\mathcal{B}^*(q)=\emptyset$, $P_B(p)\ll P_B(q)$ for every $B\in\mathcal{B}$, and $I_B(p,q)=0$ for every $B\in\mathcal{B}^*(p)$. Moreover, both p and q belong to $\mathcal{M}$, so π is consistent under both distributions. Because each $\mathcal{S}_B$ is finite and $P_B(p)\ll P_B(q)$ for every B, we have $P_{B,S}(q)>0$ whenever $P_{B,S}(p)>0$, for every $S\in\mathcal{S}_B$. Thus $I_B(p,q)<+\infty$ for every B. Applying Theorem 1 and then evaluating its information ratio along $(T_n)_{n\geq 1}$ gives

$$1\leq\liminf_{T\to\infty}\frac{\sum_{B\in\mathcal{B}}\mathbb{E}_p^\pi[\tau^\pi(B,T)]I_B(p,q)}{\log T}\leq\lim_{n\to\infty}\frac{\sum_{B\in\mathcal{B}}\mathbb{E}_p^\pi[\tau^\pi(B,T_n)]I_B(p,q)}{\log T_n}=\sum_{B\notin\mathcal{B}^*(p)}x_B I_B(p,q).$$

The second inequality holds because every subsequential limit dominates the full-sequence limit inferior; the equality uses the coordinatewise limits defining x and the zero information of the p-optimal actions. Since q was arbitrary, $x=(x_B)_{B\notin\mathcal{B}^*(p)}$ is a finite-cost feasible allocation for (1). Returning to the regret ratio, the decomposition above, the choice of $(T_n)_{n\geq 1}$, and the coordinatewise limits defining x give the equality below, while feasibility of x gives the inequality:

$$\liminf_{T\to\infty}\frac{\mathcal{R}^\pi(T,p)}{\log T}=\sum_{B\notin\mathcal{B}^*(p)}x_B\Delta_B(p)\geq\kappa^*_{\mathcal{M}}(p).$$ ∎

Appendix B: Proofs for the Bayesian Learning Policies

This appendix proves the results of Section 5 in the order they appear there. We first settle the full-feedback benchmark: Proposition 1 exhibits a two-scenario instance on which no policy beats $\sqrt{T}$, and Proposition 2 shows that full-feedback Thompson sampling matches that rate, so the two together pin the benchmark from both sides. We then derive the censored posterior of Proposition 3, which is what makes path-only Thompson sampling well defined at all, and Proposition 4 gives the Gibbs conditionals that turn it into something samplable. Finally, Proposition 5 shows that exact path-only Thompson sampling can nonetheless incur linear regret, which is the negative result the rest of the paper responds to.

Proof of Proposition 1. Fix a horizon $T\geq 2$. We use a fixed two-scenario interdiction instance and compare two nearby latent distributions on its scenario support. Only the separation between these distributions is calibrated to T; the network, cost support, and feasible actions are fixed across horizons.

Step 1 (A Two-Scenario Interdiction Instance). Consider a graph consisting of two parallel arcs a_1 and a_2 from the source to the sink, with interdiction budget $K=1$. Let $B_1:=\{a_2\}$ force the

evader onto a_1, and let $B_2 := \{a_1\}$ force the evader onto a_2. The empty interdiction is also feasible. In scenario 1, the costs of (a_1, a_2) are $(2,1)$, while in scenario 2 they are $(1,2)$. Hence

$$r_{B_1} = (2,1)^\top, \qquad r_{B_2} = (1,2)^\top, \qquad r_\emptyset = (1,1)^\top.$$

For $\varepsilon \in (0, 1/4)$, consider $p^+ = (1/2+\varepsilon, 1/2-\varepsilon)$ and $p^- = (1/2-\varepsilon, 1/2+\varepsilon)$. Under p^+, actions B_1 and B_2 have values $3/2+\varepsilon$ and $3/2-\varepsilon$, respectively, while the empty interdiction has value 1. Thus B_1 is uniquely optimal, B_2 has gap 2ε, and the empty interdiction has gap $1/2+\varepsilon \geq 2\varepsilon$. By symmetry, B_2 is uniquely optimal under p^-, and every action other than B_2 has gap at least 2ε.

Step 2 (KL Divergence between the Experiments). Let $\mathcal{F}_t^{\mathrm{F}} := \sigma(H^{\mathrm{F},t-1}, B^t)$ be the σ-field generated by the full-feedback history available immediately after the policy selects B^t, so that it records the first $t-1$ actions and scenario observations together with B^t. For a realization $h^{\mathrm{F},t-1} = (b^1, k^1, \ldots, b^{t-1}, k^{t-1})$ and action b^t, its mass under p^+ or p^- factors as

$$\mathbb{P}^\pi_{p^\pm}(h^{\mathrm{F},t-1}, b^t) = \prod_{s=1}^{t-1} \left[\pi(b^s \mid h^{\mathrm{F},s-1}) p^\pm_{k^s}\right] \pi(b^t \mid h^{\mathrm{F},t-1}),$$

where the conditioning history in the first product is empty when $s=1$. Evaluating these mass functions at the random pre-decision history gives

$$\begin{aligned} D_{\mathcal{F}_t^{\mathrm{F}}}(\mathbb{P}^\pi_{p^+} \| \mathbb{P}^\pi_{p^-}) &\overset{(a)}{=} \mathbb{E}^\pi_{p^+}\left[\log \frac{\mathbb{P}^\pi_{p^+}(H^{\mathrm{F},t-1}, B^t)}{\mathbb{P}^\pi_{p^-}(H^{\mathrm{F},t-1}, B^t)}\right] \\ &\overset{(b)}{=} \mathbb{E}^\pi_{p^+}\left[\log \prod_{s=1}^{t-1} \frac{p^+_{k^s}}{p^-_{k^s}}\right] \\ &= (t-1) D(p^+ \| p^-). \end{aligned}$$

Equality (a) is the definition of KL divergence for the finite pre-decision history laws, while (b) substitutes the two factorizations and cancels their common policy factors. The final equality is the direct-scenario-observation specialization of the adaptive relative-entropy calculation in (EC.1): under p^+, the first $t-1$ scenario indices are i.i.d. with distribution p^+, so each completed period contributes $D(p^+ \| p^-)$.

For the two distributions in Step 1, the one-period divergence is

$$D(p^+ \| p^-) = \left(\frac{1}{2}+\varepsilon\right) \log \frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}-\varepsilon} + \left(\frac{1}{2}-\varepsilon\right) \log \frac{\frac{1}{2}-\varepsilon}{\frac{1}{2}+\varepsilon} = 2\varepsilon \log \frac{1+2\varepsilon}{1-2\varepsilon}.$$

For $x \in (0, 1/2)$, the tangent bound for the logarithm and $1-x \geq 1/2$ give $\log((1+x)/(1-x)) \leq (1+x)/(1-x) - 1 = 2x/(1-x) \leq 4x$. Taking $x = 2\varepsilon$ therefore yields $D(p^+ \| p^-) \leq 16\varepsilon^2$, and hence

$$D_{\mathcal{F}_t^{\mathrm{F}}}(\mathbb{P}^\pi_{p^+} \| \mathbb{P}^\pi_{p^-}) \leq 16(t-1)\varepsilon^2.$$

Step 3 (Testing the Two Environments). By the Bretagnolle–Huber inequality (Lattimore and Szepesvári 2020, Theorem 14.2), for any event A, $P(A)+Q(A^c)\geq \frac{1}{2}\exp\{-D(P\|Q)\}$. Apply it to the event $\{B^t\neq B_1\}$ under the two pre-decision laws; the event is measurable with respect to these laws precisely because they include the period-t action. This gives

$$\mathbb{P}^{\pi}_{p^+}(B^t\neq B_1)+\mathbb{P}^{\pi}_{p^-}(B^t=B_1)\geq \frac{1}{2}\exp\{-16(t-1)\varepsilon^2\}.$$

Choose $\varepsilon=1/(4\sqrt{T})$. Then $16(t-1)\varepsilon^2\leq 1$ for every $t\leq T$, so the sum of the two error probabilities is at least $1/(2e)$ in every period.

Step 4 (From Testing Error to Regret). Under p^+, every action different from B_1 incurs regret at least 2ε, so $\mathcal{R}^{\pi}(T,p^+)\geq 2\varepsilon\sum_{t=1}^{T}\mathbb{P}^{\pi}_{p^+}(B^t\neq B_1)$. Under p^-, playing B_1 incurs regret 2ε, so $\mathcal{R}^{\pi}(T,p^-)\geq 2\varepsilon\sum_{t=1}^{T}\mathbb{P}^{\pi}_{p^-}(B^t=B_1)$. Adding the two and bounding each period by the testing bound $1/(2e)$ of Step 3 gives

$$\mathcal{R}^{\pi}(T,p^+)+\mathcal{R}^{\pi}(T,p^-)\geq 2\varepsilon\sum_{t=1}^{T}\left[\mathbb{P}^{\pi}_{p^+}(B^t\neq B_1)+\mathbb{P}^{\pi}_{p^-}(B^t=B_1)\right]\geq\frac{\varepsilon T}{e}.$$

The larger of the two regrets is at least half of their sum, and $\varepsilon=1/(4\sqrt{T})$ gives $\varepsilon T/(2e)=T/(8e\sqrt{T})=\sqrt{T}/(8e)$. Since $p^+,p^-\in\Delta^1$,

$$\sup_{p\in\Delta^1}\mathcal{R}^{\pi}(T,p)\geq\max\left\{\mathcal{R}^{\pi}(T,p^+),\mathcal{R}^{\pi}(T,p^-)\right\}\geq\frac{1}{2}\cdot\frac{\varepsilon T}{e}=\frac{\sqrt{T}}{8e}.$$ ∎

Proof of Proposition 2. Fix $p\in\Delta^{m-1}$ and choose any $B^*\in\mathcal{B}^*(p)$. The proof first reduces regret to the distance between the Thompson draw and p, then bounds that distance using the posterior variance and the error of the posterior mean.

Step 1 (Regret Is Controlled by Posterior Error). Since $V(p)=r_{B^*}^{\top}p$, the gap satisfies $\Delta_{B^t}(p)=(r_{B^*}-r_{B^t})^{\top}p$. Adding and subtracting the sampled values gives

$$\begin{aligned}\Delta_{B^t}(p)&=(r_{B^*}-r_{B^t})^{\top}(p-\widetilde{p}^t)+(r_{B^*}-r_{B^t})^{\top}\widetilde{p}^t\\&\overset{(a)}{\leq}(r_{B^*}-r_{B^t})^{\top}(p-\widetilde{p}^t)\\&\overset{(b)}{=}\min_{a\in\mathbb{R}}(r_{B^*}-r_{B^t}-a\mathbf{1})^{\top}(p-\widetilde{p}^t)\\&\overset{(c)}{\leq}\min_{a\in\mathbb{R}}\|r_{B^*}-r_{B^t}-a\mathbf{1}\|_2\|\widetilde{p}^t-p\|_2\\&\leq\operatorname{diam}_{\perp}(r)\|\widetilde{p}^t-p\|_2.\end{aligned}\tag{EC.3}$$

In (EC.3), inequality (a) follows from the optimality of B^t under the posterior draw: $r_{B^t}^{\top}\widetilde{p}^t\geq r_{B^*}^{\top}\widetilde{p}^t$. Equality (b) uses $\mathbf{1}^{\top}(p-\widetilde{p}^t)=0$, and (c) applies the Cauchy–Schwarz inequality for each $a\in\mathbb{R}$

before minimizing. The last inequality follows from the definition of $\mathrm{diam}_\perp(r)$ for the pair (B^*, B^t). Taking expectations and summing over periods,

$$\mathcal{R}^{\mathrm{fTS}}(T,p) \le \mathrm{diam}_\perp(r) \sum_{t=1}^{T} \mathbb{E}_p^{\mathrm{fTS}}[\|\widetilde{p}^t - p\|_2]. \tag{EC.4}$$

Step 2 (Posterior Estimation Error). Recall that, conditional on $H^{\mathrm{F},t-1}$, $\widetilde{p}^t$ is Dirichlet with parameters $\alpha_k^0 + N_k(t-1)$, $k \in [m]$, total concentration $\bar{\alpha} + t - 1$, and conditional mean $\bar{p}^t$, whose coordinates are $\bar{p}_k^t := (\alpha_k^0 + N_k(t-1))/(\bar{\alpha} + t - 1)$, $k \in [m]$. Adding and subtracting $\bar{p}^t$, applying the triangle inequality pointwise, and using linearity of expectation gives

$$\mathbb{E}_p^{\mathrm{fTS}}[\|\widetilde{p}^t - p\|_2] = \mathbb{E}_p^{\mathrm{fTS}}\left[\|(\widetilde{p}^t - \bar{p}^t) + (\bar{p}^t - p)\|_2\right] \le \mathbb{E}_p^{\mathrm{fTS}}[\|\widetilde{p}^t - \bar{p}^t\|_2] + \mathbb{E}_p^{\mathrm{fTS}}[\|\bar{p}^t - p\|_2].$$

The two terms on the right are the posterior sampling fluctuation and the posterior mean error, respectively. Conditional on the observed counts, the sampling fluctuation satisfies

