FINITE STRUCTURAL QUERIES FOR ASYMPTOTICALLY OPTIMAL MATROID SEMI-BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic matroid semi-bandit algorithms typically assume continued access to an independence oracle. We show that asymptotically optimal regret is achievable with only finitely many oracle queries in expectation over an infinite horizon. For Bernoulli rewards, we introduce ECR-KL, which stores a basis and one feasible exchange partner for each outside element. The partners define groups for simultaneous exploration. Fresh samples are used to test and repair the mean comparisons. It handles arbitrary ties, uses $O(d)$ words of memory, and has a polynomial expected query bound that does not depend on the horizon. ECR-KL attains the optimal logarithmic regret coefficient of the known-matroid problem, including at nonunique optima. A structural lower bound shows that hidden matroids with the same mean vector and optimal basis can require different optimal sampling rates. When all means are distinct, a full-order variant matches the optimal $\Theta(d \log d)$ expected query complexity at balanced rank.

1 INTRODUCTION

In a stochastic *combinatorial semi-bandit*, the learner selects a feasible subset of d arms, receives their total reward, and observes each selected arm's reward (Kveton et al., 2015). We study matroid constraints, including cardinality, partition, and spanning-tree constraints (Kveton et al., 2014). For Bernoulli rewards and a known matroid, change of measure characterizes the optimal logarithmic regret coefficient, which KL-OSM attains under distinct means (Talebi & Proutière, 2016). These guarantees assume the feasible family is available. We instead hide the matroid and count independence queries separately from rewards; a certified basis can be replayed for free. This models settings where each structural query requires nontrivial computation. Recent methods reduce this cost to $O(\log \log T)$ adaptive optimization rounds (Kim et al., 2025) or $O_\mu(\log \log T)$ basis-update rounds (Delage & Gaudel, 2025), but structural access still grows with T. We ask whether the known-matroid coefficient $C^\star(\mu)$ is attainable with finitely many expected oracle queries over an infinite run.

Two natural approaches fail. Querying all $\binom{d}{r}$ rank-r sets identifies the matroid, but it is exponential at balanced rank ($r, d - r = \Theta(d)$). Certifying only an optimal basis is not enough either: Theorem 2 gives hidden matroids with the same means and the same unique optimal basis but different optimal sampling rates. ECR-KL certifies just enough exchange structure to sit between these extremes. It stores one basis and one feasible partner for each outside arm, using $O(d)$ words. The partners define a partition-matroid subfamily of feasible actions that supports parallel exploration (Bérczi & Schwarcz, 2020). When the stored comparisons are correct, this subfamily recovers the known-matroid sampling thresholds. A detected violation triggers a repair on fresh samples and a rebuild, while global KL indices keep every past observation.

Contributions.

- **Finite-query learning.** ECR-KL uses $O(d)$ words and $O(d^2 \log(r+1))$ expected oracle calls over an infinite run. Its expected last-rebuild time is finite, and $R_\mu(T)/\log T \to C^\star(\mu)$ at every mean vector, including those with nonunique optima (Theorem 1).
- **Structural lower bound.** Any universally feasible, uniformly efficient policy that attains $C^\star$ at every unique-optimum instance needs at least $(d - r)\log_2 r$ expected queries on some instance. The hard instances share their means and optimal basis (Theorem 2).

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.

- **Distinct means.** Full-Order ECR-KL uses a constant expected number of builds and $O(d \log d)$ expected queries at balanced rank, matching the lower bound in order (Theorem 3). With ties, a factor-d gap remains at balanced rank.

Exchange partners also appear in MAUB, and KL-OSM supplies the distinct-means regret benchmark (Delage & Gaudel, 2025; Talebi & Proutière, 2016). What distinguishes our work is a finite expected total of structural queries (Table 1).

2 Model and regret benchmark

Notation. Logarithms are natural unless base two is specified. Instance-dependent constants may depend on $(\mathcal{M}, \mu)$. Appendix A.1 collects notation and state variables.

Stochastic interaction. Let $E = [d] = \{1, 2, \cdots, d\}$ index $d \geq 2$ base arms, with $1 \leq r < d$. The learner knows d, r and the model class, but not the mean vector $\mu \in (0,1)^d$ or the family of feasible bases $\mathcal{B} \subseteq \binom{E}{r}$. Before choosing an action in round t, the learner may make finitely many adaptive oracle calls. Let $\mathcal{H}_t$ be the sigma-field generated by the history and policy randomness available at action selection. We require $A_t \in \mathcal{B}$ to be $\mathcal{H}_t$-measurable. The environment then draws $X_{t,i} \sim \text{Bernoulli}(\mu_i)$, independently across t and i. The learner receives the payoff and selected-arm rewards:

$$Y_t = \sum_{i \in A_t} X_{t,i}, \qquad O_t = (X_{t,i})_{i \in A_t}, \qquad \mathbb{E}_\mu[Y_t \mid \mathcal{H}_t] = \mu(A_t) := \sum_{i \in A_t} \mu_i. \tag{1}$$

Only selected-arm rewards are observed. This is standard semi-bandit feedback (Kveton et al., 2014; Cuvelier et al., 2021), but the feasible family is unknown. The observation counts, reward sums, and optimal mean reward are

$$N_i(T) = \sum_{t=1}^{T} \mathbf{1}\{i \in A_t\}, \quad S_i(T) = \sum_{t=1}^{T} \mathbf{1}\{i \in A_t\} X_{t,i}, \quad \mu^\star = \max_{B \in \mathcal{B}} \mu(B).$$

The expected pseudo-regret is

$$R_\mu(T) = T\mu^\star - \mathbb{E}_\mu \sum_{t=1}^{T} Y_t = \mathbb{E}_\mu \sum_{t=1}^{T} [\mu^\star - \mu(A_t)].$$

Expectations are taken over rewards and policy randomness on a fixed instance.

