CONTINUATION-AWARE DECISIONS FOR MATROID-CONSTRAINED RESTLESS BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Restless multi-armed bandit (RMAB) methods often reduce joint decisions to local scalar priorities. We ask whether this representation remains sufficient when feasible active sets form a matroid. An exact three-arm construction shows that state-local scalar scores can fail to represent the optimal policy under a fixed greedy selector, despite independent transitions and additive rewards. We therefore compare feasible sets using the Whittle continuation value. For binary-state RMABs, a low-order continuation approximation can be learned from known dynamics without joint-state value labels. Under explicit regularity, gap, and transition-row coverage assumptions, we prove a conditional expected finite-horizon regret bound of $T\epsilon_{\mathrm{struct}} + O(\log T)$, where $\epsilon_{\mathrm{struct}}$ measures the structural error of the true-model frozen policy. Experiments show high decision fidelity under known dynamics and lower frozen-policy Q-loss than plugin and arm-wise optimistic baselines under unknown dynamics. The improvement over the plugin policy is not robust to several distribution shifts, and sparse online fitting can distort early decisions.

1 INTRODUCTION

Restless multi-armed bandits (RMABs) extend the classical multi-armed bandit (MAB) model to sequential resource allocation problems where multiple stochastic processes evolve simultaneously and only a subset can be activated at each decision epoch. Each arm follows a Markov decision process (MDP) whose transition dynamics depend on whether the arm is active or passive, and passive arms continue to evolve. Applications include healthcare resource allocation (Wang et al., 2023a), dynamic multichannel access (Liu & Zhao, 2010), deadline scheduling (Yu et al., 2018), and demand response control (Chen & Hou, 2024). This creates a repeated prioritization problem under a shared resource constraint. A common approach is to assign each arm a scalar priority, such as the Whittle index (Whittle, 1988), and select a feasible subset based on these priorities.

Many RMAB formulations use a cardinality budget, allowing at most K arms to be activated at each decision epoch (Wang et al., 2023b). We consider the more general case where feasible actions are the independent sets of an arbitrary matroid $\mathcal{M}$. This includes cardinality and partition constraints over predefined groups as special cases, as well as graphic, laminar, and linear matroids, which encode acyclicity, hierarchical capacity, and linear independence, respectively. Given scalar arm scores, selecting a feasible set under a matroid is a standard combinatorial optimization problem (Edmonds, 1971). However, since the current action changes future arm states, arm-local scalar scores may not contain enough information to determine the joint decision.

This limitation is not specific to the Whittle index. Even with arbitrary arm-specific scalar scores based only on local states, a fixed greedy matroid selector may fail to represent the optimal joint decision. We show this with an exact rational three-arm example under a rank-two uniform matroid. A contextual extension represents the same policy. This representational obstruction already appears in the cardinality special case.

We use Whittle continuation information to compare feasible actions through immediate reward and expected future value. The action is constructed sequentially subject to matroid feasibility, followed by at most one add, delete, or swap correction. We then study how continuation approximation error affects the ranking, stopping, and local correction decisions made during this construction.

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

Sparse second-order continuations preserve the tested known-model decisions with high fidelity and can be learned without exact joint-state value labels. We also provide a conditional regret analysis connecting transition estimation, continuation error, and decision stability under explicit coverage and gap assumptions. In the main online study, posterior-guided exploration has lower frozen-policy Q-loss than the plugin and arm-wise optimistic baselines, although its difference from random exploration is not statistically resolved. Several distribution shifts reverse the improvement over the plugin baseline, and causal sparse fitting exhibits substantial early distortion, so these results do not establish end-to-end scalable online performance.

2 Related Work

Whittle index policies and online learning. The Whittle index decomposes RMABs into single-arm problems (Whittle, 1988). UCWhittle learns optimistic Whittle indices online and proves sublinear regret for a Lagrangian relaxation of the cardinality-constrained RMAB (Wang et al., 2023b). Thompson sampling also obtains scalable regret relative to the Whittle policy (Akbarzadeh & Mahajan, 2024). GINO-Q learns gain indices without standard Whittle indexability and targets asymptotic optimality in large RMABs (Chen et al., 2026). Decision-focused RMAB learning shows that transition prediction accuracy need not track policy quality (Wang et al., 2023a). Classical priority policies have asymptotic guarantees under population scaling (Weber & Weiss, 1990). We instead study exact finite-N representability by arbitrary state-local scores under a fixed greedy matroid selector.

Combinatorial and structured RMAB decisions. Matroid bandits learn modular objectives under matroid constraints, where weighted optimization admits greedy selection (Kveton et al., 2014). Our setting is dynamic, with additive immediate rewards but continuation coupling through future feasible selections. SEQUOIA learns joint action values for combinatorial restless bandits and selects actions by mixed-integer optimization (Xu et al., 2025). Global reward RMABs study nonseparable immediate rewards and develop adaptive index and planning methods (Raman et al., 2024). Contextual RMABs incorporate exogenous global context into arm dynamics and allocation (Chen & Hou, 2024). Here, transitions remain independent and rewards additive. Coupling arises through future matroid-constrained decisions.

Decision-focused value approximation. Reinforcement learning work studies how prediction and value approximation errors affect downstream decisions. Action-gap analysis shows that approximation error matters relative to the margin between competing actions (Farahmand, 2011). Factored Markov decision process (MDP) methods exploit local transition structure to compute expected future values (Guestrin et al., 2003). Value-aware model learning and value equivalence evaluate models through their effects on planning and value estimation (Farahmand et al., 2017; Grimm et al., 2020). RMAB rollout methods use a base policy continuation value to improve current decisions (Mittal et al., 2023). We apply these ideas to the comparisons in matroid-feasible construction, identifying continuation errors that change ranking, stopping, and posterior comparisons.

Related to these directions, we show a limitation of state-local scores under a fixed greedy matroid selector, and develop a continuation-based feasible-set rule without learning a general joint action-value function or solving a mixed-integer program at each decision.

3 Problem Setting

3.1 Restless Multi-Armed Bandits and Matroid Constraints

Consider N restless arms indexed by $\mathcal{E} := \{1, \dots, N\}$. Arm $i \in \mathcal{E}$ has finite state space $\mathcal{S}$, binary action space $\mathcal{A} := \{0, 1\}$, reward function $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and transition kernel $P_i(s_i' \mid s_i, a_i)$. Let $P := (P_i)_{i \in \mathcal{E}}$ be the transition model. Action $a_i = 1$ activates arm i, and $a_i = 0$ leaves it passive. Both active and passive arms continue to evolve. The joint state is $\mathbf{s} = (s_1, \dots, s_N) \in \mathcal{S}^N$, and an action is an active set $A \subseteq \mathcal{E}$ with $a_i(A) := \mathbb{1}\{i \in A\}$. Rewards are additive across arms $R(\mathbf{s}, A) = \sum_{i \in \mathcal{E}} R_i\big(s_i, a_i(A)\big)$. Arms evolve independently conditional on the current state and action with $P(\mathbf{s}' \mid \mathbf{s}, A) = \prod_{i \in \mathcal{E}} P_i\big(s_i' \mid s_i, a_i(A)\big)$. Feasible active sets form a matroid

$\mathcal{M} = (\mathcal{E}, \mathcal{I})$, where the ground set $\mathcal{E}$ consists of the arms and $\mathcal{I} \subseteq 2^{\mathcal{E}}$ is the family of independent sets. It satisfies $\emptyset \in \mathcal{I}$, $B \subseteq A \in \mathcal{I} \Rightarrow B \in \mathcal{I}$, and the exchange property: if $A, B \in \mathcal{I}$ and $|A| < |B|$, then some $j \in B \setminus A$ satisfies $A \cup \{j\} \in \mathcal{I}$. We assume access to an independence oracle for $\mathcal{M}$.

Given P, a stationary policy is a mapping $\pi : \mathcal{S}^N \to \mathcal{I}$. Let $\mathbf{s}_t$ and A_t denote the joint state and action set at decision epoch $t \in \mathbb{Z}_{\geq 0}$, respectively. For a discount factor $\gamma \in (0, 1)$, the value of policy π from state $\mathbf{s}$ is

$$V_P^\pi(\mathbf{s}) = \mathbb{E}_P^\pi \left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{s}_t, A_t) \,\middle|\, \mathbf{s}_0 = \mathbf{s} \right].$$

The corresponding state-action value is the immediate reward from A plus the discounted value of following π:

$$Q_P^\pi(\mathbf{s}, A) = R(\mathbf{s}, A) + \gamma \sum_{\mathbf{s}' \in \mathcal{S}^N} P(\mathbf{s}' \mid \mathbf{s}, A) V_P^\pi(\mathbf{s}').$$

Define $V_P^\star(\mathbf{s}) := \sup_\pi V_P^\pi(\mathbf{s})$. Since the state and action spaces are finite and $\gamma < 1$, an optimal deterministic stationary policy $\pi_P^\star$ exists. Define $Q_P^\star(\mathbf{s}, A) := Q_P^{\pi_P^\star}(\mathbf{s}, A)$. Unless stated otherwise, all value functions use the same discount factor γ.

3.2 Online Learning in Matroid-Constrained RMABs

In the online setting, the reward functions R_i and the matroid $\mathcal{M}$ are known, but the transition model P is unknown. At decision epoch t, the learner observes $\mathbf{s}_t$, selects $A_t \in \mathcal{I}$, receives $R(\mathbf{s}_t, A_t)$, and observes $\mathbf{s}_{t+1}$. Each arm therefore contributes one transition $(s_{i,t}, a_i(A_t), s_{i,t+1})$, so both active and passive transition rows can be estimated online. With history $\mathcal{H}_t = (\mathbf{s}_0, A_0, \ldots, A_{t-1}, \mathbf{s}_t)$, an online policy selects $A_t \in \mathcal{I}$ according to a possibly randomized rule based on $\mathcal{H}_t$. The decision rule is based on the discounted objective above, whereas realized cumulative reward and the regret analysis use an undiscounted finite horizon.

3.3 Whittle Index Policy

For each arm i with transition kernel P_i, the Whittle index is defined through a single-arm problem with activation penalty $w \in \mathbb{R}$ (Whittle, 1988). Let $V_i^w(s_i)$ denote its optimal discounted value:

$$V_i^w(s_i) = \max_{a_i \in \{0,1\}} \left\{ R_i(s_i, a_i) - w a_i + \gamma \sum_{s_i' \in \mathcal{S}} P_i(s_i' \mid s_i, a_i) V_i^w(s_i') \right\}.$$

Let $\mathcal{B}_i(w)$ be the set of states where $a_i = 0$ is optimal in this penalized single-arm problem. Arm i is indexable if its passive set expands monotonically, $\mathcal{B}_i(w_1) \subseteq \mathcal{B}_i(w_2)$ for all $w_1 \leq w_2$, with $\mathcal{B}_i(w) = \emptyset$ for sufficiently small w and $\mathcal{B}_i(w) = \mathcal{S}$ for sufficiently large w. For an indexable arm, the Whittle index at state s_i is the threshold $W_i(s_i) := \inf\{w \in \mathbb{R} : s_i \in \mathcal{B}_i(w)\}$. The matroid-feasible Whittle policy $\pi_P^W : \mathcal{S}^N \to \mathcal{I}$ considers arms in decreasing $W_i(s_i)$, breaks ties by arm index, and adds an arm only if its index is positive and independence is preserved. We assume that all arms considered by the Whittle-based policies are indexable.