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{p}^t - \bar{p}^t\|_2^2 \,\middle|\, N_1(t-1), \ldots, N_m(t-1)\right] &= \sum_{k=1}^{m} \mathrm{Var}\left(\widetilde{p}_k^t \,\middle|\, N_1(t-1), \ldots, N_m(t-1)\right) \\
&= \frac{\sum_{k=1}^{m} \bar{p}_k^t(1-\bar{p}_k^t)}{\bar{\alpha} + t} \\
&= \frac{1 - \|\bar{p}^t\|_2^2}{\bar{\alpha} + t} \\
&\le \frac{1}{\bar{\alpha} + t}.
\end{aligned}$$

The first equality expands the squared Euclidean norm and uses that $\bar{p}^t$ is the conditional mean of $\widetilde{p}^t$; no covariance terms appear because the norm is a sum of coordinatewise squares. The second equality is the Dirichlet marginal-variance formula, whose denominator is one plus the total concentration, $\bar{\alpha} + t$. The remaining equality uses $\sum_{k=1}^{m} \bar{p}_k^t = 1$, and the inequality uses $\|\bar{p}^t\|_2^2 \ge 0$. The law of total expectation and conditional Jensen's inequality give

$$\begin{aligned}
\mathbb{E}_p^{\mathrm{fTS}}[\|\widetilde{p}^t - \bar{p}^t\|_2] &= \mathbb{E}_p^{\mathrm{fTS}}\left[\mathbb{E}\left[\|\widetilde{p}^t - \bar{p}^t\|_2 \,\middle|\, N_1(t-1), \ldots, N_m(t-1)\right]\right] \\
&\le \mathbb{E}_p^{\mathrm{fTS}}\left[\left(\mathbb{E}\left[\|\widetilde{p}^t - \bar{p}^t\|_2^2 \,\middle|\, N_1(t-1), \ldots, N_m(t-1)\right]\right)^{1/2}\right] \\
&\le \frac{1}{\sqrt{\bar{\alpha} + t}}.
\end{aligned}$$

For the posterior-mean term, write $N(t-1) := (N_1(t-1), \ldots, N_m(t-1))$. Since $\bar{p}^t = (\alpha^0 + N(t-1))/(\bar{\alpha} + t - 1)$, subtracting p gives $\bar{p}^t - p = [N(t-1) - (t-1)p + \alpha^0 - \bar{\alpha}p]/(\bar{\alpha} + t - 1)$. This identity separates a random count fluctuation from a deterministic prior-bias term. Because the scenarios are i.i.d. with law p and their distribution is unaffected by the selected actions, $N(t-1) \sim \mathrm{Multinomial}(t-1, p)$. In particular, each coordinate has mean $(t-1)p_k$ and variance $(t-1)p_k(1-p_k)$, so

$$\mathbb{E}_p^{\mathrm{fTS}}[\|N(t-1) - (t-1)p\|_2^2] = \sum_{k=1}^{m} \mathbb{E}_p^{\mathrm{fTS}}[(N_k(t-1) - (t-1)p_k)^2]$$

$$
\begin{aligned}
&= \sum_{k=1}^{m} \mathrm{Var}_p(N_k(t-1)) \\
&= (t-1)\sum_{k=1}^{m} p_k(1-p_k) \\
&= (t-1)(1-\|p\|_2^2) \le t-1.
\end{aligned}
$$

The deterministic prior term satisfies $\|\alpha^0 - \bar{\alpha}p\|_2 \le \|\alpha^0\|_1 + \bar{\alpha}\|p\|_1 = 2\bar{\alpha}$. Applying the triangle and Jensen inequalities to the decomposition above now gives

$$
\begin{aligned}
\mathbb{E}_p^{\mathrm{fTS}}[\|\bar{p}^t - p\|_2] &\le \frac{\mathbb{E}_p^{\mathrm{fTS}}[\|N(t-1)-(t-1)p\|_2] + \|\alpha^0 - \bar{\alpha}p\|_2}{\bar{\alpha}+t-1} \\
&\le \frac{\left(\mathbb{E}_p^{\mathrm{fTS}}[\|N(t-1)-(t-1)p\|_2^2]\right)^{1/2} + \|\alpha^0 - \bar{\alpha}p\|_2}{\bar{\alpha}+t-1} \\
&\le \frac{\sqrt{t-1}+2\bar{\alpha}}{\bar{\alpha}+t-1}.
\end{aligned}
$$

Combining this posterior-mean bound with the sampling-fluctuation bound gives

$$
\mathbb{E}_p^{\mathrm{fTS}}[\|\widetilde{p}^t - p\|_2] \le \mathbb{E}_p^{\mathrm{fTS}}[\|\widetilde{p}^t - \bar{p}^t\|_2] + \mathbb{E}_p^{\mathrm{fTS}}[\|\bar{p}^t - p\|_2] \le \frac{1}{\sqrt{\bar{\alpha}+t}} + \frac{\sqrt{t-1}}{\bar{\alpha}+t-1} + \frac{2\bar{\alpha}}{\bar{\alpha}+t-1}.
$$

Since $\bar{\alpha} > 0$, for $t \ge 2$ we have $1/\sqrt{\bar{\alpha}+t} \le 1/\sqrt{t}$ and $\sqrt{t-1}/(\bar{\alpha}+t-1) \le 1/\sqrt{t-1} \le \sqrt{2}/\sqrt{t}$. The second summand is zero when $t=1$. Thus the first two terms are bounded by $(1+\sqrt{2})/\sqrt{t}$ for every $t \ge 1$. Since $1+\sqrt{2} > 2$, a single universal constant $C \ge 1+\sqrt{2}$ also covers the factor 2 in the prior-mass term, giving

$$
\mathbb{E}_p^{\mathrm{fTS}}[\|\widetilde{p}^t - p\|_2] \le \frac{C}{\sqrt{t}} + \frac{C\bar{\alpha}}{\bar{\alpha}+t-1}.
$$

Step 3 (Summing the Posterior Error). Using $\sum_{t=1}^{T} t^{-1/2} \le 2\sqrt{T}$, we next sum the prior-mass term. Its $t=1$ summand equals one. For each $t \ge 2$, the function $1/(\bar{\alpha}+x)$ is decreasing in x, so its integral over the unit interval $[t-2, t-1]$ is at least its value at the right endpoint:

$$
\frac{\bar{\alpha}}{\bar{\alpha}+t-1} = \bar{\alpha}\int_{t-2}^{t-1} \frac{1}{\bar{\alpha}+t-1}dx \le \bar{\alpha}\int_{t-2}^{t-1} \frac{dx}{\bar{\alpha}+x}.
$$

Summing these period-specific bounds over $t = 2, \ldots, T$, so that the intervals tile $[0, T-1]$, gives

$$
\sum_{t=1}^{T} \frac{\bar{\alpha}}{\bar{\alpha}+t-1} \le 1 + \bar{\alpha}\int_{0}^{T-1} \frac{dx}{\bar{\alpha}+x} = 1 + \bar{\alpha}\log\left(1+\frac{T-1}{\bar{\alpha}}\right) \le 1 + \bar{\alpha}\log\left(1+\frac{T}{\bar{\alpha}}\right).
$$

Substituting these two summation bounds into (EC.4), and enlarging C to absorb the factor 2 from $\sum_{t=1}^{T} t^{-1/2} \le 2\sqrt{T}$, yields

$$
\mathcal{R}^{\mathrm{fTS}}(T,p) \le C\,\mathrm{diam}_{\perp}(r)\left[\sqrt{T} + 1 + \bar{\alpha}\log\left(1+\frac{T}{\bar{\alpha}}\right)\right].
$$

For fixed $\bar{\alpha}$, the last two terms are $O(\sqrt{T})$, uniformly over $p \in \Delta^{m-1}$. Finally, if $0 \le r_{B,k} \le r_{\max}$, then every coordinate of $r_B - r_{B'}$ has absolute value at most $r_{\max}$. Orthogonal projection cannot increase Euclidean norm, so $\mathrm{diam}_{\perp}(r) \le r_{\max}\sqrt{m}$. Substituting this bound proves the last claim. ■

B.1. Censored-Feedback Gibbs Representation

Proof of Proposition 3. Fix a possible realization $h^t = (B^1, S^1, \ldots, B^t, S^t)$ of H^t, and let h^{s-1} denote its prefix through period $s-1$. The probability of this history under p factors as

$$\mathbb{P}_p^\pi(H^t = h^t) = \prod_{s=1}^{t} \pi(B^s \mid h^{s-1}) P_{B^s,S^s}(p) = \prod_{s=1}^{t} \pi(B^s \mid h^{s-1}) \left(\sum_{k \in \mathcal{Z}^s} p_k \right).$$

The first equality follows from the sequential construction of the history: after history h^{s-1}, the policy selects B^s according to $\pi(\cdot \mid h^{s-1})$, and the resulting path has distribution $P_{B^s}(p)$. The second equality follows from the definitions of $P_{B^s}(p)$ and $\mathcal{Z}^s$. For the fixed history h^t, each policy factor $\pi(B^s \mid h^{s-1})$ is independent of the candidate distribution p. Let f_0 denote the Dirichlet prior density. Bayes' rule and the history factorization established above give

$$f_t(p) = \frac{f_0(p)\mathbb{P}_p^\pi(H^t = h^t)}{\displaystyle\int_{\Delta^{m-1}} f_0(q)\mathbb{P}_q^\pi(H^t = h^t)\, dq} = \frac{f_0(p) \displaystyle\prod_{s=1}^{t} \left(\sum_{k \in \mathcal{Z}^s} p_k \right)}{\displaystyle\int_{\Delta^{m-1}} f_0(q) \prod_{s=1}^{t} \left(\sum_{k \in \mathcal{Z}^s} q_k \right) dq}.$$

The second equality follows because the common policy factor $\prod_{s=1}^{t} \pi(B^s \mid h^{s-1})$ appears in both the numerator and denominator. It is strictly positive for a possible history h^t and therefore cancels. Finally, substituting the Dirichlet prior kernel $\prod_{k=1}^{m} p_k^{\alpha_k^0 - 1}$ and omitting normalizing factors that do not depend on p yields

$$f_t(p) \propto \prod_{k=1}^{m} p_k^{\alpha_k^0 - 1} \prod_{s=1}^{t} \left(\sum_{k \in \mathcal{Z}^s} p_k \right), \qquad p \in \Delta^{m-1}.$$

■

Proof of Proposition 4. The proof has four parts: we write the joint posterior kernel of (p, Z), verify that its p-marginal is the censored posterior of (2), and then read off the two conditional laws, the second of which we also present in collapsed form. Fix a possible realization h^t of the path-feedback history H^t and a candidate assignment vector $z \in [m]^t$. By Bayes' rule, the joint posterior kernel of (p, Z) is the Dirichlet prior kernel multiplied by the complete-data likelihood. The latter factors as

$$\mathbb{P}_p^\pi(H^t = h^t, Z = z) = \prod_{s=1}^{t} \pi(B^s \mid h^{s-1}) p_{z^s} \mathbb{1}\{z^s \in \mathcal{Z}^s\}.$$

For the fixed history h^t, the policy factors are independent of both p and z and can therefore be absorbed into the posterior normalizing constant. Hence the joint posterior kernel is

$$\prod_{k=1}^{m} p_k^{\alpha_k^0 - 1} \prod_{s=1}^{t} p_{z^s} \mathbb{1}\{z^s \in \mathcal{Z}^s\} = \prod_{k=1}^{m} p_k^{\alpha_k^0 - 1 + n_k(z)} \prod_{s=1}^{t} \mathbb{1}\{z^s \in \mathcal{Z}^s\}.$$

The factor $\prod_{k=1}^{m} p_k^{\alpha_k^0-1}$ is the Dirichlet prior kernel. For each period s, p_{z^s} is the probability of scenario assignment z^s, while the indicator enforces consistency with the path observed in that period. Moreover, $n_k(z)$ is the number of periods assigned to scenario k. Hence p_k appears exactly $n_k(z)$ times in $\prod_{s=1}^{t} p_{z^s}$, which gives the equality.

Before deriving the conditional laws, we verify that the augmented posterior has the censored posterior as its marginal in p. Summing the joint posterior kernel over all assignment vectors gives

$$\sum_{z\in[m]^t}\left[\prod_{k=1}^{m} p_k^{\alpha_k^0-1}\prod_{s=1}^{t} p_{z^s}\mathbb{1}\{z^s\in\mathcal{Z}^s\}\right]=\prod_{k=1}^{m} p_k^{\alpha_k^0-1}\sum_{z\in[m]^t}\prod_{s=1}^{t} p_{z^s}\mathbb{1}\{z^s\in\mathcal{Z}^s\}=\prod_{k=1}^{m} p_k^{\alpha_k^0-1}\prod_{s=1}^{t}\sum_{j\in\mathcal{Z}^s} p_j.$$

The first equality moves the Dirichlet prior kernel outside the sum because it does not depend on z. The second equality factors the sum over assignment vectors into period-specific sums. For each period s, summing over the possible values of z^s gives $\sum_{j\in\mathcal{Z}^s} p_j$. Thus the final expression is the Dirichlet prior kernel multiplied by the censored path-feedback likelihood, and hence is exactly the posterior kernel in (2).