Definition 1 (Uniform efficiency, Talebi & Proutière (2016))**.** A policy is uniformly efficient if, on every fixed instance and for every $a > 0$, $R_\mu(T) = o(T^a)$.

Matroid constraints and structural access. A matroid $\mathcal{M} = (E, \mathcal{I})$ has an independent-set family $\mathcal{I} \subseteq 2^E$ containing $\varnothing$ and closed under subsets. It satisfies augmentation: if $A, B \in \mathcal{I}$ and $|A| < |B|$, some $e \in B \setminus A$ satisfies $A \cup \{e\} \in \mathcal{I}$ (Oxley, 2011, Section 1.1). Its maximal independent sets, or *bases*, have common size r, the rank. We set $\mathcal{B} = \mathcal{B}(\mathcal{M})$ and assume *looplessness*: every singleton is independent. Since $\mu_i > 0$, every optimal independent set is a basis.

The learner accesses $\mathcal{M}$ only through a noiseless independence oracle. A query $C_q \subseteq E$ returns the bit $Z_q = \mathbf{1}\{C_q \in \mathcal{I}\}$ (Talebi & Proutière, 2016, Section 3.1.1). Queries may occur before or between reward rounds and depend on all previous answers and observations. Every call counts; no representation of $\mathcal{M}$ is revealed. A policy is *universally feasible* if it plays a basis almost surely at every round on every admissible instance. Writing Q_T for the number of calls by round T, we require

$$\mathbb{E}_\mu Q_\infty < \infty, \qquad Q_\infty = \sup_{T \geq 1} Q_T.$$

This implies finitely many queries almost surely, but imposes no deterministic cap. Learning continues after the last query, which the learner need not recognize as such.

Greedy selection in decreasing weight order returns a maximum-weight basis (Edmonds, 1971). For $B \in \mathcal{B}$ and $i \notin B$, write

$$K_B(i) = \{j \in B : B - j + i \in \mathcal{B}\}, \qquad B - j + i = (B \setminus \{j\}) \cup \{i\}.$$

The nonempty set $K_B(i)$ consists of the feasible removals: every element of B in a uniform matroid, the selected member of i's group in a capacity-one partition matroid, or the tree edges on the cycle created by adding i in a graphic matroid.

Table 1: Oracle access and regret guarantees.

Method	Reported access bound	Regret
KL-OSM (Talebi & Proutière, 2016)	$O(dT)$ calls	Exact $C^\star$
MAUB (Delage & Gaudel, 2025)	$O_\mu(\log\log T)$ update rounds	$O_\mu(\log T)$
SROQ-CMAB (Kim et al., 2025)	$O(d\log\log T)$ optimization calls	$\widetilde{O}(r\sqrt{dT})$
ECR-KL	$O(d^2\log(r+1))$ expected total calls	Exact $C^\star$
Full-Order ECR-KL	$O(d\log(r+1))$ expected total calls	Exact $C^\star$

Calls are binary independence queries except for SROQ-CMAB's action-returning optimization oracle ($O(\log\log T)$ adaptive rounds for linear rewards). KL-OSM assumes distinct means; its count is for uncached greedy. MAUB's analysis assumes distinct basis values; one update may involve several calls. The ECR-KL bounds are expected totals over the infinite run; ECR-KL allows ties, whereas Full-Order ECR-KL requires distinct means.

Known-matroid benchmark. Define the insertion losses, strictly suboptimal arms, and insertion thresholds by

$$\Delta_i := \mu^\star - \max_{A\in\mathcal{B}:\, i\in A} \mu(A), \qquad E_+ := \{i : \Delta_i > 0\}, \qquad m_i := \mu_i + \Delta_i.$$

Looplessness guarantees that every arm belongs to some basis, so an arm lies in E_+ exactly when it belongs to no optimal basis. For any optimal basis $B^\star$ and $i \notin B^\star$,

$$m_i = \min_{j\in K_{B^\star}(i)} \mu_j.$$

Thus m_i does not depend on which optimal basis is chosen (Lemma C.1). At a unique optimum, $E_+ = E \setminus B^\star$ and every outside exchange has positive loss.

Let $I_i = \mathrm{kl}(\mu_i, m_i)$ for $i \in E_+$, where $\mathrm{kl}(p,u) = p\log(p/u) + (1-p)\log((1-p)/(1-u))$. Every uniformly efficient policy, even one that knows the matroid, satisfies

$$\liminf_{T\to\infty} \frac{R_\mu(T)}{\log T} \geq C^\star(\mu) := \sum_{i\in E_+} \frac{\Delta_i}{I_i}.$$

Arms with zero insertion loss do not contribute. Talebi & Proutière (2016, Theorem 2) prove the distinct-means case; Appendix D.3 extends it to arbitrary means. Raising only μ_i above m_i makes i belong to every optimal basis. A change of measure then forces at least $I_i^{-1}\log T$ observations to first order, while exchange matching charges at least Δ_i regret per observation.

3 The ECR-KL algorithm

ECR-KL alternates between *learning episodes* and *repair batches*. An episode starts with a *build*: oracle queries certify a basis and feasible exchanges, but not their mean comparisons. The learner tests these comparisons using the episode's observations. A trigger ends the episode, starts a fresh repair batch, and leads to a new priority and build. Every observation enters the global statistics; only builds query the oracle.

3.1 Certificate construction

The learner maintains ordered mean blocks $\mathcal{P} = (C_1, \ldots, C_b)$ and a priority σ that lists the blocks consecutively. The block order represents proposed comparisons between mean levels; comparisons within a block remain unresolved. Initially $\mathcal{P} = ([d])$ and $\sigma = (1, \ldots, d)$.