4 Continuation-Based Matroid Decision

We first define continuation-based scores for feasible action sets, use them to construct and locally improve a matroid-feasible action, and extend the rule to online learning with unknown dynamics.

4.1 Whittle Continuation and Action Scores

Let π_P^W denote the Whittle policy in Section 3.3 with continuation value $V_P^W(\mathbf{s}) := V_P^{\pi_P^W}(\mathbf{s})$. Although rewards are additive and transitions are independent, V_P^W need not decompose across arms

because future actions remain matroid constrained. For a feasible action set $A \in \mathcal{I}$, we define the continuation-based set score

$$F_P(\mathbf{s}, A) := R(\mathbf{s}, A) + \gamma \sum_{\mathbf{s}' \in \mathcal{S}^N} P(\mathbf{s}' \mid \mathbf{s}, A) V_P^W(\mathbf{s}'). \tag{1}$$

Equivalently, $F_P(\mathbf{s}, A) = Q_P^{\pi_P^W}(\mathbf{s}, A)$, the value of taking A once and then following the Whittle policy. We use it to compare feasible sets.

4.2 Matroid-Feasible Construction and Local Correction

Let $G \in \mathcal{I}$ be the current selected set and define its feasible additions by $\mathcal{C}(G) := \{i \in \mathcal{E} \setminus G : G \cup \{i\} \in \mathcal{I}\}$. For each $i \in \mathcal{C}(G)$, define the continuation gain $\Delta_i(\mathbf{s}, G) := F_P(\mathbf{s}, G \cup \{i\}) - F_P(\mathbf{s}, G)$. Starting from $G = \emptyset$, we repeatedly add the feasible arm with largest $\Delta_i(\mathbf{s}, G)$ whenever that gain is strictly positive. Ties between feasible additions are broken by arm index. In the local correction, the incumbent G is preferred in a tie, and the remaining candidates follow a fixed total order. Since F_P need not be modular, the resulting G need not maximize it over $\mathcal{I}$, so we apply one local correction. Define the neighborhood used for local correction by

$$\begin{aligned}\mathcal{N}(G) :=& \{G\} \cup \{G \cup \{j\} : j \notin G,\ G \cup \{j\} \in \mathcal{I}\} \cup \{G \setminus \{i\} : i \in G\} \\ &\cup \{(G \setminus \{i\}) \cup \{j\} : i \in G,\ j \notin G,\ (G \setminus \{i\}) \cup \{j\} \in \mathcal{I}\}.\end{aligned}$$

The final action $A_P(\mathbf{s})$ is selected from $\arg\max_{A \in \mathcal{N}(G)} F_P(\mathbf{s}, A)$ using the tie rule above.

4.3 Posterior Learning and Exploration

When P is unknown, let $\widehat{P}_t$ denote the plugin estimate used at decision epoch t. The plugin estimate and posterior models are updated when an observed transition-row count reaches a power of two and reused between updates. At decision epoch t, we apply the sequential construction in Section 4.2 using $F_{\widehat{P}_t}$ and denote the resulting set by G_t, then form $\mathcal{N}(G_t)$ for posterior-guided selection. For candidate $A \in \mathcal{N}(G_t)$, define the plugin margin

$$\widehat{m}_t(A) := F_{\widehat{P}_t}(\mathbf{s}_t, A) - F_{\widehat{P}_t}(\mathbf{s}_t, G_t). \tag{2}$$

Let $\tau(t) \le t$ be the most recent model update and $\Pi_{\tau(t)}$ denote the posterior given $\mathcal{H}_{\tau(t)}$. At time $\tau(t)$, draw M independent models $\widetilde{P}_{\tau(t)}^{(1)}, \ldots, \widetilde{P}_{\tau(t)}^{(M)} \sim \Pi_{\tau(t)}$ and reuse them until the next update. For $k \in [M]$, define the anchored margin

$$m_t^{(k)}(A) := F_{\widetilde{P}_{\tau(t)}^{(k)}}(\mathbf{s}_t, A) - F_{\widetilde{P}_{\tau(t)}^{(k)}}(\mathbf{s}_t, G_t). \tag{3}$$

Here, $F_{\widetilde{P}_{\tau(t)}^{(k)}}$ uses the Whittle policy and continuation value computed under the same sampled model $\widetilde{P}_{\tau(t)}^{(k)}$. Let $\overline{m}_t(A)$ and $\sigma_t(A)$ denote the sample mean and sample standard deviation of these posterior margins with denominator $M - 1$. For each nonincumbent candidate $A \in \mathcal{N}(G_t)$, define the exploration bonus

$$b_t(A) := |\overline{m}_t(A) - \widehat{m}_t(A)| + \beta \sigma_t(A), \tag{4}$$

where $\beta > 0$ controls posterior dispersion. The candidate score is $U_t(A) = \widehat{m}_t(A) + b_t(A)$ for $A \neq G_t$, with $U_t(G_t) = 0$. The bonus combines plugin-to-posterior mean shift with posterior dispersion. With β fixed throughout, the bonus is used to guide exploration empirically and is not calibrated as a confidence bound. The selected action is $A_t \in \arg\max_{A \in \mathcal{N}(G_t)} U_t(A)$ under the tie rule in Section 4.2. The online rule uses exact joint-state continuations for the plugin and posterior models, isolating exploration from continuation approximation error but retaining exponential dependence on the joint state space. Sparse approximation is studied in Section 7.4. Algorithm 1 summarizes the online decision rule.

5 Representation and Decision Stability

5.1 Limits of Local Scalar Scores

Let $\Gamma_{\mathcal{M}}$ be the fixed matroid selector that orders arms by decreasing score, breaks ties by arm index, ignores nonpositive scores, and adds an arm only if independence is preserved. For

Algorithm 1 Continuation-Based Matroid Decision with Posterior-Guided Exploration

Require: Current state $\mathbf{s}_t$, history $\mathcal{H}_t$, matroid $\mathcal{M}$
if $t = 0$ or a model update is triggered **then**
  Update $\widehat{P}_t$, draw $\widetilde{P}_{\tau(t)}^{(1:M)}$, and cache the corresponding Whittle policies and continuations
else
  Reuse the current models and their continuations
end if
Construct G_t using $F_{\widehat{P}_t}$ as in Section 4.2 and form $\mathcal{N}(G_t)$
for each $A \in \mathcal{N}(G_t) \setminus \{G_t\}$ **do**
  Compute $\widehat{m}_t(A)$ and posterior margins $m_t^{(1:M)}(A)$
  Compute $\overline{m}_t(A)$, $\sigma_t(A)$, $b_t(A)$, and $U_t(A)$
end for
Set $U_t(G_t) = 0$
Select $A_t \in \arg\max_{A \in \mathcal{N}(G_t)} U_t(A)$
Execute A_t, observe $\mathbf{s}_{t+1}$, and update $\mathcal{H}_{t+1}$

functions $d_i : \mathcal{S} \to \mathbb{R}$, define $\pi_d(\mathbf{s}) = \Gamma_{\mathcal{M}}\big(d_1(s_1), \ldots, d_N(s_N)\big)$. Define the class $\Pi_0 := \{\pi_d : d_i : \mathcal{S} \to \mathbb{R} \text{ for all } i \in \mathcal{E}\}$. The functions d_i may depend on the fixed model parameters but not on other arms' states.

Proposition 5.1. *There exists a three-arm binary RMAB with $\gamma = 19/20$ under a rank-two uniform matroid whose unique optimal policy cannot be represented by any $\pi_d \in \Pi_0$.*

Proof sketch. In the constructed instance, the unique optimal actions at $\mathbf{s} = (1, 0, 1)$ and $\mathbf{s}' = (1, 1, 1)$ are $\{2\}$ and $\{1\}$, respectively. Selecting them requires $d_1(1) \leq 0$ at $\mathbf{s}$ to exclude arm 1 and $d_1(1) > 0$ at $\mathbf{s}'$, a contradiction. The rational instance and Bellman verification are in Appendix A.1. □

The obstruction already occurs under the cardinality constraint $|A| \leq 2$. For this instance, it disappears when arm scores can use information about feasible alternatives. Appendix A.2 defines the first-opportunity contextual class Π_1 and gives an exact eight-state witness, showing $\Pi_0 \subsetneq \Pi_1$ for this fixed instance and baseline. The separation is about the information available to the score rather than how it is estimated. We next consider continuation-based comparisons.

5.2 Continuation Error and Decision Preservation

Fix a transition model P and let $\widetilde{V}$ approximate the Whittle continuation value V_P^W, with error $e(\mathbf{s}) := \widetilde{V}(\mathbf{s}) - V_P^W(\mathbf{s})$. For a feasible action $A \in \mathcal{I}$, define the approximate continuation score

$$\widetilde{F}_P(\mathbf{s}, A) := R(\mathbf{s}, A) + \gamma \sum_{\mathbf{s}' \in \mathcal{S}^N} P(\mathbf{s}' \mid \mathbf{s}, A)\widetilde{V}(\mathbf{s}').$$

For any two feasible actions $A, B \in \mathcal{I}$,

$$\begin{aligned} &\Big[\widetilde{F}_P(\mathbf{s}, A) - \widetilde{F}_P(\mathbf{s}, B)\Big] - \Big[F_P(\mathbf{s}, A) - F_P(\mathbf{s}, B)\Big] \\ =&\gamma \sum_{\mathbf{s}' \in \mathcal{S}^N} \big[P(\mathbf{s}' \mid \mathbf{s}, A) - P(\mathbf{s}' \mid \mathbf{s}, B)\big] e(\mathbf{s}'). \end{aligned} \tag{5}$$

Equation 5 shows that continuation error affects the ordering of A and B through its projection onto the difference between their next state distributions.

Proposition 5.2. *For two feasible actions $A, B \in \mathcal{I}$, let the comparison gap be $\Delta_P(\mathbf{s}; A, B) := F_P(\mathbf{s}, A) - F_P(\mathbf{s}, B)$ and suppose $\Delta_P(\mathbf{s}; A, B) \neq 0$. If*

$$\gamma \left| \sum_{\mathbf{s}' \in \mathcal{S}^N} \big[P(\mathbf{s}' \mid \mathbf{s}, A) - P(\mathbf{s}' \mid \mathbf{s}, B)\big] e(\mathbf{s}') \right| < |\Delta_P(\mathbf{s}; A, B)|,$$

then $\widetilde{F}_P$ and F_P induce the same ordering between A and B.