We next derive the conditional distribution of p. An assignment vector that violates $z^s\in\mathcal{Z}^s$ for some s has zero posterior probability. For any compatible assignment vector, all the indicator factors equal one. The conditional kernel of p is therefore

$$\prod_{k=1}^{m} p_k^{\alpha_k^0-1+n_k(z)},$$

which is the kernel of a Dirichlet distribution with parameters $\alpha_k^0+n_k(z)$, $k\in[m]$.

To derive the second conditional law, return to the joint kernel, fix $s\in[t]$, write $Z^{-s}:=(Z^r: r\in[t]\setminus\{s\})$, and condition on p and $Z^{-s}=z^{-s}$. All factors that do not involve Z^s are common across its possible values and cancel upon normalization. Hence, for every $k\in[m]$,

$$\mathbb{P}(Z^s=k\mid p, Z^{-s}=z^{-s}, H^t=h^t)=\frac{p_k\mathbb{1}\{k\in\mathcal{Z}^s\}}{\sum_{j=1}^{m} p_j\mathbb{1}\{j\in\mathcal{Z}^s\}}=\frac{p_k\mathbb{1}\{k\in\mathcal{Z}^s\}}{\sum_{j\in\mathcal{Z}^s} p_j}.$$

This expression does not depend on z^{-s}. Applying the law of total expectation over Z^{-s} therefore gives the stated conditional distribution of Z^s given p and H^t.

It remains to derive the collapsed conditional. Fix a compatible realization z^{-s} and define, for every $i\in[m]$, $\tilde{\alpha}_i:=\alpha_i^0+\sum_{r\in[t]\setminus\{s\}}\mathbb{1}\{z^r=i\}$. With z^{-s} fixed and $Z^s=k$, the joint kernel is proportional to $p_k\prod_{i=1}^{m} p_i^{\tilde{\alpha}_i-1}\mathbb{1}\{k\in\mathcal{Z}^s\}$: the factors common to all candidate values of Z^s cancel upon normalization, and the remaining assignment counts are absorbed into the exponents $\tilde{\alpha}_i$. Integrating out p over the simplex and normalizing over k, we obtain

$$\mathbb{P}(Z^s=k\mid Z^{-s}=z^{-s}, H^t=h^t)=\frac{\mathbb{1}\{k\in\mathcal{Z}^s\}\displaystyle\int_{\Delta^{m-1}} p_k\prod_{i=1}^{m} p_i^{\tilde{\alpha}_i-1}\,dp}{\displaystyle\sum_{j\in\mathcal{Z}^s}\int_{\Delta^{m-1}} p_j\prod_{i=1}^{m} p_i^{\tilde{\alpha}_i-1}\,dp}=\frac{\tilde{\alpha}_k\mathbb{1}\{k\in\mathcal{Z}^s\}}{\sum_{j\in\mathcal{Z}^s}\tilde{\alpha}_j}.$$

Because $\alpha_i^0 > 0$ and the assignment counts are nonnegative, $\tilde{\alpha}_i > 0$ for every $i \in [m]$. The second equality follows from the mean formula for the Dirichlet distribution. Indeed, normalizing the kernel $\prod_{i=1}^m p_i^{\tilde{\alpha}_i - 1}$ gives a Dirichlet distribution with parameters $(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_m)$, whose kth coordinate has mean $\tilde{\alpha}_k / \sum_{i=1}^m \tilde{\alpha}_i$. The common normalizing integral and the total-concentration factor cancel between the numerator and denominator. Substituting the definition of $\tilde{\alpha}_k$ and returning to random-variable notation gives, for every $s \in [t]$ and $k \in [m]$,

$$\mathbb{P}(Z^s = k \mid Z^{-s}, H^t) = \frac{\left(\alpha_k^0 + \sum_{r \in [t]\setminus\{s\}} \mathbb{1}\{Z^r = k\}\right) \mathbb{1}\{k \in \mathcal{Z}^s\}}{\sum_{j \in \mathcal{Z}^s} \left(\alpha_j^0 + \sum_{r \in [t]\setminus\{s\}} \mathbb{1}\{Z^r = j\}\right)}. \tag{EC.5}$$

These are the full conditionals of the marginal posterior $\mathbb{P}(Z \mid H^t)$; hence each sweep leaves this marginal invariant and, at stationarity, the subsequent draw from $p \mid Z, H^t$ recovers the joint augmented posterior of (p, Z) given H^t. This completes the proof. ∎

Partially collapsed sampler. A sweep sequentially re-imputes each Z^s from (EC.5); after the sweep, a Dirichlet draw of p is taken conditional on the updated assignment vector. This differs from alternating $Z^s \mid p$ and $p \mid Z$ after every coordinate update. Algorithm 2 implements the partially collapsed update. Its Gibbs step imputes each unobserved scenario within the path-compatible set $\mathcal{Z}^t$, and $R_t \in \mathbb{Z}_+$ denotes the number of sweeps performed before the Thompson draw in period t.

Algorithm 2 Censored Thompson Sampling with Gibbs Imputation

Input: horizon T, actions $\mathcal{B}$, response maps $\{S(B, c_k)\}$, payoffs $(r_B)_{B \in \mathcal{B}}$, prior α^0, sweep schedule $R = (R_t : t \in [T])$, fixed tie-breaking; compatible sets $(\mathcal{Z}^s)$ and latent assignments Z start empty.
for $t = 1, \ldots, T$ **do**
  Run R_t sweeps, each sequentially resampling $Z^1, \ldots, Z^{t-1}$ using (EC.5). (*partially collapsed*)
  Sample $\widetilde{p}^t \sim \text{Dirichlet}(\alpha_1^0 + n_1(Z), \ldots, \alpha_m^0 + n_m(Z))$, where $n_k(Z) = \sum_{s=1}^{t-1} \mathbb{1}\{Z^s = k\}$.
  Play $B^t \in \arg\max_{B \in \mathcal{B}} \sum_{k=1}^m \widetilde{p}_k^t r_{B,k}$ and observe path S^t.
  Form $\mathcal{Z}^t \leftarrow \{k \in [m] : S(B^t, c_k) = S^t\}$ and draw $Z^t \in \mathcal{Z}^t$ with probability proportional to $\alpha_k^0 + n_k(Z)$; retain both for later sweeps.
end for

The chain can be warm-started from the latent assignments sampled in the previous period. The latent assignments should be re-imputed rather than permanently fixed after their first draw; treating an imputed scenario as observed would understate posterior uncertainty.

B.2. Linear Regret Under Pure Path-Only Thompson Sampling

Proof of Proposition 5. Recall from Section 4.2 that B_0 and B_3 are the diagnostic actions, whereas B_1 and B_2 are scenario-blind: B_1 always produces Γ_2 and B_2 always produces Γ_1. The proof exploits the corresponding value ranking in four links: interior posterior draws exclude the diagnostic actions; observations generated by the remaining actions contribute a likelihood factor equal to one; the posterior therefore stays fixed; and the suboptimal action is selected with a

constant positive probability. Specifically, the value-ranking calculation in that section gives, for every $q \in \operatorname{relint}(\Delta^2)$, $\mu_{B_1}(q) - \mu_{B_3}(q) = 9q_2$ and $\mu_{B_2}(q) - \mu_{B_3}(q) = 9q_3$; the same comparison holds for B_0 because $\mu_{B_0}(q) = \mu_{B_3}(q)$. Since $q_2, q_3 > 0$, both B_3 and B_0 are strictly dominated in value by B_1 and B_2, so neither can maximize $\mu_B(q)$ over $B \in \mathcal{B}$.

Let $f_s(\cdot) := f(\cdot \mid H^s)$ denote the path-only posterior density after period s, with f_0 the initial Dirichlet density. We prove by induction that $f_s = f_0$, $s = 0, \dots, T$, almost surely. The assertion is immediate for $s = 0$. Suppose that $f_{s-1} = f_0$ for some $s \in [T]$. Because $\alpha_k^0 > 0$ for every $k \in [3]$, the draw $\widetilde{p}^s$ belongs to $\operatorname{relint}(\Delta^2)$ almost surely. The value comparisons above then imply $\mathbb{P}_{p^\varepsilon}^{\mathrm{TS}}\big(B^s \in \{B_1, B_2\} \mid H^{s-1}\big) = 1$. For either selected action, all three scenarios are compatible with the observed path. Hence $\mathcal{Z}^s = [3]$ and, for every $q \in \Delta^2$, $P_{B^s,S^s}(q) = \sum_{k \in \mathcal{Z}^s} q_k = \sum_{k=1}^3 q_k = 1$. Bayes' rule therefore gives

$$f_s(q) = \frac{f_{s-1}(q) P_{B^s,S^s}(q)}{\int_{\Delta^2} f_{s-1}(u) P_{B^s,S^s}(u)\, du} = \frac{f_{s-1}(q)}{\int_{\Delta^2} f_{s-1}(u)\, du} = f_{s-1}(q) = f_0(q).$$

This shows that the policy selects only B_1 and B_2.

Consequently, at every period, $\widetilde{p}^t \sim \mathrm{Dirichlet}(\alpha^0)$. Let $\widetilde{p}$ denote a generic draw from this common law. Moreover, $\mu_{B_1}(\widetilde{p}^t) - \mu_{B_2}(\widetilde{p}^t) = 9(\widetilde{p}_2^t - \widetilde{p}_3^t)$. Since the Dirichlet distribution with strictly positive parameters is absolutely continuous, the tie event $\{\widetilde{p}_2^t = \widetilde{p}_3^t\}$ has probability zero. Thus the deterministic tie-breaking rule does not affect the selection probabilities, and in every period the sampler selects B_2 with probability $\mathbb{P}(\widetilde{p}_3 > \widetilde{p}_2)$. This probability is strictly positive because the event is a nonempty relatively open subset of $\operatorname{relint}(\Delta^2)$ and the Dirichlet density is positive there.

Finally, the same calculation gives $\mathcal{B}^*(p^\varepsilon) = \{B_1\}$ and $\Delta_{B_2}(p^\varepsilon) = 9/2 - 18\varepsilon > 0$. Since the policy selects only B_1 and B_2, action B_2 is the only suboptimal action it plays. Linearity of expectation therefore gives

$$\mathcal{R}^{\mathrm{TS}}(T, p^\varepsilon) = \Delta_{B_2}(p^\varepsilon)\, \mathbb{E}_{p^\varepsilon}^{\mathrm{TS}}[\tau^{\mathrm{TS}}(B_2, T)] = \Delta_{B_2}(p^\varepsilon) \sum_{t=1}^{T} \mathbb{P}_{p^\varepsilon}^{\mathrm{TS}}(B^t = B_2) = \left(\frac{9}{2} - 18\varepsilon\right) \mathbb{P}(\widetilde{p}_3 > \widetilde{p}_2)\, T.$$

For each fixed $\varepsilon \in (0, 1/4)$, both factors multiplying T are strictly positive and independent of T, so $\mathcal{R}^{\mathrm{TS}}(T, p^\varepsilon) = \Theta(T)$. The displayed coefficient is uniformly bounded above over this instance family, while any fixed $\varepsilon \in (0, 1/4)$ supplies a positive lower coefficient. Hence its worst-case regret is also $\Theta(T)$. ∎

Appendix C: Proofs for Dominated Information and Identifying Exploration

This appendix proves the results of Section 6. The first two propositions establish the consequences of avoiding diagnostic actions. Proposition 6 shows that any policy avoiding the two diagnostic actions suffers linear regret on one of two confounded environments, while Proposition 7 shows

that value optimism can be absorbed at scenario-blind actions and never reach them. The subsequent results develop the identification conditions and regret guarantee underlying Identifying-Exploration Greedy: Lemma 1 equates identification over a convex model class with injectivity on its direction space, Corollary 2 specializes this characterization to the full simplex and gives a left-inverse certificate, and Theorem 2 uses the identifying operator to bound the regret of Identifying-Exploration Greedy.

Proof of Proposition 6. Under either environment, action B_1 always produces Γ_2, whereas B_2 always produces Γ_1. Hence $P_{B,S}(p^\varepsilon) = P_{B,S}(q^\varepsilon)$ for every $B \in \{B_1, B_2\}$ and $S \in \mathcal{S}_B$. By assumption, the policy never selects B_0 or B_3. Therefore, the induction in the proof of Lemma EC.1, specifically (EC.2) restricted to the support $\{B_1, B_2\}$ of the policy, shows that $\mathbb{P}^\pi_{p^\varepsilon}(H^t = h^t) = \mathbb{P}^\pi_{q^\varepsilon}(H^t = h^t)$ for every possible history h^t and every $t \in [T]$. Since $\tau^\pi(B_i, T)$ is determined by H^T, it follows that $\mathbb{E}^\pi_{p^\varepsilon}[\tau^\pi(B_i, T)] = \mathbb{E}^\pi_{q^\varepsilon}[\tau^\pi(B_i, T)]$, $i \in \{1, 2\}$. Moreover, because the policy only plays B_1 and B_2, these two expected counts add to T under either environment.