A greedy scan in order σ returns $B = (b_1, \ldots, b_r)$. For $i \notin B$, let $P_k = \{b_1, \ldots, b_k\}$. Dependence of $P_k \cup \{i\}$ is monotone in k. Both endpoints are known: $\{i\}$ is independent by looplessness, and $B \cup \{i\}$ is dependent by rank. Binary search locates the first dependent prefix:

$$k_i = \min\{k \in [r] : P_k \cup \{i\} \notin \mathcal{I}\}, \qquad j_\sigma(i) := b_{k_i} = \arg\max_{j\in K_B(i)} \mathrm{pos}_\sigma(j).$$

Thus $j_\sigma(i)$ is the feasible removal that comes last in σ. It need not have the smallest mean, since σ need not agree with the mean order. A build uses at most $c_{d,r} = d + (d-r)\lceil \log_2 r \rceil$ independence queries.

The stored certificate is

$$\mathcal{E}_\sigma = \{(b_{k+1}, b_k) : k < r\} \cup \{(i, j_\sigma(i)) : i \notin B\}.$$

An edge (a, b) represents $\mu_a \le \mu_b$: the basis edges form a chain, and each outside element points to one feasible exchange partner.

These partners define structural groups $\mathcal{G}_\sigma = (G_1, \dots, G_r)$, distinct from the mean blocks, and an action family

$$G_k = \{b_k\} \cup \{i \notin B : j_\sigma(i) = b_k\}, \qquad \mathcal{B}_\sigma = \{\{a_1, \dots, a_r\} : a_k \in G_k\}. \tag{2}$$

Lemma 1 (Certificate and partition reduction). *Every action in $\mathcal{B}_\sigma$ is a basis, regardless of the means. If every inequality in $\mathcal{E}_\sigma$ holds, then B is optimal and $j_\sigma(i)$ minimizes the mean over $K_B(i)$ for every $i \notin B$. Otherwise, either an adjacent basis pair is inverted, or the chain is correctly ordered and a stored exchange has positive true gain. Ties are allowed.*

The groups are successive differences of the flats spanned by basis prefixes, a standard partition reduction (Bérczi & Schwarcz, 2020, Lemma 10). Appendix C proves the lemma. A certificate is *correct* when its mean inequalities hold; correctness is not directly observable.

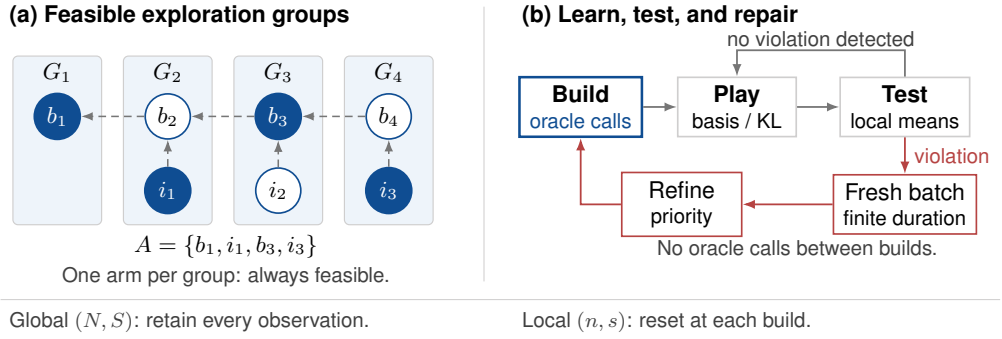


Figure 1: Structural groups and certificate repair ($r = 4$, $d = 7$). Each group holds one basis element and the outside arms pointing to it. Filled nodes mark one feasible action, which swaps in two outside arms. Dashed arrows $a \to b$ are the tested comparisons $\mu_a \le \mu_b$. Groups guarantee feasibility; correct comparisons also give optimal exploration thresholds. Only builds query the oracle.

3.2 Learning and local tests

Global counts and reward sums (N_i, S_i) include every observation. Local statistics (n_i, s_i) reset after each build and include only that episode's learning observations. The clock t counts all reward rounds, including covers and repairs. At action selection, (N_i, S_i) denotes $(N_i(t-1), S_i(t-1))$.

Let $E \setminus B = \{i_1 < \cdots < i_q\}$, where $q = d - r$, and set $v = q + 1$. The basis and partners determine the cover $\mathcal{C}_\sigma = (B, B - j_\sigma(i_1) + i_1, \dots, B - j_\sigma(i_q) + i_q)$. A *cover cycle* plays each action once and observes every coordinate. The initial cycle updates both global and local statistics. Repair cycles update only the global statistics and each supply one *designated* observation per coordinate: from B for basis elements, and from the corresponding exchange for each outside element.

The learner then alternates a forced play of B with an optimistic play. With $\bar{\mu}_i = S_i/N_i$, define

$$U_i(t) = \sup\{u \in [\bar{\mu}_i, 1] : N_i \,\mathrm{kl}(\bar{\mu}_i, u) \le f(t)\}, \quad f(t) = \log(t+1) + 6\log\log(\mathrm{e}(t+1)), \tag{3}$$

using $U_i(t) = 1$ when $N_i = 0$. On an optimistic round, select a maximum-index arm separately in each group, breaking ties by arm label:

$$a_k(t) \in \arg\max_{i \in G_k} U_i(t), \qquad A_t^{\mathrm{opt}} = \{a_1(t), \dots, a_r(t)\}. \tag{4}$$

Forced plays observe every basis element; optimistic plays can explore outside arms simultaneously. Lemma 1 guarantees feasibility without further queries.