Proof. The approximation error in the score difference is smaller than the magnitude of the reference gap, so its sign cannot change. □

Corollary 5.3. *If the condition in Proposition 5.2 holds for every strict comparison in the reference construction and local correction, and the choice at every reference tie remains unchanged under the fixed tie rule, then the complete construction path and final action are preserved.*

Conditional online guarantee. Under the stated Whittle regularity and gap conditions, transition concentration, bounded frozen-policy span, and sufficient transition-row coverage, the expected reward loss relative to the true-model frozen policy is $O(\log T)$. Relative to the finite-horizon optimum, it is $T\epsilon_{\text{struct}} + O(\log T)$, which need not be sublinear. Coverage is assumed, not established by the policy. Details are given in Appendix A.4.

6 Sparse Continuation Approximation

6.1 Second-Order Continuation Model

Tabular evaluation of V_P^W stores one value per joint state in $\mathcal{S}^N$. For binary arm states, we approximate it with first- and second-order state features. Let $x_i := 2s_i - 1 \in \{-1, 1\}$. For a set of retained arm pairs $\mathcal{J} \subseteq \{(i,j) : 1 \le i < j \le N\}$, define

$$\widetilde{V}_\theta(\mathbf{s}) = c + \sum_{i=1}^{N} b_i x_i + \sum_{(i,j)\in\mathcal{J}} b_{ij} x_i x_j, \tag{6}$$

where θ collects the coefficients c, b_i, and b_{ij}. The additive model is recovered by taking $\mathcal{J} = \emptyset$.

Conditional independence in the factorized transition model gives a closed-form expression for the expected continuation value. For arm i, define $q_i(\mathbf{s}, A) := 2P_i\big(1 \mid s_i, a_i(A)\big) - 1$. Then,

$$\mathbb{E}_P\Big[\widetilde{V}_\theta(\mathbf{s}') \mid \mathbf{s}, A\Big] = c + \sum_{i=1}^{N} b_i q_i(\mathbf{s}, A) + \sum_{(i,j)\in\mathcal{J}} b_{ij} q_i(\mathbf{s}, A) q_j(\mathbf{s}, A). \tag{7}$$

Thus, the approximate continuation score can be evaluated without summing over all 2^N next states:

$$\widetilde{F}_P(\mathbf{s}, A) = R(\mathbf{s}, A) + \gamma \mathbb{E}_P\Big[\widetilde{V}_\theta(\mathbf{s}') \mid \mathbf{s}, A\Big].$$

For fixed support $\mathcal{J}$, one score costs $O(N+|\mathcal{J}|)$, and rank-r construction followed by one correction costs $O(Nr(N + |\mathcal{J}|))$ plus oracle calls. These decision-time bounds exclude fitting and support discovery.

6.2 Analytic Bellman Fitting

The coefficients can be fitted without exact continuation labels since V_P^W satisfies

$$V_P^W(\mathbf{s}) = R\big(\mathbf{s}, \pi_P^W(\mathbf{s})\big) + \gamma \mathbb{E}_P\big[V_P^W(\mathbf{s}') \mid \mathbf{s}, \pi_P^W(\mathbf{s})\big].$$

Define the Bellman residual

$$\ell_\theta(\mathbf{s}) := \widetilde{V}_\theta(\mathbf{s}) - R(\mathbf{s}, \pi_P^W(\mathbf{s})) - \gamma \mathbb{E}_P[\widetilde{V}_\theta(\mathbf{s}') \mid \mathbf{s}, \pi_P^W(\mathbf{s})].$$

For B sampled current states $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(B)}$ with repetitions allowed, the fit is

$$\widehat{\theta}_{\mathcal{J}} \in \arg\min_\theta \left\{ \frac{1}{B} \sum_{m=1}^{B} \ell_\theta(\mathbf{s}^{(m)})^2 + \lambda \|\theta\|_2^2 \right\}. \tag{8}$$

This regularized fit does not guarantee exact recovery of V_P^W if the feature class is misspecified or the sampled states do not identify the coefficients. We use $\lambda = 10^{-8}$ and regularize the intercept. To obtain sparse support, we first fit all $\binom{N}{2}$ pair features, keep the N largest $|b_{ij}|$, and refit on them. Since the initial fit has $O(N^2)$ pair features, support discovery is not linear in N. Sparsification reduces the fitted representation and subsequent action-score cost. The procedure uses the known transition model and Whittle policy, but not exact V_P^W values, optimal actions, or oracle pair support. Following Proposition 5.2, we evaluate final action and construction path agreement in Section 7 since continuation error matters when it changes the relevant comparisons. Online use is examined separately since transition uncertainty also affects the fit.

7 EXPERIMENTS

The experiments examine decision fidelity under known dynamics, online learning under unknown dynamics, and failures under distribution shift or causal sparse approximation. Additional generator, fitting, evaluation, and statistical details are given in Appendix B.

7.1 EXPERIMENTAL SETUP

All experiments use binary arm states, discount factor $\gamma = 0.95$, and uniform, graphic, laminar, and linear matroids. Small known-model instances use exact dynamic programming for reference continuations and decisions, and the larger study uses the analytic Bellman fit in Section 6.2. The main online study assumes known rewards and unknown transitions, uses $N \in \{5, 6\}$ and $T = 5000$, and evaluates paired policies on the same model and environment draws. The online study evaluates exact Whittle continuations for each plugin or posterior model. Causal sparse replacement is studied separately in Section 7.4. Let $f_P(\mathbf{s}) := A_P(\mathbf{s})$ be the continuation-based policy under the true transition model. The primary online action-quality metric is cumulative frozen-policy Q-loss:

$$L_T^{\mathrm{fr}} := \sum_{t=0}^{T-1} \left[Q_P^\star(\mathbf{s}_t, f_P(\mathbf{s}_t)) - Q_P^\star(\mathbf{s}_t, A_t) \right]. \tag{9}$$

This signed metric compares the executed and frozen-policy actions at each encountered state using the same $Q_P^\star$. It is not reward regret, and its cumulative value depends on the learner's trajectory. Realized undiscounted reward is reported separately. Details are given in Appendices B.3–B.7.

7.2 REPRESENTATION AND CONTINUATION FIDELITY

Beyond the three-arm counterexample, the finite-state diagnostic contains 51 certified cases in which no state-local scalar score reproduces the reference optimum. Exact Whittle continuation resolves 31 during sequential construction and 37 after local correction. Details are in Appendix B.2.

Across three independently generated models per size, each evaluated under four matroids, Figure 1(a) shows that the fitted pairwise coefficients have squared energy of 0.24%, 0.04%, and 0.22% of exact continuation variance for $N = 5, 6, 8$, respectively. In Figure 1(b), sparse pairwise terms increase complete construction-path agreement from 84.38% to 96.88% at $N = 5$. At $N = 6$ and $N = 8$, both first- and second-order models achieve high agreement. The energy and held-out results use different fitting protocols, so they are not a matched ablation of the same fitted coefficients. Details are in Appendix B.3.

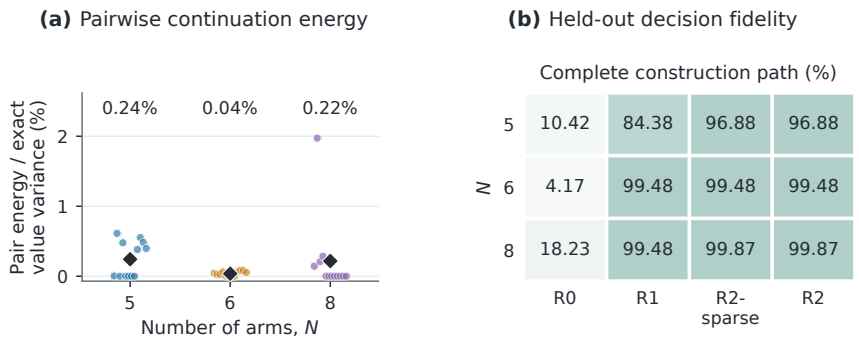

Figure 1: Known-model continuation structure and decision fidelity. **(a)** Pairwise energy relative to exact continuation variance. Each point represents one model–matroid pair. Diamonds show per-size means. **(b)** Held-out path agreement for constant (R0), additive (R1), sparse second-order (R2-sparse), and dense second-order (R2) fits.

The preceding fit uses exact continuation values on training states. We next fit a sparse continuation model from known factorized dynamics without such labels. Table 1 reports analytic Bellman acquisition. Learned support selects the N pair terms with the largest absolute coefficients from a dense

second-order fit and refits without exact continuation values, optimal actions, or oracle pair support. The Bellman fitting states may overlap the evaluation states, so this tests label-free acquisition on fixed models. Fixed support reuses the support selected by historical exact-label fitting. Details are in Appendix B.4. Learned support achieves $99.31\%/98.96\%$ final-action/path agreement at $N = 5$

	Fixed pair support		Learned pair support	
N	Final (%)	Path (%)	Final (%)	Path (%)
5	98.96	98.96	99.31	98.96
6	100.00	100.00	100.00	100.00
8	100.00	99.87	100.00	100.00

Table 1: Analytic sparse continuation fidelity relative to exact Whittle continuation.

and 100% for both at $N = 6$ and $N = 8$, without exact continuation labels or oracle pair support.

7.3 Online Learning Performance

The confirmatory online study compares plugin, random feasible-set exploration, arm-wise optimism, and posterior-guided exploration on 162 independently generated $N \in \{5, 6\}$ models spanning three competition regimes, four paired matroids, and two paired environments per model. Details are in Appendix B.5.

Figure 2 shows cumulative frozen-policy Q-loss from equation 9 and paired endpoint differences at $T = 5000$. Posterior-guided exploration reduces mean loss from 497.956 to 364.858 relative to plugin, with paired difference -133.098 and 95% CI $[-216.971, -46.983]$. The paired differences are -1405.428 for arm-wise optimism, with CI $[-1643.994, -1174.638]$, and -19.965 for random exploration, with CI $[-62.509, 26.876]$. The latter remains unresolved. For realized cumulative reward, the paired difference with plugin is 195.343 with CI $[103.190, 286.440]$.

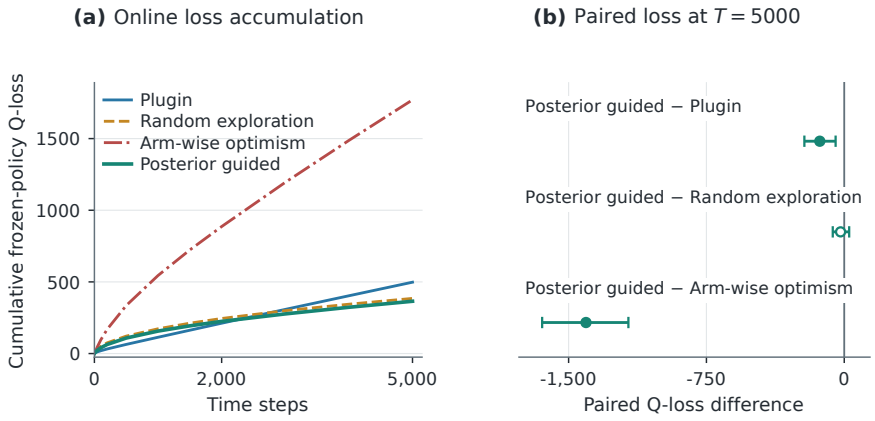

Figure 2: Online learning with unknown transitions. **(a)** Mean cumulative frozen-policy Q-loss over 162 models. **(b)** Paired differences at $T = 5000$ with 95% model-cluster bootstrap CIs. Negative values indicate lower posterior-guided loss.