Under p^ε, the unique optimal action is B_1, and the regret gap of B_2 is $\Delta_{B_2}(p^\varepsilon) = 9/2 - 18\varepsilon$. Under q^ε, the unique optimal action is B_2, and the regret gap of B_1 is the same. Therefore, using the equality of the count laws,

$$\mathcal{R}^\pi(T, p^\varepsilon) + \mathcal{R}^\pi(T, q^\varepsilon) = \left(\frac{9}{2} - 18\varepsilon\right)\left(\mathbb{E}^\pi_{p^\varepsilon}[\tau^\pi(B_2, T)] + \mathbb{E}^\pi_{p^\varepsilon}[\tau^\pi(B_1, T)]\right).$$

The term in parentheses equals T. At least one of the two regrets must therefore be at least half of this sum:

$$\max\{\mathcal{R}^\pi(T, p^\varepsilon), \mathcal{R}^\pi(T, q^\varepsilon)\} \geq \left(\frac{9}{4} - 9\varepsilon\right) T.$$

Because $\varepsilon \in (0, 1/4)$, the coefficient is strictly positive. ∎

Proof of Proposition 7. Fix $p \in \mathcal{M}$. We prove by induction that $B^t \in \mathcal{B}^0_{\mathrm{UCB}}$ for every $t \in [T]$, with probability one under p. At $t = 1$, no action has been sampled, so no empirical constraint is imposed and $\Theta_1 = \mathcal{M}$. The maximizing step therefore gives $B^1 \in \mathcal{B}^0_{\mathrm{UCB}}$.

Now fix $t \geq 2$ and suppose that every action selected through period $t-1$ belongs to $\mathcal{B}^0_{\mathrm{UCB}}$. Each sampled action B is then scenario-blind over $\mathcal{M}$, so every observation generated by B equals S_B with probability one. Thus, for every sampled B, every $S \in \mathcal{S}_B$, and every $q \in \mathcal{M}$,

$$\widehat{P}_{B,S}(t-1) = \mathbb{1}\{S = S_B\} = P_{B,S}(q).$$

It follows that $\|M_B q - \widehat{P}_B(t-1)\|_\infty = 0 \leq \mathrm{rad}_B(t)$ for every $q \in \mathcal{M}$ and every sampled B, and hence $\Theta_t = \mathcal{M}$. The Value-UCB optimization at period t is therefore the initial optimistic optimization defining $\mathcal{B}^0_{\mathrm{UCB}}$, so $B^t \in \mathcal{B}^0_{\mathrm{UCB}}$. This completes the induction; in particular, the empty-set fallback is never invoked.

If $\mathcal{B}^0_{\mathrm{UCB}}$ is a singleton with unique element $\widehat{B}$, the preceding result gives $\tau^{\mathrm{UCB}}(\widehat{B},T)=T$ with probability one. The action-count representation of regret then yields $\mathcal{R}^{\mathrm{UCB}}(T,p)=T\Delta_{\widehat{B}}(p)$. ∎

Proof of Lemma 1. We prove both directions by contraposition. Suppose first that $\mathcal{E}$ is not identifying over $\mathcal{M}$. Then there exist distinct $p,q\in\mathcal{M}$ such that $M_Bp=M_Bq$ for every $B\in\mathcal{E}$. The nonzero difference $v=q-p$ belongs to $\mathcal{V}_{\mathcal{M}}$, satisfies $\mathbf{1}^\top v=0$, and obeys $M_Bv=0$ for every $B\in\mathcal{E}$. Hence $M_{\mathcal{E}}v=0$, so the restriction of $M_{\mathcal{E}}$ to $\mathcal{V}_{\mathcal{M}}$ is not injective.

For the reverse contraposition, suppose that there exists a nonzero $v\in\ker(M_{\mathcal{E}})\cap\mathcal{V}_{\mathcal{M}}$. By the definition of $\mathcal{V}_{\mathcal{M}}$, write $v=\sum_{i=1}^J\xi_i(p^{(i)}-q^{(i)})$ for some $p^{(i)},q^{(i)}\in\mathcal{M}$. By swapping $p^{(i)}$ and $q^{(i)}$ whenever $\xi_i<0$ and replacing ξ_i by $|\xi_i|$, we may assume that $\xi_i\geq 0$ without changing this representation. Because $v\neq 0$, we have $\sum_{i=1}^J\xi_i>0$. Define

$$\bar{p}:=\frac{\sum_{i=1}^J\xi_ip^{(i)}}{\sum_{i=1}^J\xi_i},\qquad \bar{q}:=\frac{\sum_{i=1}^J\xi_iq^{(i)}}{\sum_{i=1}^J\xi_i}.$$

Convexity of $\mathcal{M}$ gives $\bar{p},\bar{q}\in\mathcal{M}$, while $v=(\sum_{i=1}^J\xi_i)(\bar{p}-\bar{q})$ and $v\neq 0$ imply $\bar{p}\neq\bar{q}$. Moreover, $M_Bv=0$ gives $M_B\bar{p}=M_B\bar{q}$ for every $B\in\mathcal{E}$. Thus $\mathcal{E}$ is not identifying over $\mathcal{M}$, proving the stated equivalence.

When $\mathcal{M}=\Delta^{m-1}$, $\mathcal{V}_{\mathcal{M}}=\{v\in\mathbb{R}^m:\mathbf{1}^\top v=0\}$. Because $\mathbf{1}^\top$ is the final row of $M_{\mathcal{E}}$, every vector in $\ker(M_{\mathcal{E}})$ already belongs to $\mathcal{V}_{\mathcal{M}}$. Consequently, $\ker(M_{\mathcal{E}})\cap\mathcal{V}_{\mathcal{M}}=\{0\}$ if and only if $\ker(M_{\mathcal{E}})=\{0\}$, equivalently, if and only if $M_{\mathcal{E}}$ has full column rank m. ∎

Proof of Corollary 2. By Lemma 1, $\mathcal{E}$ is identifying over Δ^{m-1} if and only if $M_{\mathcal{E}}$ has full column rank. This holds if and only if $M_{\mathcal{E}}$ admits a left inverse. Partitioning any such left inverse according to the row blocks $(M_B)_{B\in\mathcal{E}}$ and $\mathbf{1}^\top$ gives matrices $(W_B)_{B\in\mathcal{E}}$ and a vector w_0 satisfying $\sum_{B\in\mathcal{E}}W_BM_B+w_0\mathbf{1}^\top=\mathbb{I}_m$. Conversely, if these blocks satisfy the identity, their conformable block matrix is a left inverse of $M_{\mathcal{E}}$. Hence $M_{\mathcal{E}}$ has full column rank, and Lemma 1 implies that $\mathcal{E}$ is identifying over Δ^{m-1}. ∎

Proof of Theorem 2. The proof has five steps. We first count the forced-exploration plays and show that every greedy mistake requires a fixed estimation error. We then use identification to translate latent-law error into empirical path-frequency error, apply Hoeffding's inequality to the forced samples, and assemble the two regret contributions.

Step 1 (Forced-Exploration Count). Fix $B\in\mathcal{E}$. If B is never selected for forced exploration, then $\tau^{\mathrm{fe}}(B,T)=0$. Otherwise, let $t\leq T$ be its last forced-exploration period. Then $\tau^{\mathrm{fe}}(B,t-1)<g_t$, and integrality of both quantities implies $\tau^{\mathrm{fe}}(B,t)\leq g_t\leq g_T$. No subsequent period increments this counter, so $\tau^{\mathrm{fe}}(B,T)\leq g_T$. Therefore, the total forced-exploration regret is bounded by

$$g_T\sum_{B\in\mathcal{E}}\Delta_B(p).$$

Step 2 (Mistakes Require Estimation Error). Suppose that at a greedy period t, the action selected under $\widehat{p}^{t-1}$ is some $B \notin \mathcal{B}^*(p)$. Choose $B^* \in \mathcal{B}^*(p)$ attaining the inner minimum in the definition of $\rho_*(p)$. Then $(r_{B^*} - r_B)^\top \widehat{p}^{t-1} \le 0$, while $(r_{B^*} - r_B)^\top p = \Delta_B(p) \ge \Delta_{\min}(p)$. Subtracting these two inequalities gives

$$(r_{B^*} - r_B)^\top (p - \widehat{p}^{t-1}) \ge \Delta_{\min}(p).$$

Applying Cauchy–Schwarz gives $(r_{B^*} - r_B)^\top (p - \widehat{p}^{t-1}) \le \|r_{B^*} - r_B\|_2 \, \|\widehat{p}^{t-1} - p\|_2$. Its left-hand side is at least $\Delta_{\min}(p) > 0$, so $\|r_{B^*} - r_B\|_2 > 0$. Using the choice of B^* in the definition of $\rho_*(p)$, we obtain

$$\|\widehat{p}^{t-1} - p\|_2 \ge \frac{\Delta_{\min}(p)}{\|r_{B^*} - r_B\|_2} \ge \frac{\Delta_{\min}(p)}{\rho_*(p)},$$

Thus a greedy mistake can occur only if $\|\widehat{p}^{t-1} - p\|_2 \ge \Delta_{\min}(p)/\rho_*(p)$.

Step 3 (Least-Squares Stability under Identifying Feedback). We next relate this estimation error to empirical path-frequency errors. Recall that $\underline{\sigma}_{\mathcal{E}}$ is the infimum of $\|M_{\mathcal{E}} v\|_2$ over unit vectors $v \in \mathcal{V}_{\mathcal{M}}$. It therefore measures the smallest change in the stacked path probabilities generated by a unit perturbation allowed by the model class. By Lemma 1, identification implies $M_{\mathcal{E}} v \neq 0$ for every nonzero $v \in \mathcal{V}_{\mathcal{M}}$. Hence $\|M_{\mathcal{E}} v\|_2$ is strictly positive on the unit sphere of $\mathcal{V}_{\mathcal{M}}$. Since this sphere is compact and the norm is continuous, the minimum defining $\underline{\sigma}_{\mathcal{E}}$ is attained and is strictly positive. Applying its definition to $v/\|v\|_2$ for every nonzero $v \in \mathcal{V}_{\mathcal{M}}$ gives

$$\|v\|_2 \le \frac{1}{\underline{\sigma}_{\mathcal{E}}} \|M_{\mathcal{E}} v\|_2, \qquad \forall v \in \mathcal{V}_{\mathcal{M}}.$$

The case $v = 0$ is immediate. For any $q \in \mathcal{M}$, the difference $q - p$ belongs to $\mathcal{V}_{\mathcal{M}}$. Applying this bound to $v = q - p$ and using $\mathbf{1}^\top (q - p) = 0$ gives

$$\|q - p\|_2 \le \frac{1}{\underline{\sigma}_{\mathcal{E}}} \|M_{\mathcal{E}}(q - p)\|_2 = \frac{1}{\underline{\sigma}_{\mathcal{E}}} \left(\sum_{B \in \mathcal{E}} \|M_B(q - p)\|_2^2 \right)^{1/2}. \tag{EC.6}$$

By construction, $\widehat{p}^{t-1}$ is selected as a minimizer of the least-squares problem over $\mathcal{M}$, so $\widehat{p}^{t-1} \in \mathcal{M}$. The theorem assumes $p \in \mathcal{M}$; hence (EC.6) applies with $q = \widehat{p}^{t-1}$. At a greedy period t, every $B \in \mathcal{E}$ has at least $g_t \ge \beta \log t$ forced-exploration samples by time $t - 1$. Let $e_B(t-1) := \widehat{P}_B^{\mathrm{fe}}(t-1) - M_B p$ denote the forced-exploration path-frequency error for action B, namely, the difference between the empirical frequencies of the observed paths and their true probability vector $M_B p$. By optimality of the least-squares estimator and feasibility of p,

$$\sum_{B \in \mathcal{E}} \|e_B(t-1) - M_B(\widehat{p}^{t-1} - p)\|_2^2 \le \sum_{B \in \mathcal{E}} \|e_B(t-1)\|_2^2. \tag{EC.7}$$

Expanding each squared norm on the left,

$$\|e_B(t-1) - M_B(\widehat{p}^{t-1} - p)\|_2^2 = \|e_B(t-1)\|_2^2 - 2\, e_B(t-1)^\top M_B(\widehat{p}^{t-1} - p) + \|M_B(\widehat{p}^{t-1} - p)\|_2^2,$$

substituting this expansion into (EC.7), summing over $B \in \mathcal{E}$, and cancelling the common term $\sum_{B\in\mathcal{E}} \|e_B(t-1)\|_2^2$ gives the first inequality below. Applying the Cauchy–Schwarz inequality to the sum over B gives the second:

$$\begin{aligned}\sum_{B\in\mathcal{E}} \|M_B(\widehat{p}^{t-1}-p)\|_2^2 &\le 2\sum_{B\in\mathcal{E}} e_B(t-1)^\top M_B(\widehat{p}^{t-1}-p)\\ &\le 2\left(\sum_{B\in\mathcal{E}} \|e_B(t-1)\|_2^2\right)^{1/2}\left(\sum_{B\in\mathcal{E}} \|M_B(\widehat{p}^{t-1}-p)\|_2^2\right)^{1/2}.\end{aligned}$$