Let ℓ_{ep} count learning rounds since the build, including the initial cover. For $\alpha = 1/(8d)$, set

$$\rho(n, \ell_{\mathrm{ep}}) = \sqrt{\frac{\log(4d\ell_{\mathrm{ep}}^6/\alpha)}{2n}}, \qquad \widehat{\mu}_i = s_i/n_i.$$

After the initial cover and each later learning action, check chain edges in basis order, then outside edges in increasing arm order. Trigger a repair at the first edge (a, b) satisfying

$$\widehat{\mu}_a - \widehat{\mu}_b > 2[\rho(n_a, \ell_{\mathrm{ep}}) + \rho(n_b, \ell_{\mathrm{ep}})]. \tag{5}$$

Tests use local means; indices use global estimates.

3.3 Refinement

At a trigger, set $g = (\widehat{\mu}_a - \widehat{\mu}_b)/2 > 0$, $\eta = 1/(8d)$, and

$$w = \frac{g}{4d}, \qquad m_{\mathrm{rep}} = \left\lceil \frac{8d^2}{g^2} \log \frac{2d}{\eta} \right\rceil. \tag{6}$$

The learner runs m_{rep} fresh cover cycles without queries and updates (N, S) with every observation. Each coordinate receives exactly m_{rep} designated samples, whose mean x_i defines the interval $J_i = [x_i - w, x_i + w]$.

Every certificate edge points to an earlier element in σ. A cross-block trigger may indicate an incorrect block order, so the learner resets $\mathcal{P}$ to $([d])$. Within each parent block, it connects coordinates whose intervals overlap and replaces the block by the connected components, ordered from high to low by their disjoint interval hulls. Parent blocks keep their relative order, and coordinates within a component keep their old priority; concatenation gives the next priority σ.

The repair ends after vm_{rep} rounds, even if no block splits or a false trigger involves equal means.

In Algorithm 1, $\mathrm{Observe}(A, 1)$ plays A, updates both global and local statistics, and increments t and ℓ_{ep}. $\mathrm{Observe}(A, 0)$, used during repairs, updates only (N, S) and t. Neither call changes ℓ, which counts learning rounds after the cover. Only Build queries the oracle; Appendix A gives the full routines. The arrays and interval workspace use $O(d)$ words; cover actions are reconstructed from the basis and partners. Refinement sorts d intervals in $O(d \log d)$ comparisons. An optimistic step computes d KL indices and r group maxima without independence queries.

4 Finite structural queries and optimal regret

Theorem 1 (ECR-KL)**.** *Under the assumptions of Section 2, let $L = |\{\mu_i : i \in [d]\}|$. Algorithm 1* (i) *is uniformly efficient and* (ii) *uses $O(d)$ words of memory.* (iii) *Its total expected query count satisfies*

$$\mathbb{E}Q_\infty \le \frac{4dL}{d + L + 1} c_{d,r} < 2dc_{d,r}. \tag{7}$$

(iv) *The expected time of its last rebuild, measured in reward rounds, is finite.* (v) *For every mean vector there is a finite instance-dependent A_μ such that*

$$R_\mu(T) \le C^\star(\mu) f(T) + A_\mu(\sqrt{f(T)} + 1), \qquad \lim_{T\to\infty} \frac{R_\mu(T)}{\log T} = C^\star(\mu). \tag{8}$$

If $E_+ = \varnothing$, then $\sup_T R_\mu(T) < \infty$. The algorithm requires neither L nor a positive reward gap as input.

The query bound is independent of the gaps between distinct mean levels, although small gaps can lengthen repairs.

Call a completed refinement *bad* if a local confidence bound fails at its trigger or some designated batch mean satisfies $|x_i - \mu_i| > w$. Otherwise, call it good.

Algorithm 1 ECR-KL: core policy with uniform repair

Initialize $t = 1$, $(N, S) = 0$, $\mathcal{P} = ([d])$, and $\sigma = (1, \dots, d)$.
loop
  $(B, j_\sigma, \mathcal{G}_\sigma, \mathcal{E}_\sigma) \leftarrow \mathrm{Build}(\sigma)$. {certify feasibility}
  Reset $(n, s) = 0$ and $\ell_{\mathrm{ep}} = \ell = 0$.
  for $A \in \mathcal{C}_\sigma$ in cover order **do** $\mathrm{Observe}(A, 1)$. {retain observations}
  $(a, b) \leftarrow \mathrm{Trig}(\mathcal{E}_\sigma, n, s, \ell_{\mathrm{ep}})$. {first violated test}
  while $(a, b) = \varnothing$ **do**
    $A \leftarrow B$ if ℓ is even; otherwise use A_t^{opt} from Equation 4.
    $\mathrm{Observe}(A, 1)$; $\ell \leftarrow \ell + 1$.
    $(a, b) \leftarrow \mathrm{Trig}(\mathcal{E}_\sigma, n, s, \ell_{\mathrm{ep}})$. {first violated test}
  end while
  $g \leftarrow (s_a/n_a - s_b/n_b)/2$.
  $(J_i)_{i \in E} \leftarrow \mathrm{Repair}(B, j_\sigma, g)$. {fresh finite batch}
  $(\mathcal{P}, \sigma) \leftarrow \mathrm{Refine}(\mathcal{P}, \sigma, a, b, (J_i)_{i \in E})$. {update blocks}
end loop

Lemma 2 (Refinement progress)**.** *Among the first R completed refinements, let H_{err} be the number of bad refinements. The $K_R = R + 1$ certificate builds satisfy*

$$K_R \le L + (3d - L - 1)H_{\mathrm{err}}. \tag{9}$$

A good refinement separates the triggering pair without splitting equal means within a block. Each good reset can be charged to a distinct earlier bad refinement. Counting the resulting block gains and losses proves Equation 9; in particular, an error-free run uses at most L builds (Appendix D.1).