7.4 Robustness and Online Approximation Limits

Across 150 new models and 840 shift cases with fixed policy parameters, posterior-guided exploration has higher Q-loss than plugin under the persistence, reward heterogeneity, and mixed-population shifts, with CIs excluding zero. The intervention strength and four matroid-family summaries remain unresolved. Values in Figure 3(a) are normalized by reward span for visualization, and conclusions use raw endpoints within each family. Details are in Appendix B.6.

Figure 3(b) compares causal sparse and exact continuation on exact-backend histories from 18 models. Both shared support and per-kernel support show substantial early posterior distortion and fail the prespecified plugin and posterior-action fidelity gates in Appendix B.7. Support differences

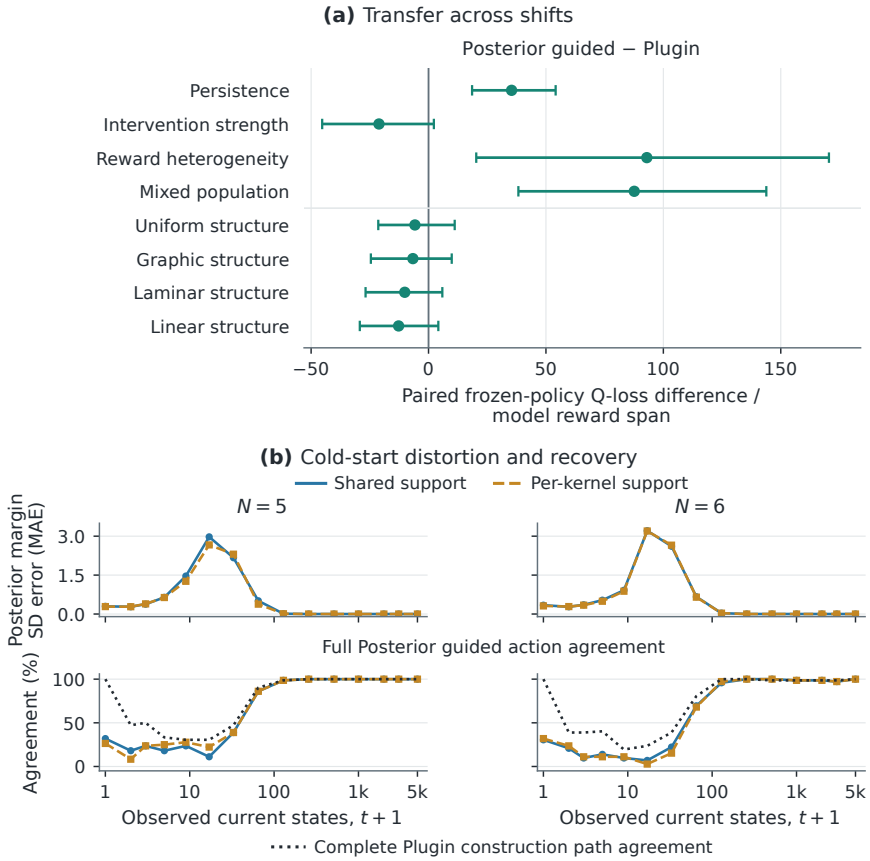

Figure 3: Robustness and causal sparse approximation. **(a)** Q-loss differences under distribution shifts, normalized by reward span, with 95% bootstrap CIs. Positive values indicate higher posterior-guided loss. **(b)** Shared and per-kernel support compared with exact continuation: posterior margin SD MAE, action agreement, and plugin path agreement.

across posterior kernels do not fully explain the early distortion: shared support still has large errors, and dense fits already show substantial move-margin error before sparsification. Over the late checkpoints $t \in \{1000, 2000, 3000, 5000\}$, agreement between the sparse and exact backends on plugin final actions and complete paths averages 100%/100% for $N = 5$ and 99.31%/98.96% for $N = 6$. Posterior-guided action agreement between the two backends averages 100% and 98.61%, respectively. Late recovery does not remove the early failures, so the current sparse backend does not establish end-to-end scalable online performance. Details are in Appendix B.7.

8 Conclusion

We showed that state-local scalar priorities under a fixed greedy selector can be insufficient for matroid-constrained RMABs even with independent transitions and additive rewards. An exact three-arm example separates this class from a contextual extension. We used Whittle policy continuations to compare feasible sets and gave conditions under which approximation preserves construction decisions. Sparse second-order continuations achieve high fidelity on the tested known-model instances without exact joint-state value labels. Under unknown dynamics, posterior-guided exploration has lower frozen-policy Q-loss than plugin and arm-wise optimism, although its difference from random exploration remains statistically unresolved. Distribution shifts can reverse the improvement over plugin, and causal sparse fitting shows substantial early distortion. These results suggest that RMAB methods beyond local priorities should preserve information relevant to feasible action comparisons.

AI USE STATEMENT

In this work, generative AI tools were used to assist with editing and improving the readability and presentation of the manuscript. We did not use generative AI tools for any tasks requiring mandatory disclosure under the ICLR 2027 AI Policy. The authors reviewed the AI-assisted material and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs for the theoretical results are given in Appendices A.1–A.4. Experimental generation, fitting, evaluation, and statistical procedures are described in Appendix B, including the known-model continuation studies, online confirmation, external-validity evaluation, and causal sparse online backend. The experiments use prespecified protocols, fixed random seeds, and paired evaluation designs.

REFERENCES

Nima Akbarzadeh and Aditya Mahajan. On learning whittle index policy for restless bandits with scalable regret. *IEEE Transactions on Control of Network Systems*, 11(3):1190–1202, 2024. doi: 10.1109/TCNS.2023.3333402.

Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, New York, 2000.

Gongpu Chen, Soung Chang Liew, and Deniz Gündüz. GINO-Q: Learning an asymptotically optimal index policy for restless multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20032–20040, 2026. doi: 10.1609/aaai.v40i24.39088.

Xin Chen and I-Hong Hou. Contextual restless multi-armed bandits with application to demand response decision-making. In *2024 IEEE 63rd Conference on Decision and Control*, 2024. doi: 10.1109/CDC56724.2024.10886713.

Jack Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1:127–136, 1971.

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Amir-Massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1486–1494. PMLR, 2017.

Christopher Grimm, Andre Barreto, Satinder P. Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003. doi: 10.1613/jair.1000.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 420–429, 2014.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010. doi: 10.1109/TIT.2010.2068950.

Vishesh Mittal, Rahul Meshram, Deepak Dev, and Surya Prakash. Indexability of finite state restless multi-armed bandit and rollout policy. *arXiv preprint arXiv:2305.00410*, 2023.

Naveen Raman, Zheyuan Ryan Shi, and Fei Fang. Global rewards in restless multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24625–24658, 2024. doi: 10.52202/079017-0777.

Kai Wang, Shresth Verma, Aditya Mate, Sanket Shah, Aparna Taneja, Neha Madhiwalla, Aparna Hegde, and Milind Tambe. Scalable decision-focused learning in restless multi-armed bandits with application to maternal and child health. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pp. 12138–12146, 2023a. doi: 10.1609/aaai.v37i10.26431.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic Whittle index policy: Online learning for restless bandits. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pp. 10131–10139, 2023b. doi: 10.1609/aaai.v37i8.26207.

Richard R. Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990. doi: 10.2307/3214547.

P. Whittle. Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.*, 25(A): 287–298, 1988. doi: 10.2307/3214163.

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

Zhe Yu, Yunjian Xu, and Lang Tong. Deadline scheduling as restless bandits. *IEEE Transactions on Automatic Control*, 63(8):2343–2358, 2018. doi: 10.1109/TAC.2018.2807924.

APPENDICES

A ADDITIONAL THEORETICAL DETAILS

A.1 EXACT COUNTEREXAMPLE

We give the complete rational instance used in Proposition 5.1. The three arms are indexed by $\mathcal{E} = \{1, 2, 3\}$, consistent with Section 3.1. The feasible family is the rank-two uniform matroid $\mathcal{I} = \{A \subseteq \mathcal{E} : |A| \leq 2\}$, and the discount factor is $\gamma = 19/20$. Thus, the obstruction already occurs under a cardinality constraint and does not require more complex matroid structure.

For local state $x \in \{0, 1\}$ and action $a \in \{0, 1\}$, let $p_i^{xa} := P_i(1 \mid x, a)$ and $r_i^{xa} := R_i(x, a)$. The transition probabilities and rewards are given in Table 2. The joint MDP has eight states and seven

i	p^{00}	p^{01}	p^{10}	p^{11}	r^{00}	r^{01}	r^{10}	r^{11}
1	$1/4$	$3/5$	$11/20$	$3/4$	0	$-1/5$	$7/5$	$11/10$
2	$1/2$	$3/4$	$1/2$	$1/2$	0	0	$3/2$	0
3	$1/2$	$1/2$	$1/2$	$1/2$	0	1	1	0

Table 2: Exact rational parameters of the three-arm counterexample.

feasible actions. Exact rational Bellman evaluation gives a unique optimal action at every joint state. Two states suffice for the impossibility argument. At $\mathbf{s} = (1, 0, 1)$, the unique optimal action is $\{2\}$. At $\mathbf{s}' = (1, 1, 1)$, the unique optimal action is $\{1\}$.

Recall that the selector $\Gamma_{\mathcal{M}}$ in Section 5.1 orders arms by decreasing score, ignores nonpositive scores, breaks ties by arm index, and adds an arm only when independence is preserved. If a policy in Π_0 selects $\{2\}$ at $\mathbf{s}$, then it must have $d_1(1) \leq 0$. Otherwise arm 1 would also be added since the rank-two constraint still has available capacity. Selecting $\{1\}$ at $\mathbf{s}'$ requires $d_1(1) > 0$. The two requirements are incompatible. Here, the contradiction is not caused by an optimal-action tie. Exact rational Bellman evaluation gives strictly positive gaps for the two relevant comparisons:

$$Q_P^\star(\mathbf{s}, \{2\}) - Q_P^\star(\mathbf{s}, \{1, 2\}) \approx 0.002635488175,$$
$$Q_P^\star(\mathbf{s}', \{1\}) - Q_P^\star(\mathbf{s}', \emptyset) \approx 0.004358007571.$$

Thus, the obstruction comes from opposite decision requirements for arm 1 at the same local state.