If the last square-root factor is zero, the desired bound is immediate; otherwise divide by it to obtain

$$\left(\sum_{B\in\mathcal{E}} \|M_B(\widehat{p}^{t-1}-p)\|_2^2\right)^{1/2} \le 2\left(\sum_{B\in\mathcal{E}} \|e_B(t-1)\|_2^2\right)^{1/2}. \tag{EC.8}$$

Combining (EC.6) and (EC.8) gives

$$\|\widehat{p}^{t-1}-p\|_2 \le \frac{2}{\underline{\sigma}_{\mathcal{E}}}\left(\sum_{B\in\mathcal{E}} \|e_B(t-1)\|_2^2\right)^{1/2}.$$

Set the proof threshold $\varepsilon_* := \underline{\sigma}_{\mathcal{E}}\Delta_{\min}(p)/(2\rho_*(p)\sqrt{|\mathcal{E}|})$. If $\|e_B(t-1)\|_2 < \varepsilon_*$ for every $B \in \mathcal{E}$, then $\|\widehat{p}^{t-1}-p\|_2 < \Delta_{\min}(p)/\rho_*(p)$. Therefore, a greedy mistake at t can occur only if $\|e_B(t-1)\|_2 \ge \varepsilon_*$ for some $B \in \mathcal{E}$. Applying the union bound over $\mathcal{E}$ gives

$$\mathbb{P}_p^{\mathrm{IE}}\big(t \text{ is a greedy period and } B^t \notin \mathcal{B}^*(p)\big) \le \sum_{B\in\mathcal{E}} \mathbb{P}_p^{\mathrm{IE}}(\|e_B(t-1)\|_2 \ge \varepsilon_*). \tag{EC.9}$$

Step 4 (Concentration of Path Frequencies). For a fixed $B \in \mathcal{E}$, the forced-exploration observations used to construct $\widehat{P}_B^{\mathrm{fe}}(t-1)$ are i.i.d. categorical samples with distribution $M_B p$. Although the path-specific counts $N_{B,S}^{\mathrm{fe}}(t)$ are random, the schedule depends only on their totals $\tau^{\mathrm{fe}}(B,t)$. Since only forced plays increment these totals, the deterministic thresholds g_t and fixed ordering determine, by induction on t, the forced-play schedule, the greedy periods, and each sample size $\tau^{\mathrm{fe}}(B,t-1)$ independently of the realized paths. This permits the direct Hoeffding bound below. For each path $S \in \mathcal{S}_B$, the coordinate $\widehat{P}_{B,S}^{\mathrm{fe}}(t-1)$ is therefore the empirical mean of $\tau^{\mathrm{fe}}(B,t-1)$ Bernoulli random variables with mean $(M_B p)_S$. Hoeffding's inequality gives, for every $s>0$,

$$\mathbb{P}_p^{\mathrm{IE}}\left(\left|\widehat{P}_{B,S}^{\mathrm{fe}}(t-1) - (M_B p)_S\right| \ge s\right) \le 2\exp(-2\tau^{\mathrm{fe}}(B,t-1)s^2).$$

Indeed, if every coordinate deviation were smaller than $\varepsilon_*/\sqrt{|\mathcal{S}_B|}$, then the squared Euclidean norm would be smaller than $|\mathcal{S}_B|\varepsilon_*^2/|\mathcal{S}_B| = \varepsilon_*^2$. Therefore, applying the union bound over $S \in \mathcal{S}_B$ and Hoeffding's inequality with $s = \varepsilon_*/\sqrt{|\mathcal{S}_B|}$ gives

$$\mathbb{P}_p^{\mathrm{IE}}\left(\|\widehat{P}_B^{\mathrm{fe}}(t-1) - M_B p\|_2 \ge \varepsilon_*\right) \le \sum_{S\in\mathcal{S}_B} \mathbb{P}_p^{\mathrm{IE}}\left(\left|\widehat{P}_{B,S}^{\mathrm{fe}}(t-1) - (M_B p)_S\right| \ge \frac{\varepsilon_*}{\sqrt{|\mathcal{S}_B|}}\right)$$

$$\leq 2|\mathcal{S}_B| \exp\left(-\frac{2\tau^{\text{fe}}(B,t-1)\varepsilon_*^2}{|\mathcal{S}_B|}\right).$$

Since $|\mathcal{S}_B| \leq s_{\max}$ and $\tau^{\text{fe}}(B,t-1) \geq \beta \log t$ at greedy periods,

$$\mathbb{P}_p^{\text{IE}}(\|\widehat{P}_B^{\text{fe}}(t-1) - M_B p\|_2 \geq \varepsilon_*) \leq 2s_{\max} t^{-2\beta\varepsilon_*^2/s_{\max}}. \tag{EC.10}$$

Substituting (EC.10) into (EC.9) and summing the $|\mathcal{E}|$ resulting terms gives

$$\mathbb{P}_p^{\text{IE}}\big(t \text{ is a greedy period and } B^t \notin \mathcal{B}^*(p)\big) \leq 2|\mathcal{E}|s_{\max} t^{-2\beta\varepsilon_*^2/s_{\max}}.$$

Step 5 (Regret Assembly). The condition imposed on β gives $2\beta\varepsilon_*^2/s_{\max} = \beta\underline{\sigma}_{\mathcal{E}}^2\Delta_{\min}(p)^2/(2s_{\max}|\mathcal{E}|\rho_*(p)^2) > 1$, so the sum over t is finite. Greedy mistakes therefore contribute at most the finite constant

$$C(p,\mathcal{E},\beta) := \max_{B\in\mathcal{B}} \Delta_B(p) \sum_{t=1}^{\infty} 2|\mathcal{E}|s_{\max} t^{-2\beta\varepsilon_*^2/s_{\max}}.$$

Adding the forced-exploration regret to the regret from greedy mistakes gives

$$\mathcal{R}^{\text{IE}}(T,p) \leq g_T \sum_{B\in\mathcal{E}} \Delta_B(p) + C(p,\mathcal{E},\beta),$$

which is the stated bound. Since $g_T = \max\{1, \lceil \beta \log T \rceil\}$ and $C(p,\mathcal{E},\beta)$ does not depend on T, the right-hand side is $O(\log T)$. ∎

Appendix D: Selecting a Low-Cost Identifying Set

Theorem 2 holds for any identifying set, but its leading forced-exploration term depends on the selected actions. Over the full simplex, Corollary 2 therefore yields an exact formulation for selecting a minimum-cost one. Directly activating each reconstruction block introduces the bilinear product $y_B W_B$; for computation, we use the equivalent bounded MILP below. Given nonnegative selection weights ℓ_B—for example, $\ell_B = \Delta_B(p)$ when the objective is to minimize the leading forced-exploration regret—let y_B indicate whether action B is selected, set $W_1^{\max} := 1$ and $W_m^{\max} := (m-1)^{(m-1)/2}$ for $m \geq 2$, and solve

$$\begin{aligned} \min_{y\in\{0,1\}^{|\mathcal{B}|}, W, w_0} \quad & \sum_{B\in\mathcal{B}} \ell_B y_B \\ \text{s.t.} \quad & \sum_{B\in\mathcal{B}} W_B M_B + w_0 \mathbf{1}^\top = \mathbb{I}_m, \\ & -W_m^{\max} y_B \leq (W_B)_{i,S} \leq W_m^{\max} y_B, \qquad \forall B\in\mathcal{B},\ i\in[m],\ S\in\mathcal{S}_B. \end{aligned} \tag{EC.11}$$

The coefficient bounds force $W_B = 0$ whenever $y_B = 0$, so every feasible solution supplies a left-inverse certificate supported on $\{B : y_B = 1\}$, which is identifying by Corollary 2. Conversely, let

$\mathcal{E}$ be identifying and select m linearly independent rows of the binary matrix $M_{\mathcal{E}}$. The resulting square matrix has a nonzero integer determinant, whose absolute value is at least one. The cofactor formula and Hadamard's inequality therefore bound every entry of its inverse by $W_m^{\max}$ in absolute value. Padding this inverse with zero columns produces a feasible certificate supported on $\mathcal{E}$. Hence (EC.11) selects a minimum-cost identifying set.

For restricted model classes, the implementation applies the analogous rank certificate to a basis of $\mathcal{V}_{\mathcal{M}}$, omitting the normalization row and reducing the identity dimension to $\dim(\mathcal{V}_{\mathcal{M}})$. Each returned set is independently verified using the injectivity condition in Lemma 1.

Appendix E: Proof of the Finite Exponential-Cone Reformulation

This section proves Theorem 3 through the three reductions introduced in Section 7: covering the alternative set by finitely many challenger cells, replacing each strict cell by its closed relaxation, and dualizing the resulting cellwise information problem. We state the required results in this order, combine them to prove the theorem, and then provide their proofs.

Setup and notation. We retain the full-simplex specialization and abbreviations $\mathcal{Q}(p) := \mathcal{Q}_{\Delta^{m-1}}(p)$ and $\kappa^*(p) := \kappa^*_{\Delta^{m-1}}(p)$ from Section 7. We assume $\mathcal{B} \setminus \mathcal{B}^*(p) \neq \emptyset$; otherwise $\mathcal{Q}(p) = \emptyset$ and $\kappa^*(p) = 0$.

For each challenger $B' \notin \mathcal{B}^*(p)$, define the strict challenger cell

$$\mathcal{Q}_{B'}(p) := \left\{ q \in \Delta^{m-1} : \begin{array}{ll} M_{B^*} q = P_{B^*}(p), & \forall B^* \in \mathcal{B}^*(p), \\ (r_{B'} - r_{B^*})^\top q > 0, & \forall B^* \in \mathcal{B}^*(p), \\ (M_B q)_S > 0, & \forall B \in \mathcal{B},\ \forall S \in \mathcal{S}_B \text{ with } P_{B,S}(p) > 0 \end{array} \right\}.$$

The cell matches the path laws of the p-optimal actions, makes B' strictly preferable to each of them, and enforces the support condition for finite path-feedback information. A decision-changing alternative may have several optimal challengers, but any one of them indexes it. The first reduction is therefore the following finite cover.

Proposition EC.1 **(Finite Cover by Challenger Cells)**. *Under Assumption 1,*

$$\mathcal{Q}(p) = \bigcup_{B' \in \mathcal{B} \setminus \mathcal{B}^*(p)} \mathcal{Q}_{B'}(p).$$

Each cell is convex and described by finitely many affine equalities and linear inequalities, with strictness only in the challenger and support conditions.

We henceforth retain only the nonempty cells and write $\mathcal{C}(p) := \{B' \in \mathcal{B} \setminus \mathcal{B}^*(p) : \mathcal{Q}_{B'}(p) \neq \emptyset\}$. The strict challenger and support conditions mean that these cells need not be closed. For each $B' \in \mathcal{C}(p)$, define the closed relaxation

$$\operatorname{cl} \mathcal{Q}_{B'}(p) := \left\{ q \in \Delta^{m-1} : M_{B^*} q = P_{B^*}(p),\ (r_{B'} - r_{B^*})^\top q \geq 0,\ \forall B^* \in \mathcal{B}^*(p) \right\}.$$

This relaxation weakens the challenger inequalities and drops the strict support conditions. For $x \in \mathbb{R}_+^{\mathcal{B}\setminus\mathcal{B}^*(p)}$, let

$$\psi_{B'}(x) := \min_{q \in \mathrm{cl}\, \mathcal{Q}_{B'}(p)} \sum_{B \notin \mathcal{B}^*(p)} x_B D(P_B(p) \| M_B q).$$

Thus $\psi_{B'}(x)$ is the minimum information generated by allocation x over the closed cell. The minimum exists because the cell is compact and the objective is lower semicontinuous. We use the closed convention $0 \cdot D(P_B(p)\|M_B q) = 0$, even when the divergence is infinite. Moreover, $\psi_{B'}$ is concave because it is the pointwise infimum of functions that are linear in x. The next lemma shows that passing to the compact relaxation does not change the worst-case information, which is needed for both attainment and duality.

LEMMA EC.2 **(Closure of the Cellwise Infimum)**. *For every $B' \in \mathcal{C}(p)$ and $x \in \mathbb{R}_+^{\mathcal{B}\setminus\mathcal{B}^*(p)}$,*

$$\inf_{q \in \mathcal{Q}_{B'}(p)} \sum_{B \notin \mathcal{B}^*(p)} x_B D(P_B(p) \| M_B q) = \psi_{B'}(x).$$

The cover and closure result reduce the original continuum of information constraints to one concave superlevel constraint per nonempty cell.

PROPOSITION EC.2 **(Cellwise Form of the Lower-Bound Problem)**. *Under Assumption 1,*

$$\kappa^*(p) = \inf_{x \in \mathbb{R}_+^{\mathcal{B}\setminus\mathcal{B}^*(p)}} \left\{ \sum_{B \notin \mathcal{B}^*(p)} x_B \Delta_B(p) : \psi_{B'}(x) \geq 1 \quad \forall B' \in \mathcal{C}(p) \right\},$$

with $\kappa^*(p) = 0$ when $\mathcal{C}(p) = \emptyset$.

It remains to dualize each $\psi_{B'}$. This step requires a constraint qualification for the inner cellwise problem.