Proof sketch of Theorem 1. Query count. Conditional on a reached build's history, the probability that its episode ends in a bad refinement is at most $p = \alpha + \eta = 1/(4d)$ (Lemma D.2). Let K count all builds. Taking expectations in Lemma 2 for the first $K \wedge M$ builds gives

$$\mathbb{E}(K \wedge M) \le L + (3d - L - 1)p\,\mathbb{E}(K \wedge M).$$

Rearranging and letting $M \to \infty$ gives $\mathbb{E}K \le 4dL/(d + L + 1)$; multiplying by $c_{d,r}$ then proves the query bound. Lemma D.3 also gives an exponential tail for K.

Stabilization. Forced basis plays detect an inverted chain edge. Otherwise, an incorrect certificate has an improving outside exchange. With accurate basis indices and no index underestimation, charge each optimistic round to a selected outside arm with largest index gain. Each charge increases its count, limiting nonimproving arms to $O_\mu(f(t))$ charges by time t, apart from concentration failures. An improving arm then receives enough local observations to trigger a repair (Lemma D.5).

Let W_T count incorrect learning rounds through T. Conditional on the build history, a correct certificate triggers at age ℓ_{ep} with probability $O(\ell_{\mathrm{ep}}^{-5})$; with $\mathbb{E}K < \infty$, completed correct episodes have integrable total duration. Each repair lasts at most a constant times its triggering episode. At the last incorrect round by T, all earlier correct episodes have ended. Summing the detection bounds there gives

$$W_T \le a_\mu K \log(\mathrm{e} + W_T) + \Xi_\mu, \qquad \mathbb{E}\Xi_\mu < \infty,$$

where Ξ_μ does not depend on T. Rearranging yields $W_T \le 2a_\mu K \log(\mathrm{e} + 2a_\mu K) + 2\Xi_\mu$, which is integrable because K has an exponential tail. Hence $\mathbb{E}W_\infty < \infty$. The same decomposition bounds repair and last-rebuild times (Lemma D.6). An incorrect final certificate would force $W_\infty = \infty$.

Regret. Incorrect episodes, repairs, initial covers, and index underestimation together contribute $O_\mu(1)$ total expected regret. On every other round, the correct certificate gives

$$\mu^\star - \mu(A_t) = \sum_{i \in A_t \cap E_+} \Delta_i.$$

A selected $i \in E_+$ on such a round satisfies $U_i(t) \ge U_{j(i)}(t) \ge m_i$. Lemma D.7 bounds their count $M_i(T)$ by

$$\mathbb{E}M_i(T) \le \frac{f(T)}{I_i} + O_\mu(\sqrt{f(T)} + 1).$$

Since m_i is the same for every optimal basis and each selection increases the global count, this bound applies once per arm across episodes. Summing $\Delta_i \mathbb{E}M_i(T)$ gives the upper bound and uniform efficiency; if $E_+ = \varnothing$, regret is bounded. Lemma D.8 supplies the matching lower bound. □

Observable completion (Proposition 1). ECR-KL does not announce its last rebuild. A universally feasible policy, consistent on every instance, can declare at a finite time that no further queries will be made only after identifying the entire matroid without error. On a hidden uniform matroid with $2 \le r < d$, this requires at least $\binom{d}{r}$ queries (Appendix E.3).

5 Structural lower bound and distinct means

Theorem 2 (Allocation reveals hidden structure)**.** *Let $2 \le r < d$. Every universally feasible, uniformly efficient policy attaining $C^\star(\mu)$ at every unique-optimum instance satisfies*

$$\mathbb{E}Q_\infty \ge (d-r)\log_2 r$$

on some loopless rank-r instance. The hard family can share a common mean vector and a common unique optimal basis. The statement also holds when all coordinate means are required to be distinct.

Consider two capacity-one groups with anchors of means $\nu_1 < \nu_2$. One outside element with mean $p < \nu_1$ belongs to either group. The anchors form the optimal basis in both cases, but the outside element's optimal sampling rate is $1/\operatorname{kl}(p, \nu_1)$ or $1/\operatorname{kl}(p, \nu_2)$, depending on which anchor it can replace. Knowing the means and optimal basis is therefore not enough to explore optimally.

Proof sketch of Theorem 2. Take r anchors with means $\nu_1 < \cdots < \nu_r$ and $q = d - r$ outside elements with distinct means $p_i < \nu_1$. A label $\theta_i \in [r]$ assigns i to anchor θ_i's capacity-one group. All r^q instances share the same means and optimal basis, but

$$R_\theta(T) = \sum_i (\nu_{\theta_i} - p_i)\mathbb{E}_\theta N_i(T).$$

Raising p_i above ν_{θ_i} makes every basis omitting i suboptimal. A change of measure and uniform efficiency show that the probability of undersampling by any fixed fraction tends to zero. The normalized shortfall is bounded, so its expectation also vanishes. Since every regret weight $\nu_{\theta_i} - p_i$ is positive, attaining $C^\star$ forces the expected excess to vanish as well (Lemma E.1). Thus

$$\frac{N_i(T)}{\log T} \xrightarrow{L^1} \frac{1}{\operatorname{kl}(p_i, \nu_{\theta_i})}.$$

The limits are distinct across anchor labels, so a decoder given the common means recovers θ with vanishing error.

Let Θ be uniform on $[r]^q$ and let Ω, independent of Θ, contain both common reward streams and policy randomness. Given Ω, runs with the same answer prefix agree until their next query. The terminal answer strings W_T at horizon T are therefore prefix-free. Lemma E.2 and vanishing decoding error give, with information in bits,

$$q\log_2 r - o(1) \le I(\Theta; \Omega, \mathsf{W}_T) \le \mathbb{E}Q_T.$$

Letting $T \to \infty$ proves the bound on the family average, and hence on at least one instance. □

Distinct means. Full-Order ECR-KL retains the cover, action rule, and local test, with $\alpha = \eta = 1/8$. After a trigger, it collects fresh cover cycles until all coordinate confidence intervals are pairwise disjoint, then sorts the coordinates and rebuilds. These samples enter the global statistics, not the next episode's local tests.