A.2 First-Opportunity Contextual Scores

We now define the contextual class used to state the strict separation from Π_0. Fix the true-model Whittle baseline $w(\mathbf{s}) = \big(W_1(s_1), \dots, W_N(s_N)\big)$. For a target arm i, let $\mathcal{E}_i^+(\mathbf{s}) := \{j \in \mathcal{E} \setminus \{i\} : W_j(s_j) \geq 0\}$. Negative-weight nontarget arms are removed before the opportunity families are formed. Zero-weight nontarget arms remain available.

Define the OUT and IN families

$$\mathcal{I}_i^{\text{out}}(\mathbf{s}) := \{A \in \mathcal{I} : i \notin A,\ A \subseteq \mathcal{E}_i^+(\mathbf{s})\},$$
$$\mathcal{I}_i^{\text{in}}(\mathbf{s}) := \{A \in \mathcal{I} : i \in A,\ A \setminus \{i\} \subseteq \mathcal{E}_i^+(\mathbf{s})\}.$$

The target arm remains forced in the IN family even when its own Whittle index is negative. For either $q \in \{\text{out}, \text{in}\}$, consider the distinct objective values $\left\{\sum_{j\in A} W_j(s_j) : A \in \mathcal{I}_i^q(\mathbf{s})\right\}$ in strictly decreasing order. Let $O_i^{q,(k)}(\mathbf{s})$ denote the kth distinct level when it exists, and let $\eta_i^{q,(k)}(\mathbf{s}) \in \{0, 1\}$ be its availability indicator. If fewer than k distinct levels exist, we set the corresponding value and its availability indicator to zero. The indicator distinguishes this case from an available level whose value is zero.

The first L marked opportunity levels from each family are collected in

$$\Omega_i^{(L)}(\mathbf{s}) := \Big(\eta_i^{q,(k)}(\mathbf{s}), z_i^{q,(k)}(\mathbf{s})\Big)_{\substack{q\in\{\text{out},\text{in}\}\\ 1\leq k\leq L}}, \text{ where } z_i^{q,(k)}(\mathbf{s}) = \begin{cases} O_i^{q,(k)}(\mathbf{s}), & \eta_i^{q,(k)}(\mathbf{s}) = 1, \\ 0, & \eta_i^{q,(k)}(\mathbf{s}) = 0. \end{cases}$$

For $L \geq 1$, define

$$\Pi_L := \left\{\mathbf{s} \mapsto \Gamma_{\mathcal{M}}\big(g_1(s_1, \Omega_1^{(L)}(\mathbf{s})), \dots, g_N(s_N, \Omega_N^{(L)}(\mathbf{s}))\big)\right\},$$

where each g_i is an arbitrary arm-specific real-valued function. The same fixed selector $\Gamma_{\mathcal{M}}$ is used for every class. Since g_i may ignore the contextual coordinates, $\Pi_0 \subseteq \Pi_L$.

For the exact instance in Appendix A.1, a policy in Π_1 reproduces the unique optimal action at all eight joint states. The exact witness is shown in Table 3.

s	Optimal action	Contextual action	Score vector
000	$\{1, 3\}$	$\{1, 3\}$	$(1, \frac{1}{10}, 1)$
001	$\{1, 2\}$	$\{1, 2\}$	$(1, \frac{1}{10}, -1)$
010	$\{1, 3\}$	$\{1, 3\}$	$(1, -1, 1)$
011	$\{1\}$	$\{1\}$	$(1, -1, -1)$
100	$\{2, 3\}$	$\{2, 3\}$	$(-\frac{12111}{102400}, \frac{1}{10}, 1)$
101	$\{2\}$	$\{2\}$	$(-\frac{3311}{25600}, \frac{1}{10}, -1)$
110	$\{1, 3\}$	$\{1, 3\}$	$(\frac{1}{10}, -1, 1)$
111	$\{1\}$	$\{1\}$	$(\frac{1}{10}, -1, -1)$

Table 3: Exact first-opportunity contextual witness for the three-arm counterexample.

A closed-form realization of the witness is

$$g_1\Big(s_1, \Omega_1^{(1)}(\mathbf{s})\Big) = \begin{cases} 1, & s_1 = 0, \\ \frac{1}{10} - z + z^2, & s_1 = 1, \end{cases} \qquad g_2(s_2) = \frac{1}{10} - \frac{11}{10}s_2, \qquad g_3(s_3) = 1 - 2s_3,$$

where

$$b = O_1^{\text{out},(1)}(\mathbf{s}), \qquad c = O_1^{\text{in},(1)}(\mathbf{s}) + \frac{41}{3430}, \qquad z = \frac{b}{1 + \mathbb{1}\{b > c\}}.$$

For this instance, $\frac{41}{3430} = -W_1(1)$, thus $b = c$ exactly at states 101, 110, and 111, and $\mathbb{1}\{b > c\} = 0$ there. Hence, arms 2 and 3 ignore the contextual coordinates, whereas arm 1 uses only the first OUT/IN opportunity levels. All first-opportunity levels used here are available, so the corresponding availability indicators equal one. Exact rational arithmetic evaluates all 56 feasible state-action values and verifies all eight selector outputs. The minimum modular selection margin is $1/10$. Therefore, the optimal policy belongs to Π_1 but not to Π_0, proving $\Pi_0 \subsetneq \Pi_1$ for this fixed instance and baseline.

A.3 Local-Score Infeasibility Certificates

The larger diagnostic study tests local-score representability using linear certificates. For local scores $d_i(s_i)$, define $D_d(\mathbf{s}, A) := \sum_{i \in A} d_i(s_i)$. For a fixed joint state $\mathbf{s}$ and a prescribed optimal action $A_{\mathbf{s}}$, strict modular separation requires $D_d(\mathbf{s}, A_{\mathbf{s}}) - D_d(\mathbf{s}, B) > 0$ for every suboptimal feasible action B. Let $\mathbf{d}$ collect all local scores $d_i(s)$ and let each row of C encode one such inequality. Stacking these inequalities over the finite state table gives a strict linear system $C\mathbf{d} > 0$.

This strict system is consistent with the fixed selector used in Π_0. In a finite state table, an arbitrarily small arm-dependent lexicographic perturbation can encode the fixed arm-index tie rule without changing any strict score ordering. Likewise, a zero score on an excluded arm can be replaced by a small negative score without changing the selected set. Hence, if the fixed selector reproduces a prescribed unique action at every state, there is an equivalent score table satisfying the strict inequalities above. If $C\mathbf{d} > 0$ is infeasible, no such Π_0 representation exists.

By Gordan's theorem of the alternatives (Borwein & Lewis, 2000, Theorem 2.2.1), $C\mathbf{d} > 0$ is infeasible whenever there exists a nonzero vector $y \geq 0$ satisfying $C^\top y = 0$. We normalize $\mathbf{1}^\top y = 1$ and reconstruct the certificate over the rationals. When several optimal actions are available at a state, the procedure considers every admissible choice of one optimal action. A case is called a certified local-score barrier only when all admissible selections are ruled out.

The exact three-arm contradiction reduces to $d_1(1) \leq 0$ and $d_1(1) > 0$. The larger certificate study uses the strict perturbation described above so that the numerical infeasibility test and the fixed-selector policy class have the same selector behavior. For the larger diagnostic instances, the rational certificate establishes infeasibility of the score system constructed from numerically verified optimal-action labels. Those labels are checked separately using Bellman residuals and action gaps. Unlike the three-arm example, these larger certificates use numerically verified optimal actions rather than a fully rational Bellman solution.

A.4 Conditional Regret Analysis

For the binary-state online setting, we give a conditional finite-horizon expected reward guarantee for the posterior-guided policy. The result separates the structural error of the true-model frozen policy from the additional loss due to learning the transition dynamics.

Let f_P denote the policy defined by the continuation-based construction and local correction in Section 4.2 under the true transition model P. Let π^{on} denote the posterior-guided online policy in Section 4.3. For any policy π, define the expected undiscounted finite-horizon reward

$$J_T(\pi) := \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} R(\mathbf{s}_t, A_t) \right].$$

All policies in this comparison start from the same initial distribution μ_0, and expectations include both transition randomness and any randomization in the online policy. Let $\pi_T^\star$ be a finite-horizon optimal policy and define $\mathcal{R}_T^\star := J_T(\pi_T^\star) - J_T(\pi^{\mathrm{on}})$. Then

$$\mathcal{R}_T^\star = [J_T(\pi_T^\star) - J_T(f_P)] + [J_T(f_P) - J_T(\pi^{\mathrm{on}})]. \tag{10}$$

Write $f := f_P$. For $h \geq 0$, let v_h^f be the h-step undiscounted value of f_P, with $v_0^f = 0$. Assume

$$H_f := \sup_{h \geq 0} \operatorname{span}(v_h^f) < \infty. \tag{11}$$

Let D_R denote the range of the one-step joint reward and set $L_f := D_R + H_f$. Define the structural Bellman error

$$\epsilon_{\mathrm{struct}} := \sup_{h \geq 1, \mathbf{s}} \max_{A \in \mathcal{I}} \left\{ R(\mathbf{s}, A) + K_{P,A} v_{h-1}^f(\mathbf{s}) - v_h^f(\mathbf{s}) \right\}, \tag{12}$$

where $K_{P,A}$ is the joint transition operator under action A. Bellman telescoping gives

$$J_T(\pi_T^\star) - J_T(f_P) \leq T \epsilon_{\mathrm{struct}}. \tag{13}$$

For binary arms, there are $q = 4N$ local state-action transition rows. For a transition model P', define

$$d_{\mathrm{row}}(P', P) := \max_{i,s,a} \| P_i'(\cdot \mid s, a) - P_i(\cdot \mid s, a) \|_1 .$$

We next state the local stability conditions needed to identify f_P. Let $g_W > 0$ be the smallest strict Whittle sign or ordering margin used by the reference policy, $g_{\mathrm{con}} > 0$ the smallest strict ranking or stopping gap along the reference sequential construction, and $g_{\mathrm{move}} > 0$ the smallest gap between the selected local correction and its competitors. Assume that there exist $\rho > 0$ and $L_W < \infty$ such that, whenever $d_{\mathrm{row}}(P', P) \le \rho$, every Whittle index is well defined and

$$|W_i^{P'}(s) - W_i^P(s)| \le L_W d_{\mathrm{row}}(P', P)$$

for every arm i and local state s, where $W_i^{P'}$ denotes the Whittle index under P'.

The stability conditions below are uniform over the joint state space, and the margins g_W, g_{con}, and g_{move} are minima over strict reference comparisons. If a reference comparison is tied under P, we assume that its selected outcome remains unchanged throughout the perturbation neighborhood.