ASSUMPTION EC.1 **(Relative Slater Condition)**. *For each $B' \in \mathcal{C}(p)$, there exists a point $\tilde{q} \in \mathrm{relint}\{q \in \Delta^{m-1} : M_{B^*} q = P_{B^*}(p),\ \forall B^* \in \mathcal{B}^*(p)\}$ such that $(r_{B'} - r_{B^*})^\top \tilde{q} > 0$ for every $B^* \in \mathcal{B}^*(p)$, and $(M_B \tilde{q})_S > 0$ for every $B \in \mathcal{B}$ and every $S \in \mathcal{S}_B$ with $P_{B,S}(p) > 0$.*

This is the relative Slater qualification used below to establish strong duality and dual attainment. For each $B' \in \mathcal{C}(p)$, introduce $\lambda_{B',B,S} \geq 0$ for $B \notin \mathcal{B}^*(p)$ and $S \in \mathcal{S}_B$, $\eta_{B',B^*,S} \in \mathbb{R}$ for $B^* \in \mathcal{B}^*(p)$ and $S \in \mathcal{S}_{B^*}$, and $\nu_{B',B^*} \geq 0$ for $B^* \in \mathcal{B}^*(p)$. Define

$$\Phi_{B'}(x, \lambda_{B'}, \eta_{B'}, \nu_{B'}) := \sum_{B \notin \mathcal{B}^*(p)} \sum_{\substack{S \in \mathcal{S}_B: \\ P_{B,S}(p) > 0}} P_{B,S}(p) \left[x_B \log \frac{\lambda_{B',B,S}}{x_B} + x_B \right] + \sum_{B^* \in \mathcal{B}^*(p)} \sum_{S \in \mathcal{S}_{B^*}} P_{B^*,S}(p) \eta_{B',B^*,S}, \tag{EC.12}$$

where the logarithmic perspective uses its closed extension at $x_B = 0$. The dual-feasibility constraints are

$$\sum_{B \notin \mathcal{B}^*(p)} \lambda_{B',B,S(B,c_k)} + \sum_{B^* \in \mathcal{B}^*(p)} \left[\eta_{B',B^*,S(B^*,c_k)} + \nu_{B',B^*}(r_{B'} - r_{B^*})_k\right] \leq 0, \qquad k \in [m]. \tag{EC.13}$$

Variables $\lambda_{B',B,S}$ with $P_{B,S}(p) = 0$ remain in (EC.13) but do not enter (EC.12).

Proposition EC.3 **(Cellwise Dual Representation)**. *Fix $B' \in \mathcal{C}(p)$. Under Assumption EC.1, for every $x \in \mathbb{R}_+^{\mathcal{B}\setminus\mathcal{B}^*(p)}$,*

$$\psi_{B'}(x) = \max_{\lambda_{B'},\eta_{B'},\nu_{B'}} \Phi_{B'}(x, \lambda_{B'}, \eta_{B'}, \nu_{B'})$$

subject to (EC.13)*, $\lambda_{B',B,S} \geq 0$, $\nu_{B',B^*} \geq 0$, and $\eta_{B',B^*,S} \in \mathbb{R}$.*

Proof of Theorem 3. By Proposition EC.2, the LBP has the cellwise constraints $\psi_{B'}(x) \geq 1$ for all $B' \in \mathcal{C}(p)$. Fix $B' \in \mathcal{C}(p)$. By strong duality and dual attainment in Proposition EC.3, this constraint holds if and only if there exist dual-feasible $(\lambda_{B'}, \eta_{B'}, \nu_{B'})$ satisfying $\Phi_{B'}(x, \lambda_{B'}, \eta_{B'}, \nu_{B'}) \geq 1$. Thus duality eliminates the inner minimization over q, while attainment replaces the dual maximization by the existence of an above-threshold certificate.

It remains to represent this certificate conically. The only nonlinear terms in $\Phi_{B'}$ are $x_B \log(\lambda_{B',B,S}/x_B)$. For $x > 0$, $(u, x, \lambda) \in \mathcal{K}_{\exp}$ is equivalent to $u \leq x \log(\lambda/x)$; at $x = 0$, its closure gives $u \leq 0$ and $\lambda \geq 0$. Suppose first that the B'-block of (3) is feasible. Its unit-threshold and exponential-cone constraints give

$$1 \overset{(a)}{\leq} \sum_{B \notin \mathcal{B}^*(p)} \sum_{\substack{S \in \mathcal{S}_B: \\ P_{B,S}(p) > 0}} P_{B,S}(p)(u_{B',B,S} + x_B) + \sum_{B^* \in \mathcal{B}^*(p)} \sum_{S \in \mathcal{S}_{B^*}} P_{B^*,S}(p)\eta_{B',B^*,S} \overset{(b)}{\leq} \Phi_{B'}(x, \lambda_{B'}, \eta_{B'}, \nu_{B'}).$$

Here (a) is the block's unit-threshold constraint, while (b) follows by weighting each cone inequality $u_{B',B,S} \leq x_B \log(\lambda_{B',B,S}/x_B)$ by $P_{B,S}(p)$ and summing. The remaining block constraints are precisely the dual-feasibility constraints. Hence the block provides the required dual certificate.

Conversely, take a dual-feasible certificate with $\Phi_{B'} \geq 1$. We lift this certificate to the conic program. Finiteness of $\Phi_{B'}$ gives $\lambda_{B',B,S} > 0$ whenever $x_B > 0$ and $P_{B,S}(p) > 0$. Set $u_{B',B,S} := x_B \log(\lambda_{B',B,S}/x_B)$ when $x_B > 0$ and $u_{B',B,S} := 0$ when $x_B = 0$. The cone constraints then hold with equality or by their boundary convention, respectively. Dual feasibility supplies the linear block constraints, and the unit-threshold constraint is exactly $\Phi_{B'} \geq 1$.

Because $\mathcal{C}(p)$ is finite and the certificate variables are indexed by cell, these constructions can be made simultaneously. The cellwise and conic programs therefore have the same feasible vectors x and the same objective, so their optimal values coincide as extended reals. Finally, the remaining

constraints and objective are linear and $\mathcal{K}_{\exp}$ is convex, making (3) a finite convex conic program. ■

We now prove the auxiliary results in dependency order.

Proof of Proposition EC.1. For $\subseteq$, fix $q \in \mathcal{Q}(p)$ and verify the three defining conditions. First, the information inequality gives $I_B(p,q) \geq 0$, with equality exactly when $M_B q = P_B(p)$ (Cover and Thomas 2006, Theorem 2.6.3); hence the zero-information condition gives $M_{B^*} q = P_{B^*}(p)$ for every $B^* \in \mathcal{B}^*(p)$. Second, disjointness makes any $B' \in \mathcal{B}^*(q)$ lie outside $\mathcal{B}^*(p)$ and every $B^* \in \mathcal{B}^*(p)$ strictly suboptimal under q; hence $r_{B'}^\top q > r_{B^*}^\top q$. Third, because each response set is finite, absolute continuity gives $(M_B q)_S > 0$ whenever $P_{B,S}(p) > 0$. Therefore $q \in \mathcal{Q}_{B'}(p)$.

For $\supseteq$, fix $q \in \mathcal{Q}_{B'}(p)$ for some $B' \notin \mathcal{B}^*(p)$. Its observation equalities give zero information for every $B^* \in \mathcal{B}^*(p)$. Because each response set is finite, its support inequalities give $P_B(p) \ll P_B(q)$ for every B. Its strict challenger inequalities make the optimal sets disjoint. Assumption 1 then rules out zero information for every action, so all the conditions defining $\mathcal{Q}(p)$ hold. Hence $q \in \mathcal{Q}(p)$. This proves the union. Each cell is convex as the intersection of the simplex with finitely many affine equalities and linear strict inequalities. ■

Proof of Lemma EC.2. Write $F_x(q) := \sum_{B \notin \mathcal{B}^*(p)} x_B D(P_B(p) \| M_B q)$ under the closed convention above. Since $\mathcal{Q}_{B'}(p) \subseteq \operatorname{cl} \mathcal{Q}_{B'}(p)$, $\psi_{B'}(x) \leq \inf_{q \in \mathcal{Q}_{B'}(p)} F_x(q)$. If $\psi_{B'}(x) = +\infty$, every point of the closed cell, and hence of the strict cell, has infinite value, so equality follows.

Suppose instead that $\psi_{B'}(x) < +\infty$, and let $\bar{q} \in \operatorname{cl} \mathcal{Q}_{B'}(p)$ attain this minimum. Choose $\tilde{q} \in \mathcal{Q}_{B'}(p)$, which exists because $B' \in \mathcal{C}(p)$, and set $q_\theta := (1-\theta)\bar{q} + \theta \tilde{q}$ for $\theta \in (0,1)$. For every $B^* \in \mathcal{B}^*(p)$, linearity gives $M_{B^*} q_\theta = (1-\theta) P_{B^*}(p) + \theta P_{B^*}(p) = P_{B^*}(p)$ and $(r_{B'} - r_{B^*})^\top q_\theta = (1-\theta)(r_{B'} - r_{B^*})^\top \bar{q} + \theta (r_{B'} - r_{B^*})^\top \tilde{q} > 0$. Likewise, for every $B \in \mathcal{B}$ and $S \in \mathcal{S}_B$ with $P_{B,S}(p) > 0$, $(M_B q_\theta)_S = (1-\theta)(M_B \bar{q})_S + \theta (M_B \tilde{q})_S > 0$. Together with convexity of the simplex, these relations give $q_\theta \in \mathcal{Q}_{B'}(p)$; moreover, $q_\theta \to \bar{q}$ as $\theta \downarrow 0$.

Because $F_x(\bar{q}) < +\infty$, every action with $x_B > 0$ satisfies $(M_B \bar{q})_S > 0$ whenever $P_{B,S}(p) > 0$, so its divergence is continuous at $\bar{q}$. Terms with $x_B = 0$ vanish by convention, and the action and response sets are finite. Hence $F_x(q_\theta) \to F_x(\bar{q}) = \psi_{B'}(x)$. Since every q_θ belongs to the strict cell, $\inf_{q \in \mathcal{Q}_{B'}(p)} F_x(q) \leq \psi_{B'}(x)$. Combining the two inequalities proves the result. ■

Proof of Proposition EC.2. Fix $x \in \mathbb{R}_+^{\mathcal{B} \setminus \mathcal{B}^*(p)}$ and write $F_x(q) := \sum_{B \notin \mathcal{B}^*(p)} x_B D(P_B(p) \| M_B q)$. By (1), x is feasible for the original LBP precisely when $F_x(q) \geq 1$ for every $q \in \mathcal{Q}(p)$. The finite cover in Proposition EC.1, the definition of $\mathcal{C}(p)$, and Lemma EC.2 give

$$F_x(q) \geq 1 \quad \forall q \in \mathcal{Q}(p) \iff \inf_{q \in \mathcal{Q}_{B'}(p)} F_x(q) \geq 1 \quad \forall B' \in \mathcal{C}(p) \iff \psi_{B'}(x) \geq 1 \quad \forall B' \in \mathcal{C}(p).$$

Thus the original and cellwise programs have the same feasible allocations; their objectives are identical, proving the stated formulation. If $\mathcal{C}(p) = \emptyset$, Proposition EC.1 gives $\mathcal{Q}(p) = \emptyset$, so the LBP has no information constraints and $x = 0$ yields $\kappa^*(p) = 0$. ■

Proof of Proposition EC.3. Fix $B' \in \mathcal{C}(p)$ and $x \in \mathbb{R}_+^{\mathcal{B}\setminus\mathcal{B}^*(p)}$. To separate the logarithmic divergence terms from the latent law q, introduce the auxiliary path probabilities $z_{B,S} = (M_B q)_S$ for $B \notin \mathcal{B}^*(p)$ and $S \in \mathcal{S}_B$. By the definition of $\operatorname{cl} \mathcal{Q}_{B'}(p)$, $\psi_{B'}(x)$ is the minimum of

$$\sum_{B\notin\mathcal{B}^*(p)} \sum_{S\in\mathcal{S}_B} x_B P_{B,S}(p) \log \frac{P_{B,S}(p)}{z_{B,S}}$$

subject to (i) $z_{B,S} = (M_B q)_S$ for $B \notin \mathcal{B}^*(p)$ and $S \in \mathcal{S}_B$; (ii) $M_{B^*} q = P_{B^*}(p)$ for $B^* \in \mathcal{B}^*(p)$; (iii) $(r_{B'} - r_{B^*})^\top q \geq 0$ for $B^* \in \mathcal{B}^*(p)$; and (iv) $q \geq 0$. No separate normalization constraint is needed: summing the rows in (ii) gives $\mathbf{1}^\top q = 1$. Associate $\lambda_{B',B,S}$, $\eta_{B',B^*,S}$, and $\nu_{B',B^*} \geq 0$ with (i)–(iii), respectively, written as $z_{B,S} - (M_B q)_S = 0$, $P_{B^*,S}(p) - (M_{B^*} q)_S = 0$, and $-(r_{B'} - r_{B^*})^\top q \leq 0$; constraint (iv) remains in the domain. Since $(M_B)_{S,k} = \mathbb{1}\{S(B, c_k) = S\}$, define the resulting coefficient

$$a_{B',k} := \sum_{B\notin\mathcal{B}^*(p)} \lambda_{B',B,S(B,c_k)} + \sum_{B^*\in\mathcal{B}^*(p)} \left[\eta_{B',B^*,S(B^*,c_k)} + \nu_{B',B^*}(r_{B'} - r_{B^*})_k\right], \qquad k \in [m].$$