Theorem 3 (Pairwise-distinct means)**.** *If all coordinate means are distinct, Full-Order ECR-KL uses $O(d)$ words of memory, has finite expected last-rebuild time, satisfies Equation 8, and obeys*

$$\mathbb{E}Q_\infty \le \frac{113}{49} c_{d,r}.$$

The optimal worst-case expected query complexity is therefore $\Theta(d \log d)$ on this submodel at balanced rank.

Proof sketch of Theorem 3. At a reached ordering phase, the returned order is wrong with conditional probability at most η. A correct order gives a correct certificate, whose fresh tests trigger with probability at most α. Thus another phase is reached with probability at most p_{next}, giving

$$p_{\mathrm{next}} = \eta + (1-\eta)\alpha = 15/64, \qquad \mathbb{E}K \le 1 + (1-p_{\mathrm{next}})^{-1} = 113/49,$$

where K counts builds. Successive conditioning gives K a geometric tail. The uniform conditional duration bound in Lemma F.1 then makes the total ordering time integrable. Combining this bound with the detection and retained-count estimates proves stabilization and regret (Appendix F). □

At balanced rank, ECR-KL uses $O(Ld\log d)$ expected queries (Table 4). A known L permits a constant expected number of builds (Appendix F). With unknown L and arbitrary ties, the general bounds remain $\Omega(d\log d)$ and $O(d^2\log d)$; exact recognition matches the former on matroids with bounded-size components (Appendix G.3).

6 Experiments

We compare the regret and query costs of practical ECR-KL variants, then examine how parallel exploration and scheduled rebuilds affect those costs. These are finite-horizon comparisons, not tests of asymptotic convergence. The core repair can itself exceed the evaluation horizon: at $d = 8, r = 4, g = .1$, one uniform batch takes 1 774 460 rounds by Equation 6.

Settings and protocol. Suite A has eight partition and graphic instances (Appendix H.3); Suite B has twelve binary-linear, capacity-two partition, and sparse-paving instances; Suite C has four exact-tie instances. The refresh study adds 16 connected graphic and binary-linear instances. These comparisons use $T = 10^6$, eight shared-stream seeds, and Equation 3. Practical ECR-KL uses groupwise learning after capacity-one recognition, or a finite prefix and bounded direct rebuilds if recognition fails. Refreshes follow rounds 8192, 32768, and 131072; the schedule was chosen from earlier results, before observing the refresh-study rewards. KL-greedy uses matched indices and caches. All reward rounds and queries are counted; intervals use eight seed-level instance averages. Appendix H.2 gives the protocols and post-hoc shared component-recognition accounting.

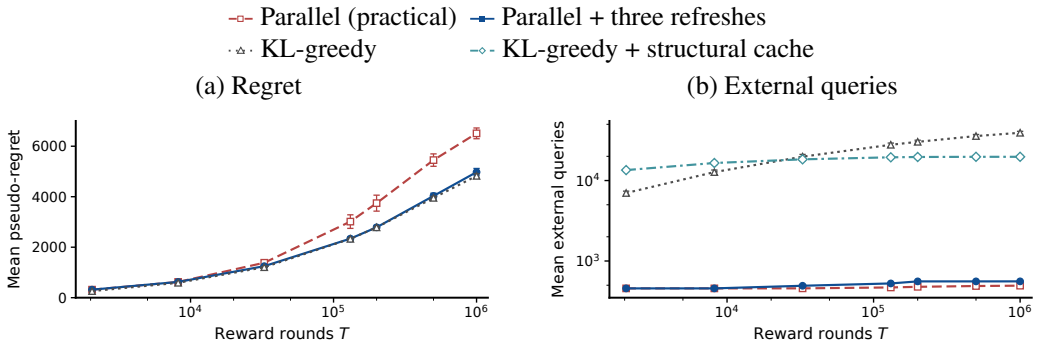


Figure 2: Regret and query costs on the same 16 connected instances in the refresh study. Query counts include failed shared recognition, initialization, and all rebuilds; the query axis is logarithmic. Error bars are unadjusted 95% t-intervals over eight seed-level instance averages. The two greedy variants have identical regret.

Results. With shared recognition, practical parallel ECR-KL uses 514 versus 114 101 external calls on graphic instances—about a 222-fold reduction—at the cost of 25.5% higher regret (Table 2). On binary-linear instances, 465 versus 10 957 calls gives roughly a 24-fold reduction, with 21.6% higher regret. Against structural-cache greedy, the reductions are 16.5-fold and 40.3-fold, respectively. These savings disappear on fully recognized families. Structural caching can increase calls: failed verification requires a full basis-and-partner build rather than a greedy scan (Appendix G.2).

On Suites A and B, parallel exploration has lower mean regret than serial, but not matched KL-greedy (Appendix H.2). In the connected study, three refreshes reduce regret by 23.6% (-1536 ± 276) for

Table 2: Shared recognition at $T = 10^6$ (four instances and eight seeds per family). Regret is the parallel-to-greedy ratio of mean regrets; queries are mean external calls, including recognition and subsequent calls through a 16-entry cache. "Identified" counts fully recognized structures; "gr+str" adds structural caching. Post-hoc accounting preserves reward trajectories. Table 6 gives intervals.

Family	Identified	$R_{\mathrm{par}}/R_{\mathrm{gr}}$	Q_{par}	Q_{gr}	$Q_{\mathrm{gr+str}}$
Partition (capacity one)	$4/4$	1.000	100	100	100
Graphic	$0/4$	1.255	514	114 101	8 504
Binary linear	$0/4$	1.216	465	10 957	18 740
Partition (capacity two)	$4/4$	1.138	153	153	153
Sparse-paving direct sum	$4/4$	1.208	400	400	400

65 ± 3 extra calls. The refreshed policy makes 558 calls versus greedy's $39\,224$ and structural-cache greedy's $19\,725$, with 3.1% higher regret than either baseline (Figure 2).