Lemma A.1 (Transition-to-margin perturbation)**.** *Suppose $d_{\mathrm{row}}(P', P) \le x$ and $x \le \min\{\rho/4, g_W/(8L_W)\}$. Then every Whittle sign and ordering decision is preserved, so $\pi^W_{P'} = \pi^W_P$. Moreover, for any common incumbent G and feasible candidate A,*

$$\big|\big[F_{P'}(\mathbf{s}, A) - F_{P'}(\mathbf{s}, G)\big] - \big[F_P(\mathbf{s}, A) - F_P(\mathbf{s}, G)\big]\big| \le Cx,$$

where $C_0 := 2\gamma D_R/(1-\gamma)^2$ and $C := NC_0$.

Proof. By the Whittle-index Lipschitz bound assumption, each strict sign comparison changes by at most $L_W x$, and each strict pairwise ordering changes by at most $2L_W x$. Since $x \le g_W/(8L_W)$, all strict Whittle decisions are preserved, and the tie-stability assumption handles reference ties. Hence $\pi^W_{P'} = \pi^W_P =: \pi$. A telescoping expansion of the product distributions gives

$$\sup_{\mathbf{s},A} \|P'(\cdot \mid \mathbf{s}, A) - P(\cdot \mid \mathbf{s}, A)\|_1 \le Nx.$$

For the fixed policy π, adding the same constant to every joint reward leaves anchored F-score differences and $V^\pi_{P'} - V^\pi_P$ unchanged. We may therefore assume that the joint reward lies in $[0, D_R]$. The policy evaluation equations then give

$$\|V^\pi_{P'} - V^\pi_P\|_\infty \le \frac{\gamma N D_R}{(1-\gamma)^2} x.$$

Since $F_{P'} - F_P = \gamma P'\big(V^\pi_{P'} - V^\pi_P\big) + \gamma(P' - P)V^\pi_P$ and $\|V^\pi_P\|_\infty \le D_R/(1-\gamma)$,

$$\sup_{\mathbf{s},A} |F_{P'}(\mathbf{s}, A) - F_P(\mathbf{s}, A)| \le \frac{\gamma^2 N D_R x}{(1-\gamma)^2} + \frac{\gamma N D_R x}{1-\gamma} = \frac{\gamma N D_R x}{(1-\gamma)^2}.$$

Applying this bound to both actions in an anchored comparison proves the result with $C = 2\gamma N D_R/(1-\gamma)^2$. □

The next lemma connects this model perturbation to the posterior margins and empirical exploration bonus in equation 2–equation 4.

Lemma A.2 (Posterior-score stability)**.** *Fix an incumbent G and candidate A. Suppose that the plugin model and all M posterior models satisfy $d_{\mathrm{row}}(P', P) \le x$, and that the corresponding anchored margin errors are at most Cx. Then*

$$|U_t(A) - m_P(A)| \le \kappa_{M,\beta} Cx,$$

where $\kappa_{M,\beta} := 3 + \beta\sqrt{\frac{M}{M-1}}$ and $m_P(A) := F_P(\mathbf{s}_t, A) - F_P(\mathbf{s}_t, G)$.

Proof. Since the plugin margin and each sampled posterior margin lie within Cx of the same true-model margin, both $\overline{m}_t(A)$ and $\widehat{m}_t(A)$ lie within Cx of it, so $|\overline{m}_t(A) - \widehat{m}_t(A)| \le 2Cx$. Since all M margins lie in an interval of length $2Cx$, their variance with denominator M is at most C^2x^2. Hence the sample standard deviation with denominator $M-1$ satisfies $\sigma_t(A) \le Cx\sqrt{\frac{M}{M-1}}$. Using the definition of $U_t(A)$ and equation 4,

$$|U_t(A) - m_P(A)| \le \left(3 + \beta\sqrt{\frac{M}{M-1}}\right) Cx,$$

which is the stated bound. □

For the experiments, $M = 8$ and $\beta = 2$, so $\kappa_{8,2} = 3 + 2\sqrt{\frac{8}{7}}$. Define

$$x_\star := \min\left\{\frac{1}{4}, \frac{\rho}{4}, \frac{g_W}{8L_W}, \frac{g_{\mathrm{con}}}{4C}, \frac{g_{\mathrm{move}}}{4\kappa_{M,\beta}C}\right\}. \tag{14}$$

The terms involving ρ, g_W, and g_{con} preserve the Whittle policy and sequential-construction decisions, and the factor $1/4$ is a technical radius restriction. The last term ensures that the perturbation of any pairwise local-move score difference is at most half of the reference move gap.

For this binary-row analysis, each transition row has an independent $\mathrm{Beta}(1,1)$ prior, the plugin estimate is the posterior mean $\widehat{p}_n := (c+1)/(n+2)$, and M independent posterior models are drawn at each update. For the binary transition rows, define the logarithmic confidence factor L_T and the row-count threshold $n_\star$ by

$$L_T := \log\left(\frac{8(M+1)q(T+1)}{\delta}\right), \qquad n_\star := \left\lceil \max\left\{1, \frac{8L_T}{x_\star^2}, \frac{4}{x_\star}\right\}\right\rceil.$$

Lemma A.3 (Simultaneous plugin and posterior concentration)**.** *With probability at least $1-\delta$, at every committed update time for which every one of the q transition rows has committed count at least $n_\star$, the plugin model and all M independent Beta posterior draws satisfy $d_{\mathrm{row}}(P', P) \le x_\star$.*

Proof. For each transition row, consider its sequence of observed next-state indicators indexed by the number of visits to that row. Let $\bar{p}_n$ be the empirical frequency after n visits. Hoeffding's inequality and a union bound over rows and sample counts $n \le T$ control $|\bar{p}_n - p|$ simultaneously, including at data-dependent update times. Conditional on an update history with c successes in n observations, a posterior draw has distribution $\mathrm{Beta}(c+1, n-c+1)$. The beta-binomial identity and Hoeffding's inequality give

$$\Pr(|\widetilde{p} - \bar{p}_n| > u \mid c, n) \le 2\exp\left\{-2(n+1)u^2\right\}.$$

Set $u = x_\star/4$. For $n \ge n_\star$, the definition of $n_\star$ gives

$$\Pr\left(|\bar{p}_n - p| > \frac{x_\star}{4}\right) \le 2e^{-L_T}, \qquad \Pr\left(|\widetilde{p} - \bar{p}_n| > \frac{x_\star}{4} \mid c, n\right) \le 2e^{-L_T}.$$

The posterior mean differs from $\bar{p}_n$ by $|\widehat{p}_n - \bar{p}_n| \le 1/(n+2) \le x_\star/4$. Hence, on the corresponding event, we have $|\widehat{p}_n - p| \le \frac{x_\star}{2}$ and $|\widetilde{p} - p| \le \frac{x_\star}{2}$. For a binary transition row,

$$\|(\widehat{p}_n, 1-\widehat{p}_n) - (p, 1-p)\|_1 = 2|\widehat{p}_n - p| \le x_\star,$$

and the same ℓ_1 bound holds for each posterior draw. A union bound over at most $q(T+1)$ empirical-frequency events and $qM(T+1)$ posterior-draw events gives total failure probability

$$2q(T+1)e^{-L_T} + 2qM(T+1)e^{-L_T} \le \delta.$$

Cached models introduce no additional draws between updates. □

For fixed N, M, and positive margin and regularity parameters, $n_\star = O(\log T)$.

Let $n_r^{\mathrm{com}}(t)$ denote the number of observations of row r used by the committed model before round t, and define

$$H_T(n) := \sum_{t=0}^{T-1} \mathbb{1}\left\{\min_{1\le r\le q} n_r^{\mathrm{com}}(t) < n\right\} \tag{15}$$

be the number of rounds before all committed transition-row counts reach n.

Theorem A.4 (Conditional regret)**.** *Suppose that the span condition in equation 11, the Whittle-index Lipschitz and tie-stability conditions above, and the positive margins g_W, g_{con}, and g_{move} hold. Suppose that the simultaneous model-concentration event above holds with probability at least $1-\delta$. Then*

$$J_T(f_P) - J_T(\pi^{\mathrm{on}}) \le L_f\left\{\mathbb{E}[H_T(n_\star)] + \delta T\right\}, \tag{16}$$

and therefore

$$\mathcal{R}_T^\star \le T\epsilon_{\mathrm{struct}} + L_f\left\{\mathbb{E}[H_T(n_\star)] + \delta T\right\}. \tag{17}$$

Let $h(n)$ denote the first round at which every transition row has observed count at least n. Assume that, with probability at least $1-\delta_{\mathrm{cov}}$, $h(2n_\star) \le t_0 + \lceil \frac{2n_\star}{c} \rceil$ for some $c > 0$ and $t_0 \ge 0$. Under the power-of-two update schedule, reaching observed count $2n_\star$ guarantees committed count at least $n_\star$. Hence

$$\mathcal{R}_T^\star \le T\epsilon_{\mathrm{struct}} + L_f \left\{ \min\Big(T, t_0 + \left\lceil \frac{2n_\star}{c} \right\rceil + 1\Big) + (\delta + \delta_{\mathrm{cov}})T \right\}. \tag{18}$$

If $t_0 = O(\log T)$ and $\delta = \delta_{\mathrm{cov}} = T^{-2}$, then

$$\mathcal{R}_T^\star \le T\epsilon_{\mathrm{struct}} + O(\log T). \tag{19}$$

Proof. Let $M_T := \sum_{t=0}^{T-1} \mathbb{1}\{A_t \neq f_P(\mathbf{s}_t)\}$. Bellman telescoping along the trajectory generated by π^{on} gives

$$J_T(f_P) - J_T(\pi^{\mathrm{on}}) = \mathbb{E}\sum_{t=0}^{T-1} \left[v_{T-t}^f(\mathbf{s}_t) - R(\mathbf{s}_t, A_t) - K_{P,A_t} v_{T-t-1}^f(\mathbf{s}_t) \right].$$

The summand is zero whenever $A_t = f_P(\mathbf{s}_t)$ and is at most L_f otherwise. Therefore

$$J_T(f_P) - J_T(\pi^{\mathrm{on}}) \le L_f \mathbb{E}[M_T].$$

On the simultaneous concentration event, once every committed transition row has at least $n_\star$ observations, the bounds in equation 14 preserve the Whittle comparisons and the complete reference construction. The plugin incumbent and local neighborhood therefore agree with those under P. Lemma A.2 then preserves the posterior-guided choice within the local neighborhood. Hence the online action agrees with f_P at every later encountered state. Therefore, $\mathbb{E}[M_T] \le \mathbb{E}[H_T(n_\star)] + \delta T$, which proves equation 16. Combining this result with equation 13 gives equation 17.