Collecting the q-terms then gives the compact Lagrangian

$$L(q,z) = \sum_{B\notin\mathcal{B}^*(p)} \sum_{S\in\mathcal{S}_B} x_B P_{B,S}(p) \log \frac{P_{B,S}(p)}{z_{B,S}} + \lambda_{B',B,S} z_{B,S} + \sum_{B^*\in\mathcal{B}^*(p)} \sum_{S\in\mathcal{S}_{B^*}} P_{B^*,S}(p)\eta_{B',B^*,S} - \sum_{k=1}^{m} a_{B',k} q_k.$$

The minimization over z is coordinatewise. Fix $B \notin \mathcal{B}^*(p)$ and $S \in \mathcal{S}_B$. If $x_B P_{B,S}(p) > 0$, differentiation gives the minimizer $z = x_B P_{B,S}(p)/\lambda_{B',B,S}$ when $\lambda_{B',B,S} > 0$; if $x_B P_{B,S}(p) = 0$, the closed perspective has value zero for every $\lambda_{B',B,S} \geq 0$. Hence

$$\inf_{z\geq 0} \left\{ x_B P_{B,S}(p) \log \frac{P_{B,S}(p)}{z} + \lambda_{B',B,S} z \right\} = P_{B,S}(p) \left[x_B \log \frac{\lambda_{B',B,S}}{x_B} + x_B \right]$$

for $\lambda_{B',B,S} \geq 0$. When $x_B P_{B,S}(p) > 0$ and $\lambda_{B',B,S} = 0$, both sides equal $-\infty$; negative multipliers also give value $-\infty$ and may be excluded. Similarly, minimizing the last term over (iv) gives, for each k,

$$\inf_{q_k\geq 0} \{-a_{B',k} q_k\} = \begin{cases} 0, & a_{B',k} \leq 0, \\ -\infty, & a_{B',k} > 0. \end{cases}$$

Combining the two coordinatewise minimizations gives

$$\inf_{q\geq 0,\, z\geq 0} L(q,z) = \begin{cases} \Phi_{B'}(x, \lambda_{B'}, \eta_{B'}, \nu_{B'}), & \lambda_{B'} \geq 0,\ \nu_{B'} \geq 0,\ a_{B',k} \leq 0\ \forall k, \\ -\infty, & \text{otherwise.} \end{cases}$$

The conditions $a_{B',k} \leq 0$ are precisely (EC.13). Hence maximizing this infimum over the multipliers gives the dual problem stated in the proposition. Finally, Assumption EC.1 provides a relative-interior feasible point in the effective domain of the primal objective. Strong duality then gives equality of the primal and dual values (Rockafellar 1970, Theorem 28.2), and the existence of a Kuhn–Tucker vector gives dual attainment (Rockafellar 1970, Corollary 28.4.1). ∎

E.1. Detecting Challenger Cells

The conic program needs a block only for challengers in $\mathcal{C}(p)$. Replacing every strict inequality by a common margin δ yields the following linear test. For each $B' \in \mathcal{B} \setminus \mathcal{B}^*(p)$, consider

$$\begin{aligned}
\max_{q,\delta} \quad & \delta \\
\text{s.t.} \quad & q \in \Delta^{m-1}, \\
& M_{B^*} q = P_{B^*}(p), && \forall B^* \in \mathcal{B}^*(p), \\
& (r_{B'} - r_{B^*})^\top q \geq \delta, && \forall B^* \in \mathcal{B}^*(p), \\
& (M_B q)_S \geq \delta, && \forall B \in \mathcal{B},\ \forall S \in \mathcal{S}_B \text{ with } P_{B,S}(p) > 0.
\end{aligned}$$

Its optimal value is positive if and only if $\mathcal{Q}_{B'}(p) \neq \emptyset$. Indeed, a feasible point with $\delta > 0$ belongs to the strict cell; conversely, any point in the strict cell admits a positive δ equal to the minimum of its finitely many strict margins. The test therefore identifies $\mathcal{C}(p)$. A positive optimizer also satisfies the strict challenger and support requirements of Assumption EC.1; relative-interior membership in the affine set fixed by the optimal-action observation equalities must still be verified separately.

E.2. Two-Path Check of the Reformulation

This two-path instance checks the two least transparent steps of the reformulation: closing a strict challenger cell and constructing its dual certificate.

EXAMPLE EC.1. Take $\mathcal{M} = \Delta^1$ and consider two parallel source–sink paths $\Gamma_i := \{a_i\}$, $i = 1, 2$, with interdiction budget $K = 1$. Their costs under two scenarios and the latent distribution are

$$\begin{array}{c|cc} & c_1 & c_2 \\ \hline \Gamma_1 & 2 & 1 \\ \Gamma_2 & 1 & 3 \end{array} \qquad p = (1/3,\, 2/3)\,.$$

Write $B_0 := \emptyset$ and $B_i := \{a_i\}$ for $i = 1, 2$. The expected values of (B_1, B_2, B_0) under p are $(7/3, 4/3, 1)$, so $\mathcal{B}^*(p) = \{B_1\}$, $\Delta_{B_2}(p) = 1$, and $\Delta_{B_0}(p) = 4/3$.

For $q(y) := (y, 1-y)$, the expected values are $\mu_{B_1}(q(y)) = 3 - 2y$, $\mu_{B_2}(q(y)) = 1 + y$, and $\mu_{B_0}(q(y)) = 1$. Thus B_2 displaces the action optimal under p exactly when $y > 2/3$. Actions B_1 and B_2 each leave one path available and carry no information about y. Under B_0, ordering paths as (Γ_2, Γ_1) gives $M_{B_0} p = (1/3, 2/3)^\top$ and $M_{B_0} q(y) = (y, 1-y)^\top$. The support condition excludes $y = 1$, so

$$\mathcal{Q}(p) = \mathcal{Q}_{B_2}(p) = \{q(y) : y \in (2/3, 1)\}.$$

Only B_0 distinguishes these alternatives. Its information number is $D((1/3, 2/3) \,\|\, (y, 1-y))$, whose derivative $(y - 1/3)/(y(1-y))$ is positive on the challenger cell. The least informative alternatives therefore approach the excluded boundary $y \downarrow 2/3$. Denote the boundary information by

$$D_{\text{two}} := D\left((1/3, 2/3) \,\|\, (2/3, 1/3)\right) = \frac{1}{3} \log 2 \approx 0.231049.$$

The LBP reduces to minimizing $x_{B_2} + (4/3)x_{B_0}$ subject to $x_{B_0} D_{\text{two}} \geq 1$, and hence

$$x^*_{B_2} = 0, \qquad x^*_{B_0} = \frac{1}{D_{\text{two}}}, \qquad \kappa^*(p) = \frac{4}{3D_{\text{two}}} = \frac{4}{\log 2} \approx 5.771.$$

Thus information must be bought through the suboptimal diagnostic action B_0, and the cellwise minimum is attained only on the closure of the strict cell.

Conic verification. Only B_2 has a nonempty challenger cell. Suppress its common index $B' = B_2$, write $x := x_{B_0}$, and set

$$x_{B_2} = 0, \qquad \lambda_{B_0,\Gamma_2} = \tfrac{x}{2}, \qquad \lambda_{B_0,\Gamma_1} = 2x, \qquad \lambda_{B_2,\Gamma_1} = 0, \qquad \eta_{B_1,\Gamma_2} = -x, \qquad \nu_{B_1} = \tfrac{x}{2}.$$

Using the response map above and $r_{B_2} - r_{B_1} = (1,-2)^\top$, both dual-feasibility inequalities hold with equality. Setting each $u_{B,S}$ to its logarithmic-perspective value gives the certificate value

$$x\left[\tfrac{1}{3}\log\tfrac{1}{2} + \tfrac{2}{3}\log 2 + 1\right] - x = x \cdot \tfrac{1}{3}\log 2 = xD_{\text{two}},$$

so the conic constraint is $xD_{\text{two}} \geq 1$ and recovers the direct solution. At the optimum, $\lambda_{B_0,S} = x^*_{B_0} P_{B_0,S}(p)/(M_{B_0} q(2/3))_S$ for $S \in \{\Gamma_1, \Gamma_2\}$: the dual multipliers equal the exploration weight times the likelihood ratios against the boundary challenger.

Appendix F: Numerical Specifications and Additional Results

This section of the Electronic Companion follows the order in which the numerical material first appears in Section 8. It specifies the Value-UCB rule, records the IE-Greedy estimator variant and implementation details, reports the exploration-multiplier sensitivity, defines the lower-bound calibration instance, and closes with the layered-network study. All results use the experimental protocol of Section 8: shared scenario streams across policies within each replication, mean cumulative pseudo-regret as the performance measure, and the symmetric prior $\alpha^0 = (1, \ldots, 1)$ for the Thompson sampling policies.

F.1. Path-Only Value-UCB Specification

The UCB curves in Section 8 use a value-optimistic policy adapted to path-only feedback. It instantiates the usual confidence-set optimism principle for structured and linear bandits (Dani et al. 2008, Rusmevichientong and Tsitsiklis 2010, Abbasi-Yadkori et al. 2011), except that the confidence object is the distribution of the observed path: after playing B the leader sees one of the paths in $\mathcal{S}_B$, not the latent scenario, and M_B maps a latent law q to the induced path law $M_B q$.

Fix $\delta \in (0,1)$. Let $N_{B,S}(s)$ count the times B has produced path S through period s, and define the play count $\tau^{\text{UCB}}(B,s) := \sum_{S \in \mathcal{S}_B} N_{B,S}(s)$. When $\tau^{\text{UCB}}(B,s) > 0$, set $\widehat{P}_{B,S}(s) :=$

Algorithm 3 Path-Only Value-UCB Used in the Numerical Experiments

Input: horizon T, $\mathcal{B}, \mathcal{M}, (M_B)_{B\in\mathcal{B}}, (r_B)_{B\in\mathcal{B}}, \delta, T_{\mathrm{ref}}, \mathcal{W}_0$; $N_{B,S}(0) \leftarrow 0$ and $U_B(1) \leftarrow \sup_{q\in\mathcal{M}} r_B^\top q$ for all B, S.
for $t = 1, \dots, T$ **do**
  if $t = 1$ or $(t-1) \equiv 0 \pmod{T_{\mathrm{ref}}}$ **then**
    Compute $\widehat{P}_B(t-1)$ and $\mathrm{rad}_B(t)$ for each sampled B. (*confidence refresh*)
    Set ω_t to the first $\omega \in \mathcal{W}_0$ with $\Theta_t(\omega) \neq \emptyset$; if none exists, set $\omega_t \leftarrow 2\max \mathcal{W}_0$ and double until $\Theta_t(\omega_t) \neq \emptyset$; record $\mathrm{ext}_t \leftarrow \mathbb{1}\{\omega_t \notin \mathcal{W}_0\}$.
    Set $U_B(t) \leftarrow \sup_{q\in\Theta_t(\omega_t)} r_B^\top q$ for every $B \in \mathcal{B}$. (*value optimism*)
  else
    Set $U_B(t) \leftarrow U_B(t-1)$ for every B. (*reuse*)
  end if
  Play $B^t \in \arg\max_{B\in\mathcal{B}} U_B(t)$, observe path S^t, and increment N_{B^t,S^t}.
end for

Counts not named in a step are carried forward unchanged. ext_t is a diagnostic only and does not affect the play.

$N_{B,S}(s)/\tau^{\mathrm{UCB}}(B,s)$. The decision at t uses only counts through $t-1$; for sampled actions, $\mathrm{rad}_B(t) := \sqrt{2\log(t|\mathcal{B}||\mathcal{S}_B|/\delta)/\tau^{\mathrm{UCB}}(B,t-1)}$. For an inflation factor $\omega \geq 1$, the confidence set is

$$\Theta_t(\omega) := \left\{ q \in \mathcal{M} : \|M_B q - \widehat{P}_B(t-1)\|_\infty \leq \omega\, \mathrm{rad}_B(t) \quad \text{for every } B \text{ with } \tau^{\mathrm{UCB}}(B,t-1) > 0 \right\}.$$

Unsampled actions impose no constraint. In the implementation, the radius constant is 2, $\delta = 0.05$, optimistic values are refreshed every $T_{\mathrm{ref}} = 25$ periods, and $\mathcal{W}_0 = \{2^j : j = 0, \dots, 8\}$ is the base inflation grid. At each refresh the code takes the first $\omega \in \mathcal{W}_0$ with $\Theta_t(\omega)$ nonempty, doubling the largest inflation until feasibility is restored if none is, and records whether the base grid was extended. This terminates whenever $\mathcal{M}$ is nonempty, since the residuals are bounded and every sampled action has a positive radius; no reported run required an inflation outside the base grid. The policy uses an exact argmax without an additional forced-exploration step. The numerical implementations use their recorded solver tolerances and experiment-specific tie rules.