No main-comparison run enters ordinary repair or uses more than one direct-rebuild credit. In Suite B, all 96 parallel runs store an optimal basis, but only $57/96$ have correct mean certificates. Thus the observed query plateaus should not be read as evidence that certificate repair has finished. Appendices H.3 and H.4 examine core repair and ordering separately.

7 Relation to prior work

Known-matroid asymptotics build on change-of-measure lower bounds (Lai & Robbins, 1985; Graves & Lai, 1997) and KL indices (Garivier & Cappé, 2011). General combinatorial semi-bandits are studied by Gai et al. (2012); Chen et al. (2013); Combes et al. (2015); Cuvelier et al. (2021); KL-OSM gives the matroid benchmark (Talebi & Proutière, 2016), while MAUB reduces membership-oracle update rounds using exchange partners (Delage & Gaudel, 2025; Combes & Proutière, 2014).

Structural access connects to exact learning and matroid-oracle complexity (Angluin, 1988; Coullard & Hellerstein, 1996; Karp et al., 1988; Khanna et al., 2026; Banihashem et al., 2026). Batched and switching bandits restrict adaptivity rather than feasibility queries (Perchet et al., 2016; Gao et al., 2019; Simchi-Levi & Xu, 2019). Other computational approaches include structured confidence regions, dynamic maintenance, and oracle-efficient semi-bandits (Perrault et al., 2019; Tzeng et al., 2024; Kim et al., 2025; Blikstad et al., 2023).

8 Limitations and conclusion

Unlike the query bound, the transient bound depends on gaps between distinct mean levels, not only positive insertion losses, even with parallel exploration and packing. A factor-d query gap remains at balanced rank for unknown L with ties. The experiments do not establish finite-horizon superiority over KL-greedy. Our assumptions include a noiseless independence oracle, independent Bernoulli rewards with means in $(0, 1)$, exact KL indices, and memory measured in words rather than bits.

ECR-KL attains $C^\star$ at every mean vector with finitely many structural queries in expectation. The lower bound shows why finding an optimal basis is not enough: identical means and optimal bases can still require different optimal sampling rates.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval, manuscript organization and editing, proof checking, experimental design, and construction and checking of numerical diagnostics. The authors checked cited references against their sources and manually reviewed every AI-assisted claim, derivation, proof, and numerical conclusion against the stated assumptions and definitions. Numerical diagnostics supplemented rather than replaced the mathematical arguments. The authors take full responsibility for all text, claims, proofs, experiments, and reported results.

REPRODUCIBILITY STATEMENT

Appendix A gives the core subroutines; Appendices B–G contain the proofs and algorithm variants. Appendix H specifies the experimental instances, policy parameters, initialization rules, horizons, random-stream coupling, and uncertainty estimates. Appendix H.1 describes the structural and numerical checks.

REFERENCES

Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

Kiarash Banihashem, MohammadTaghi Hajiaghayi, Mahdi JafariRaviz, and Danny Mittal. Matroid algorithms under size-sensitive independence oracles. arXiv preprint arXiv:2605.00201, 2026. URL https://arxiv.org/abs/2605.00201.

Kristóf Bérczi and Tamás Schwarcz. Rainbow and monochromatic circuits and cuts in binary matroids. arXiv preprint arXiv:2012.05037v1, 2020. URL https://arxiv.org/abs/2012.05037v1.

Joakim Blikstad, Sagnik Mukhopadhyay, Danupon Nanongkai, and Ta-Wei Tu. Fast algorithms via dynamic-oracle matroids. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1229–1242. Association for Computing Machinery, 2023. doi: 10.1145/3564246.3585219.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 151–159. PMLR, 2013. URL https://proceedings.mlr.press/v28/chen13a.html.

Richard Combes and Alexandre Proutière. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 521–529. PMLR, 2014. URL https://proceedings.mlr.press/v32/combes14.html.

Richard Combes, Mohammad Sadegh Talebi, Alexandre Proutière, and Marc Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Collette R. Coullard and Lisa Hellerstein. Independence and port oracles for matroids, with an application to computational learning theory. *Combinatorica*, 16(2):189–208, 1996. doi: 10.1007/BF01844845.

Thibaut Cuvelier, Richard Combes, and Eric Gourdin. Asymptotically optimal strategies for combinatorial semi-bandits in polynomial time. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 505–528. PMLR, 2021. URL https://proceedings.mlr.press/v132/cuvelier21a.html.

Aurélien Delage and Romaric Gaudel. Efficient matroid bandit linear optimization leveraging unimodality. arXiv preprint arXiv:2512.00605, 2025. URL https://arxiv.org/abs/2512.00605.

Jack Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1:127–136, 1971. doi: 10.1007/BF01584082.

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

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 359–376. PMLR, 2011. URL https://proceedings.mlr.press/v19/garivier11a.html.

Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997. doi: 10.1137/S0363012994275440.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Xiaowei Huang, Shiguang Feng, and Lvzhou Li. Quantum and classical query complexities for determining connectedness of matroids. arXiv preprint arXiv:2306.12103v1, 2023. URL https://arxiv.org/abs/2306.12103v1.

Xiaowei Huang, Jingquan Luo, and Lvzhou Li. Quantum speedup and limitations on matroid property problems. *Frontiers of Computer Science*, 18(4):184905, 2024. doi: 10.1007/s11704-023-3130-9.

Per M. Jensen and Bernhard Korte. Complexity of matroid property algorithms. *SIAM Journal on Computing*, 11(1):184–190, 1982. doi: 10.1137/0211014.