On the coverage event, every transition row reaches observed count at least $2n_\star$ by round $t_0 + \lceil 2n_\star/c \rceil$. The power-of-two update schedule then guarantees committed count at least $n_\star$ for every row, so

$$H_T(n_\star) \le \min\left\{ T,\ t_0 + \left\lceil \frac{2n_\star}{c} \right\rceil + 1 \right\}.$$

Accounting for the coverage and concentration failure events gives equation 18. Since $n_\star = O(\log T)$ for fixed positive margins and regularity parameters, the logarithmic bound equation 19 follows. □

The logarithmic remainder requires positive decision gaps, the Whittle-index Lipschitz and tie-stability conditions above, and sufficient transition-row coverage. The practical posterior-guided policy is not shown to generate this coverage, so the result does not give an unconditional logarithmic-regret guarantee. The logarithmic term bounds learning loss relative to the frozen true-model policy, whereas the total bound relative to the finite-horizon optimum need not be sublinear when $\epsilon_{\mathrm{struct}} > 0$. The analysis uses exact continuation evaluation under the committed models and does not extend directly to the sparse online backend.

B Additional Experimental Details

B.1 Base Generator and Exact Evaluation

All studies use independent binary-state restless arms with additive rewards and discount factor $\gamma = 0.95$. The base generator mixes improving and crossed transition regimes. For improving arms, passive transition probabilities are sampled from $[0.02, 0.40]$ in state 0 and $[0.25, 0.90]$ in state 1. For each improving arm i and state $s \in \{0, 1\}$, the active transition probability is $p_{i,s,1} = p_{i,s,0} + u_{i,s}(1 - p_{i,s,0})$ with $u_{i,s} \sim \mathrm{Unif}[0.25, 0.90]$ sampled independently. For crossed arms, active and passive transition probabilities are sampled independently from $[0.05, 0.95]$. Each arm has quality parameter $q_i \sim \mathrm{Unif}[0.5, 1.8]$ and reward $R_i(s_i, a_i) = q_i s_i + a_i\big(b_i(s_i) - c_i\big)$. For low, medium, and high competition, respectively, costs c_i are sampled independently across arms from $\mathrm{Unif}[0.55, 1.25]$, $\mathrm{Unif}[0.10, 0.45]$, and $\mathrm{Unif}[0, 0.12]$. Benefits $b_i(s)$ are sampled independently across arms and states, independently of the costs, from $\mathrm{Unif}[-0.10, 0.15]$, $\mathrm{Unif}[0, 0.30]$, and $\mathrm{Unif}[0.25, 0.65]$. The labels describe the sampling distributions only, and instances are not filtered by realized difficulty. Write $p_{i,s,a} := P_i(1 \mid s, a)$.

Matroid families. The experiments use study-specific assignments of uniform, graphic, laminar, and linear matroids. The six-element graphic template is the cycle matroid of K_4 with ordered edges

$$(1,2),(1,3),(1,4),(2,3),(2,4),(3,4).$$

The six-element laminar template has root capacity three, capacities two on $\{1,2,3\}$ and $\{4,5,6\}$, and capacities one on $\{1,2\}$ and $\{4,5\}$. The three rank-three linear templates over $\mathbb{Q}^3$ are

$$\begin{aligned}\mathcal{V}_{\text{sums}} &= \{(1,0,0),(0,1,0),(0,0,1),(1,1,0),(0,1,1),(1,0,1)\},\\ \mathcal{V}_{\text{parallel}} &= \{(1,0,0),(2,0,0),(0,1,0),(0,0,1),(1,1,0),(0,1,1)\},\\ \mathcal{V}_{\text{plane}} &= \{(1,0,0),(0,1,0),(1,1,0),(1,-1,0),(0,0,1),(1,0,1)\}.\end{aligned}$$

For $N = 5$, the graphic, laminar, and linear templates are restricted to their first five elements. In the known-model representation and analytic-acquisition studies with $N > 6$, arms are partitioned into consecutive blocks of at most six elements and the corresponding template is applied within each block, and a final short block uses the corresponding prefix restriction. The blockwise laminar construction also has global capacity $\lfloor N/2 \rfloor$. Linear independence is evaluated over $\mathbb{Q}$. Table 4 summarizes the fixed assignments used by each experimental block.

For the small exact studies, we enumerate the complete joint state space and all feasible independent sets. Optimal values and actions are computed by value iteration and checked independently by policy iteration. Numerical optimal-action sets use tolerance 10^{-8}. The maximum Bellman residual in the primary stress test is below 10^{-11}. All arms in the frozen ground-truth instances used in the reported experiments passed numerical indexability checks.

Indexability checks. For each binary-state arm, we enumerate the four stationary deterministic policies and construct their affine optimality intervals in the activation penalty. Indexability is checked by verifying monotone expansion of the passive-optimal state set across interval boundaries and intervening cells. The primary stress and external-validity studies retain per-arm indifference and bisection cross-checks in addition, whereas the remaining frozen studies use the same affine-envelope checks for Whittle-index computation. These checks apply to the realized instances and do not give an analytic indexability guarantee for the full generator family.

For the known-model studies, let $r_{\text{km}}(N) = 2$ for $N \le 6$ and $r_{\text{km}}(N) = \lfloor N/2 + 1/2 \rfloor$ otherwise. The low-, medium-, and high-competition regimes use $\mathcal{V}_{\text{sums}}$, $\mathcal{V}_{\text{parallel}}$, and $\mathcal{V}_{\text{plane}}$, respectively. For online studies, let $r_j = 2 + (j \bmod 2)$ and $\mathcal{V}_j = (\mathcal{V}_{\text{sums}}, \mathcal{V}_{\text{parallel}}, \mathcal{V}_{\text{plane}})_{j \bmod 3}$, where j is the underlying-model replicate within an N-by-competition cell.

Study	N	Uniform	Other families
Local-score stress test	5	rank 2	none
Local-score stress test	6	ranks 2, 3	graphic, laminar, and all three linear templates
Known-model representation	5, 6, 8, 10	$r_{\text{km}}(N)$	graphic, laminar, and regime-matched linear
Analytic sparse acquisition	5, 6, 8, 10, 12, 16, 20	same as above	same as above
Online confirmation	5, 6	r_j	graphic, laminar, and linear $\mathcal{V}_j$
EV-A: arm shifts	5, 6	r_j	graphic, laminar, and linear $\mathcal{V}_j$
EV-B: structural shifts	5, 6	three variants	three variants from each remaining family
Causal sparse online	5, 6	r_j	graphic, laminar, and linear $\mathcal{V}_j$

Table 4: Frozen matroid assignments by experimental block. For online studies, j denotes the underlying-model replicate within an N-by-competition cell.

B.2 Local-Score Stress Test

The local-score diagnostic compares four score families. H1 applies a positive arm-specific scaling to the Whittle indices. H2 applies an arm-specific additive offset. H3 combines scaling and offset. H4 assigns an unrestricted score to every arm-local-state pair and corresponds to the state-local class Π_0 under the fixed selector described in Appendix A.3. The bounded formulations use $10^{-4} \le$

$\alpha_i \leq 4$ for H1, $-2 \leq \beta_i \leq 2$ for H2, the same scaling bounds and $-5 \leq \beta_i \leq 5$ for H3, and scores in $[-1, 1]$ for H4. The H4 bound is without loss of generality since a common positive rescaling preserves the selected action. For each class, we maximize a common strict-separation margin and independently verify the recovered solution. H4 infeasibility is reported only when the rational Gordan certificate described in Appendix A.3 is verified.

The primary stress test contains 180 independently generated RMAB instances and 720 evaluations across instance–matroid pairs. The 90 five-arm instances are evaluated only under the rank-two uniform matroid. Each of the 90 six-arm instances is evaluated under seven matroids, giving $90 + 90 \times 7 = 720$ evaluations. Multiple matroid structures applied to the same instance are paired observations, and the underlying RMAB instance is the statistical unit.

As a diagnostic ceiling, direct maximization of the reference joint $Q^\star$ resolves all 51 certified local-score barriers. The continuation-based sequential construction resolves 31 and the bounded local correction increases this to 37.

B.3 Known-Model Continuation Representation

The structured-continuation study contains 12 independent underlying models, three at each of $N \in \{5, 6, 8, 10\}$. The reported fidelity percentages summarize this fixed benchmark and are not population-level estimates. Each model is evaluated under uniform, graphic, laminar, and linear matroid constraints using the assignments in Table 4. State splits are frozen before fitting.

N	Train/model	Validation/model	Held-out/model
5	20	4	8
6	40	8	16
8	160	32	64
10	640	352	32

Table 5: Joint-state splits per model in the known-model continuation study. Each split is shared across the four matroid structures.

R0 contains only an intercept, R1 adds all first-order centered binary features, and R2 adds all pairwise terms. R2-sparse selects the N pairwise terms with the largest absolute coefficients in the dense training fit and refits all coefficients on the same training states. Exact continuation values are used only on the fitting states in the held-out experiment. The $N = 5, 6, 8$ cases form the primary fidelity study and $N = 10$ is diagnostic.

The pair-energy calculation in Figure 1(a) uses a centered second-order fit over the full joint state space. The decision-fidelity results use the frozen train/validation/held-out splits above. Pairwise energy is $\sum_{i<j} b_{ij}^2$ normalized by exact continuation variance. This statistic is descriptive. Proposition 5.2 depends on the action-conditioned projection of approximation error.

B.4 Analytic Sparse Acquisition

The known-model acquisition study tests whether sparse support and coefficients can be obtained without exact joint-state value labels. For a given known transition model, current states are generated by the Whittle policy and used to form the analytic Bellman equations in Section 6.2. The primary configuration uses 1000 current-state equations and ridge coefficient 10^{-8}. Repeated current states are included with multiplicity. Equivalently, they may be represented by their normalized empirical frequencies without changing the fitting objective. Each fit first estimates a dense second-order model, ranks pair coefficients by absolute magnitude, selects exactly the top N pairs, and refits on the selected support. No exact value target, optimal action, or oracle pair graph enters this procedure. We use three independent training seeds per case.

The primary study uses $N = 5, 6, 8$, with $N = 10$ as a diagnostic case. The same frozen acquisition rule is applied independently to the tested $N = 12, 16, 20$ known-model cases. The larger experiment is used only as bounded computational evidence, and $N = 20$ has no exact continuation reference. The current-state fitting streams may revisit evaluation states. Thus, this study tests acquisition without exact value labels on the same underlying models and does not test generalization restricted to unseen states or model populations.

B.5 Online Confirmation

The main online confirmation contains 162 independently generated models with $N \in \{5, 6\}$, balanced across the three competition regimes. Each underlying model is paired with four matroid structures and two environment realizations. For replicate j within an N-by-competition cell, the four structures are uniform rank r_j, the graphic template, the laminar template, and linear template $\mathcal{V}_j$. Every policy therefore has 1296 trajectories of length $T = 5000$. The four matroid cases share the same underlying arm model, initial states, and exogenous environment streams.

The online comparison includes the plugin policy, random exploration, arm-wise optimism, and posterior-guided exploration. Each binary transition row has an independent $\mathrm{Beta}(1, 1)$ prior. After observing n_1 transitions to state 1 and n_0 to state 0, the posterior is $\mathrm{Beta}(n_1 + 1, n_0 + 1)$ and the plugin estimate is $\widehat{p} = (n_1+1)/(n_0+n_1+2)$. Posterior-guided exploration forms each of its $M = 8$ transition models by sampling these row posteriors independently. Random exploration follows the plugin policy except with probability $\min\left\{1, 2/\sqrt{t+1}\right\}$ when an action is sampled uniformly from the enumerated feasible sets. For the $N = 5, 6$ online experiments, all feasible independent sets are enumerated explicitly, and the exploratory action is sampled uniformly from this finite family.