The rule stays value-driven: a diagnostic action is sampled only when it is optimistic for some law in $\Theta_t(\omega)$, so dominated-information instances can leave such actions unplayed while self-revealing instances learn from paths generated by valuable actions. This specification makes the generic radii of Section 6.2 concrete and adds radius inflation and periodic index reuse; under the hypothesis of Proposition 7 these choices are immaterial, since every refresh gives $\Theta_t(\omega) = \mathcal{M}$ for every $\omega \geq 1$ and reuse preserves the initial indices, so the absorption argument applies unchanged.

F.2. The IE-Greedy Relaxed Variant

The all-arc benchmarks of Section 8.5 plot a fifth policy, *IE-Greedy Relaxed*. It uses the same forced-exploration rule and certified identifying set as IE-Greedy, but replaces the constrained least-squares projection onto $\mathcal{M}$ by an affine least-squares estimate followed by projection onto the simplex, isolating the cost of enforcing the model-class constraint exactly. Across every benchmark of Section 8 the two policies follow almost the same regret path, as expected from their common

identifying set and forced-exploration schedule; the variant changes estimator cost rather than regret. We therefore report it for completeness but do not discuss it separately.

F.3. Implementation Details

All algorithms were coded in Julia v1.12.4 (Bezanson et al. 2017). We use JuMP v1.30.1, MathOptInterface v1.51.1, Gurobi.jl v1.9.2 with native Gurobi v13.0.0, MosekTools v0.15.10, and MOSEK v11.2.0. Gurobi is used for the all-arc graph oracle and the Value-UCB linear programs; MOSEK is used for the lower-bound and conic calibration routines and for the identifying-set MILPs in the controlled and layered-network suites. All experiments were performed on a macOS 26.4.1 machine with an Apple M5 processor and 16 GB of RAM. Julia was launched with 10 threads and BLAS was restricted to one thread. In the benchmark tables, policy time is the mean wall-clock time in seconds for one invocation of the online simulation, curve summarization, and summary construction for an accepted instance with 30 replications; instance construction, acceptance screening, oracle-equivalence checks, identifying-set selection, and checkpoint serialization fall outside the timed block. There is no warm-up pass, so first-use compilation may enter a measurement; these times are descriptive rather than controlled microbenchmarks.

F.4. Exploration Multiplier Sensitivity

Figure EC.1 varies the IE-Greedy multiplier β on the three-path instance, isolating the finite-horizon trade-off between diagnostic sampling and forced-exploration cost.

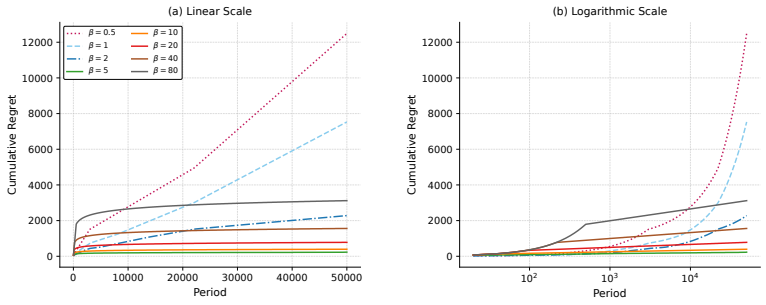

Figure EC.1 Sensitivity of IE-Greedy to the exploration multiplier β on the three-path instance with $\varepsilon = 0.1$, $n_{\text{rep}} = 300$ replications, and $T = 50{,}000$. The grid is $\{0.5, 1, 2, 5, 10, 20, 40, 80\}$. Small multipliers under-explore and can leave the policy in the dominated-information trap; large multipliers are robust but pay unnecessary forced-exploration cost.

The best final pseudo-regret occurs at $\beta = 5$, with final regret 231.7. Smaller multipliers under-explore: at $\beta = 0.5$ the final regret is 12,490.5, an order of magnitude worse. Larger multipliers are safe but increasingly conservative, with final regret rising from 394.5 at $\beta = 10$ to 3,117.6 at $\beta = 80$. The main message is that identifying exploration removes the linear-regret failure, but

its finite-horizon cost hinges on a multiplier that is not obvious ex ante. This is a finite-horizon observation about tuning, not a statement about the sufficient threshold of Theorem 2, which is instance dependent and generally conservative.

F.5. Lower-Bound Calibration Instance

The calibration in Section 8.3 needs a finite lower-bound constant and an identifying design whose exploration cost can be calibrated to it. This restricted-model instance supplies both.

EXAMPLE EC.2. Retain the three parallel paths and action convention of Section 4.2: $B_0 := \emptyset$ and $B_i := \{\Gamma_i\}$, $i = 1,2,3$, with $K = 1$ and lowest-index tie-breaking. Replace the cost support by four scenarios,

	c_1	c_2	c_3	c_4
Γ_1	1	1	2	3
Γ_2	10	1	1	1
Γ_3	10	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$

$$p = \left(\tfrac{1}{100}, \tfrac{49}{100}, \tfrac{1}{4}, \tfrac{1}{4}\right),$$

and take $\mathcal{M} = \{q \in \Delta^3 : q_3 = q_4\}$. The induced payoff vectors are

$$r_{B_0} = r_{B_2} = (1, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2})^\top, \qquad r_{B_1} = (10, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2})^\top, \qquad r_{B_3} = (1,1,1,1)^\top.$$

Actions B_0, B_1, and B_2 distinguish scenario 1 from $\{2,3,4\}$, whereas B_3 distinguishes $\{1,2\}$ from $\{3,4\}$. At p, the expected values of (B_0, B_1, B_2, B_3) are $(0.505, 0.595, 0.505, 1)$, so $\mathcal{B}^*(p) = \{B_3\}$, $\Delta_{B_0}(p) = \Delta_{B_2}(p) = 0.495$, and $\Delta_{B_1}(p) = 0.405$. A costly confounding alternative must preserve the path law of B_3, which requires $q_1 + q_2 = 1/2$. Only B_1 can then displace B_3: its expected value is $1/2 + (19/2)q_1$ and exceeds 1 exactly when $q_1 > 1/19$.

The actions B_0, B_1, and B_2 have the same information number, $D((1/100, 99/100) \,\|\, (q_1, 1-q_1))$, while B_3 carries no information against these alternatives. This divergence is increasing for $q_1 > 1/19$, and its limiting value as $q_1 \downarrow 1/19$ is

$$D_{\mathrm{cal}} := D((1/100, 99/100) \,\|\, (1/19, 18/19)) \approx 0.026969.$$

The model-class LBP therefore minimizes $0.495x_{B_0} + 0.405x_{B_1} + 0.495x_{B_2}$ subject to $(x_{B_0} + x_{B_1} + x_{B_2})D_{\mathrm{cal}} \geq 1$, giving

$$x^*_{B_0} = x^*_{B_2} = 0, \qquad x^*_{B_1} = \frac{1}{D_{\mathrm{cal}}} \approx 37.079, \qquad \kappa^*_{\mathcal{M}}(p) = \frac{0.405}{D_{\mathrm{cal}}} \approx 15.017.$$

Finally, $\mathcal{E} = \{B_1, B_3\}$ is identifying over $\mathcal{M}$: B_1 identifies q_1, B_3 then identifies q_2, and the model restriction and normalization determine $q_3 = q_4$. Thus the cheapest informative action combines with the optimal action to form the identifying set used in Section 8.3.

F.6. All-Arc Regret Time Paths

Figures EC.2 and EC.3 show the trajectories behind the terminal values in Tables 4 and 5, making visible what their standard-error columns only imply: Value-UCB holds a single optimistic action for long stretches, giving nearly deterministic regret paths.

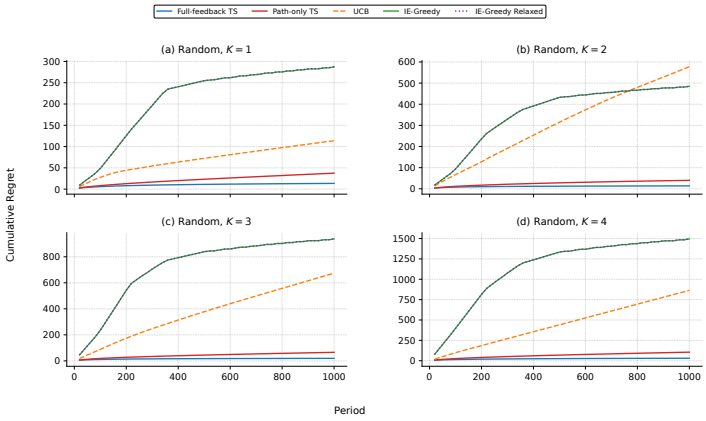

Figure EC.2 E4: mean cumulative pseudo-regret over time on the random graph benchmark.

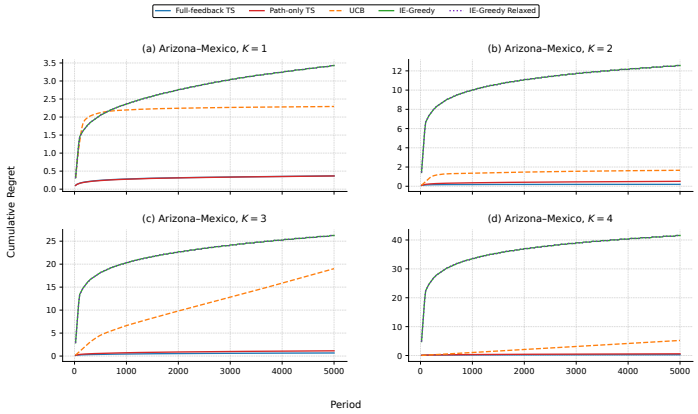

Figure EC.3 E4: mean cumulative pseudo-regret over time on the Arizona–Mexico benchmark.

F.7. Layered Graph Benchmark

Table EC.1 reports the all-arc layered graph benchmark of Section 8.5, isolating depth and width on a regular topology. It reproduces the random graph pattern: Full-feedback TS is the best informational reference, Path-only TS stays close to it, and identifying exploration is expensive at the short horizon. In the largest block Path-only TS ends at 39.4 versus 27.5.

Table EC.1 All-arc layered graph benchmark at $T = 1000$. Rows combine $L \in \{3, 4, 5\}$ layers with $(w, K) \in \{(2, 1), (4, 2), (5, 3)\}$; columns follow Table 4.

Layers	Nodes per layer	K	mean $\lvert\mathcal{B}\rvert$	Rej.	Mean regret (SE; max)					Policy time (s)				
					Full-feedback TS	Path-only TS	UCB	IE-Greedy	IE-Greedy Relaxed	Full-feedback TS	Path-only TS	UCB	IE-Greedy	IE-Greedy Relaxed
3	2	1	13	167	**18.2** (4.6; 94.9)	161.0 (40.0; 826.0)	278.8 (116.0; 2,700)	2,494 (298.6; 7,009)	2,494 (298.6; 7,009)	0.0426	3.030	0.1729	1.230	**0.0143**
3	4	2	821	0	**23.0** (3.7; 65.6)	78.5 (21.7; 614.7)	644.7 (168.4; 3,285)	984.6 (103.3; 2,355)	984.9 (103.3; 2,355)	0.0383	2.250	1.462	2.981	**0.0375**
3	5	3	36.1k	0	**33.8** (5.1; 99.1)	94.5 (24.1; 625.9)	635.2 (188.7; 4,254)	1,111 (100.5; 2,464)	1,111 (100.5; 2,464)	**1.042**	2.625	8.940	3.747	1.066
4	2	1	17	18	**25.1** (6.1; 120.5)	111.6 (29.8; 763.6)	201.5 (79.0; 1,754)	2,088 (247.4; 7,131)	2,089 (247.4; 7,131)	**0.0066**	2.631	0.1640	1.701	0.0105
4	4	2	1.6k	0	**18.8** (3.8; 80.3)	35.9 (9.4; 259.6)	562.2 (148.5; 2,907)	689.8 (78.2; 2,070)	689.8 (78.2; 2,070)	0.0864	1.367	2.355	4.019	**0.0814**
4	5	3	102.4k	0	**35.4** (5.9; 117.2)	45.3 (7.4; 144.3)	636.7 (185.3; 5,033)	897.2 (98.3; 2,482)	897.3 (98.3; 2,483)	4.135	5.245	20.04	7.346	**4.073**
5	2	1	21	6	**30.2** (6.4; 147.4)	85.6 (18.8; 438.0)	785.1 (220.8; 3,904)	1,460 (180.2; 4,350)	1,461 (180.0; 4,350)	**0.0054**	1.867	0.1883	1.971	0.0119
5	4	2	2.6k	0	**25.3** (4.8; 88.5)	36.3 (7.3; 139.8)	443.4 (117.3; 2,005)	591.4 (57.9; 1,549)	591.4 (57.9; 1,549)	0.1338	0.9853	2.966	3.768	**0.1252**
5	5	3	221.9k	0	**27.5** (4.9; 94.7)	39.4 (10.4; 285.5)	553.6 (147.4; 3,252)	957.7 (61.5; 1,921)	957.7 (61.5; 1,921)	**10.43**	10.46	34.77	13.55	10.61