Arm-wise optimism constructs a count-based optimistic transition model at each update. For a row with $n = n_0 + n_1$, it uses the plugin estimate $\widehat{p}$ and endpoints $L = \max\{0, \widehat{p} - e\}$ and $U = \min\{1, \widehat{p} + e\}$, where $e = 1$ for $n = 0$ and otherwise

$$e = \min\{1, \sqrt{\log(8NT/\delta)/(2n)} + 1/(n + 2)\}$$

with $T = 5000$ and $\delta = 0.05$. For each arm, the upper active and lower passive endpoints are selected when the zero-penalty value of the plugin single-arm model satisfies $V_i^0(1) \geq V_i^0(0)$, with the choices reversed otherwise. Whittle indices, the joint Whittle continuation, sequential construction, and local correction are then recomputed under this optimistic model. Results for arm-wise optimism refer only to this baseline. Posterior-guided exploration uses $\beta = 2$ in equation 4. Models are updated when an observed transition-row count reaches a power of two and reused between updates. For local correction ties, the incumbent is preferred and the remaining feasible actions are ordered by increasing binary action mask.

The primary online comparison changes the exploration rule while keeping the continuation-based decision architecture fixed. It does not establish superiority over methods that learn and optimize a joint action-value function. The primary online endpoint is cumulative frozen-policy Q-loss in equation 9, evaluated on each learner's own state path. Confidence intervals use 2000 percentile bootstrap resamples of the 162 underlying-model clusters, with paired matroids and environment realizations kept together within each resampled cluster. Time steps, states, matroids, and environment replicates are not treated as independent sampling units. The reported intervals are pointwise across policy contrasts and are not adjusted for simultaneous coverage.

B.6 External-Validity Study

The external-validity experiment keeps the online algorithms and their hyperparameters fixed. It contains 150 new underlying models, 840 model–shift cases, and 8400 trajectories (840 cases $\times$ 2 environment replicates $\times$ 5 evaluated policies, including the true-model frozen comparator). The arm-population shifts modify transition persistence, intervention strength, reward heterogeneity, or the mixture of arm types and use the same four-matroid replicate rule as the online confirmation. Each arm-population shift family contains 30 independent underlying models, with five replicates for each combination of $N \in \{5, 6\}$ and competition regime. A separate structural-shift block uses 30 new base models without arm-population shifts, with all 12 structural variants evaluated on each model.

Arm-population shifts. Each shifted model starts from newly generated base parameters (p^0, R^0), and unspecified parameters remain unchanged. Let $j \in \{0, \ldots, 4\}$ denote the replicate within an N-by-competition cell and $\ell \in \{0, 1, 2\}$ index the low-, medium-, and high-competition regimes, respectively.

For the persistence shift, models with $j + \ell + N$ even use $p_{i,0,0} \sim \mathrm{Unif}[0.001, 0.015]$ and $p_{i,1,0} \sim \mathrm{Unif}[0.985, 0.999]$, with $p_{i,0,1} = \min\{0.9999, p_{i,0,0} + \delta_{i,0}\}$ for $\delta_{i,0} \sim \mathrm{Unif}[0.001, 0.010]$, and

$p_{i,1,1} = \min\{0.9999, p_{i,1,0} + \delta_{i,1}\}$ for $\delta_{i,1} \sim \mathrm{Unif}[0.0001, 0.003]$. Models with $j + \ell + N$ odd use $p_{i,0,0} \sim \mathrm{Unif}[0.42, 0.48]$, $p_{i,1,0} \sim \mathrm{Unif}[0.52, 0.58]$, and $p_{i,s,1} = p_{i,s,0} + \delta_{i,s}$, where $\delta_{i,s} \sim \mathrm{Unif}[0.02, 0.06]$. Rewards are unchanged. Models with $j + \ell + N$ odd are near-memoryless, so this family includes both high-persistence and near-memoryless models.

For the intervention-strength shift, passive transitions and rewards remain unchanged and $p_{i,s,1} = p^0_{i,s,0} + u_{i,s}\big(1 - p^0_{i,s,0}\big)$. Models with $j+\ell+N$ even use $u_{i,s} \sim \mathrm{Unif}[0.96, 0.995]$, and the remaining models use $u_{i,s} \sim \mathrm{Unif}[0.005, 0.03]$.

For the reward-heterogeneity shift, transitions remain unchanged and $\log \lambda_i \sim \mathrm{Unif}[\log 0.2, \log 6]$. Let π_i denote the passive stationary probability of state 1, $\pi_i = \frac{p^0_{i,0,0}}{1 - p^0_{i,1,0} + p^0_{i,0,0}}$, and let $i_\star = \arg\min_i \min\{\pi_i, 1 - \pi_i\}$, with ties broken by arm index. Define

$$s_i^{\mathrm{tail}} = \begin{cases} 1, & \pi_i \le 1/2, \\ 0, & \pi_i > 1/2. \end{cases}$$

Draw $B_{i_\star} \sim \mathrm{Unif}[8, 12]$ and define the passive reward baseline

$$h_i(s) = \begin{cases} B_{i_\star} \mathbb{1}\{s = s_{i_\star}^{\mathrm{tail}}\}, & i = i_\star, \\ \lambda_i R_i^0(s, 0), & i \neq i_\star. \end{cases}$$

The shifted reward is

$$R_i(s, a) = h_i(s) + a\left[\lambda_i\big(R_i^0(s, 1) - R_i^0(s, 0)\big) + \xi_{i,s}\Delta_i^R\right],$$

where $\xi_{i,s} \sim \mathrm{Unif}[-0.4, 0.4]$ and $\Delta_i^R = \max\{\max_s h_i(s) - \min_s h_i(s),\ 0.1\}$.

The mixed-population shift randomly assigns one high-persistence, fast-mixing, weak-intervention, strong-intervention, and reward-tail arm when $N = 5$, with one additional high-persistence arm when $N = 6$. Each arm receives only its assigned transformation. The reward-tail arm keeps its base transitions, whereas the other transformed arms keep their base rewards.

Structural shifts. The structural-shift block evaluates the following 12 prespecified variants. Let $\mathcal{E}_k := \{1, \ldots, k\}$, $m := \lfloor N/2 \rfloor$, and let $U_{r,N}$ denote the rank-r uniform matroid on $\mathcal{E}$. For laminar matroids, (S, c) denotes $|A \cap S| \le c$. For linear matroids, let $\mathbf{v}_i$ denote the vector for arm i.

Family	Variant	Definition
Uniform	tight	$U_{1,N}$
Uniform	medium	$U_{m,N}$
Uniform	dense	$U_{N-1,N}$
Graphic	long cycle	$(1,2), (2,3), \ldots, (N-1, N), (N, 1)$
Graphic	triangle bridges	$(1,2), (2,3), (3,1)$ and $(i, i+1)$ for $i = 3, \ldots, N-1$
Graphic	parallel bundle	N distinct parallel edges between two vertices
Laminar	shallow loose	$(\mathcal{E}, N-1), (\mathcal{E}_m, m-1), (\mathcal{E} \setminus \mathcal{E}_m, N-m)$
Laminar	deep chain	$(\mathcal{E}, N-1), (\mathcal{E}_{N-1}, N-2), (\mathcal{E}_{N-2}, 1), (\mathcal{E}_2, 1)$
Laminar	unbalanced tight	$(\mathcal{E}, 3), (\mathcal{E}_{N-1}, 1), (\{N\}, 1)$
Linear	parallel rank two	$\mathbf{v}_i = (1, 0)$ for odd i and $\mathbf{v}_i = (0, 1)$ for even i
Linear	dense Vandermonde	$\mathbf{v}_i = (1, i, i^2, \ldots, i^{N-2})$
Linear	near-parallel rank three	$\mathbf{v}_i = (1000, i, i^2)$

Table 6: Prespecified matroid variants in the structural-shift study.

Linear independence is evaluated over $\mathbb{Q}$. The 12 named variants do not always define distinct labeled feasible-set systems: there are 9 distinct systems at $N = 5$ and 7 at $N = 6$. The structural variants are therefore treated as robustness conditions rather than independent replications. The four family-level structural summaries use the same underlying-model block and are correlated.

The prespecified external-validity conclusions use the raw within-family frozen-policy Q-loss endpoint. No aggregate is formed across shift families. Figure 3(a) uses per-model normalization by reward span only to place different reward populations on a common visual scale. Confidence intervals are pointwise across shift families and are not adjusted for simultaneous coverage.

B.7 CAUSAL SPARSE ONLINE BACKEND

The causal sparse online study evaluates whether the sparse continuation model developed under known dynamics can replace exact continuation in posterior-guided online learning. The study uses 18 new underlying models, nine each at $N = 5$ and $N = 6$, with four paired matroids assigned by the same replicate rule as the online confirmation and two environment realizations. Exact continuation, shared-support sparse continuation, and per-kernel-support sparse continuation produce 432 closed-loop trajectories of length 5000. The three backends evolve on separate action-dependent histories. Figure 3(b) uses a separate checkpoint comparison in which both sparse backends are evaluated on the same 144 exact-backend reference histories.

Sparse fits use only states observed by the corresponding model-update time. We maintain a reservoir of at most 1000 observed joint-state occurrences using Algorithm R reservoir sampling. Repeated states are weighted by their multiplicities within the retained reservoir. The study uses the 15 prespecified checkpoints $t \in \{0, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1000, 2000, 3000, 5000\}$. We summarize early $t \leq 64$, intermediate $t \in \{128, 256, 512\}$, and late $t \in \{1000, 2000, 3000, 5000\}$ behavior separately. Shared support uses one pair support from the plugin model across posterior kernels. Per-kernel support uses a separate pair support for each posterior model. For each posterior kernel, we also evaluate a dense second-order fit on the same causal history using the exact-backend plugin incumbent and candidate set. Signed move-margin error is decomposed into dense-minus-exact and sparse-minus-dense components. For shared support, the corresponding per-kernel dense fit is used only for this diagnostic.

Primary fidelity averages the 15 checkpoints equally within each run before aggregation over paired environments, matroids, and underlying models. The prespecified fidelity gates require plugin final action agreement of at least 98%, complete construction path agreement of at least 95%, conditional sign error of at most 2%, and full posterior-guided action agreement of at least 95%. Conditional sign error is the fraction of candidate moves with nonzero exact margin whose sparse-backend margin has a different sign. The resulting fidelity measure is not the fraction of all 5000 executed actions. The late $t = 1000$–5000 summaries are diagnostic and do not replace the full 15-checkpoint fidelity criterion. Since both sparse strategies fail the full criterion, we do not run a large-N unknown-dynamics study.