Richard M. Karp, Eli Upfal, and Avi Wigderson. The complexity of parallel search. *Journal of Computer and System Sciences*, 36(2):225–253, 1988. doi: 10.1016/0022-0000(88)90027-X.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://www.jmlr.org/papers/v17/kaufman16a.html.

Sanjeev Khanna, Aaron Putterman, and Junkai Song. Optimal parallel basis finding in graphic and related matroids. In *53rd International Colloquium on Automata, Languages, and Programming (ICALP 2026)*, volume 374 of *Leibniz International Proceedings in Informatics*, pp. 125:1–125:22. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi: 10.4230/LIPIcs.ICALP.2026.125.

Jung-hun Kim, Milan Vojnović, and Min-hwan Oh. Oracle-efficient combinatorial semi-bandits. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28269–28304. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-0840. URL https://papers.nips.cc/paper_files/paper/2025/hash/237196fdf3472b8de08bdd1103bb7b67-Abstract-Conference.html.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pp. 420–429. AUAI Press, 2014. URL https://auai.org/uai2014/proceedings/individuals/161.pdf.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 535–543. PMLR, 2015. URL https://proceedings.mlr.press/v38/kveton15.html.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. doi: 10.1016/0196-8858(85)90002-8.

James G. Oxley. *Matroid Theory*. Oxford Graduate Texts in Mathematics. Oxford University Press, Oxford, 2 edition, 2011. ISBN 978-0-19-960339-8.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016. doi: 10.1214/15-AOS1381.

Pierre Perrault, Vianney Perchet, and Michal Valko. Exploiting structure of uncertainty for efficient matroid semi-bandits. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5123–5132. PMLR, 2019. URL https://proceedings.mlr.press/v97/perrault19a.html.

David Simchi-Levi and Yunzong Xu. Phase transitions and cyclic phenomena in bandits with switching constraints. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mohammad Sadegh Talebi and Alexandre Proutière. An optimal algorithm for stochastic matroid bandit optimization. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 548–556. International Foundation for Autonomous Agents and Multiagent Systems, 2016. URL https://www.ifaamas.org/Proceedings/aamas2016/pdfs/p548.pdf.

Ruo-Chun Tzeng, Naoto Ohsaka, and Kaito Ariu. Matroid semi-bandits in sublinear time. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48855–48877. PMLR, 2024. URL https://proceedings.mlr.press/v235/tzeng24a.html.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

APPENDIX CONTENTS

Algorithm definitions are in Appendix A; the proofs of Theorems 1, 2, and 3 are in Appendices D, E, and F, respectively. Appendices G and H contain the variants and experiments.

A Algorithmic details . . . 14
- A.1 State variables . . . 14
- A.2 Builds and actions . . . 15
- A.3 Tests and repair samples . . . 16
- A.4 Mean refinement . . . 17

B Preliminaries . . . 18

C Certificates and insertion losses . . . 19

D Proof of Theorem 1 . . . 20
- D.1 Build count . . . 20
- D.2 Stabilization . . . 22
- D.3 Regret . . . 27

E Proof of Theorem 2 (Query lower bounds) . . . 28
- E.1 Sampling rates . . . 29
- E.2 Oracle information . . . 30
- E.3 Observable completion . . . 31

F Proof of Theorem 3 . . . 32

G Algorithm variants . . . 34
- G.1 Sampling . . . 34
- G.2 Rebuilds . . . 37
- G.3 Recognition . . . 40

H Experimental details . . . 42
- H.1 Setup . . . 42
- H.2 Parallel exploration and refreshes . . . 43
- H.3 Initialization and repair . . . 46
- H.4 Full-Order ECR-KL . . . 51

A ALGORITHMIC DETAILS

The subroutines below implement Algorithm 1 with uniform repair. Full-Order and the variants in Appendices F–G are specified separately.

A.1 STATE VARIABLES

The ordered blocks $\mathcal{P}$ record which mean comparisons remain unresolved. The groups $\mathcal{G}_\sigma$, in contrast, certify feasible combinations of arms. Refinement updates the blocks after a repair; a build recomputes the groups from the resulting priority. Neither partition is assumed to describe the full matroid.

Table 3: Notation and core algorithm state. Stored arrays have total length $O(d)$; the final row contains analysis-only quantities.

State	Meaning and update rule
$E, d, r, \mathcal{M}, \mathcal{B}$	Ground set, number of arms, rank, matroid, and its family of bases.
$\mu_i, \mu^\star, R_\mu(T)$	Arm mean, optimal basis reward, and expected pseudo-regret.
$B, B^\star, K_B(i), j_\sigma(i)$	Stored basis, optimal basis, feasible removals, and the stored exchange partner.
Δ_i, m_i, E_+	Insertion loss, threshold $m_i = \mu_i + \Delta_i$, and arms with positive insertion loss.
$I_i, C^\star(\mu)$	$I_i = \mathrm{kl}(\mu_i, m_i)$ and the optimal logarithmic regret coefficient.
$Q_T, Q_\infty, c_{d,r}$	Calls by round T, total calls, and the per-build query bound.
$t, (N_i, S_i)_{i\in E}$	Next reward round and global counts and sums. Initially $t = 1$ and $(N, S) = 0$; every observation is retained.
$\ell_{\mathrm{ep}}, (n_i, s_i)_{i\in E}$	Episode age and local counts and sums. Reset at a build; include the initial cover and subsequent learning rounds, but not repairs.
ℓ	Number of learning rounds after the initial cover. Reset to zero at a build. Even values select B; odd values select A_t^{opt}.
$\mathcal{P}, \sigma$	Ordered mean blocks and a priority listing each block consecutively. Refinement preserves the old priority within each new block.
$\mathcal{G}_\sigma, \mathcal{B}_\sigma, \mathcal{E}_\sigma$	Structural groups, their feasible transversals, and mean-comparison edges. Recomputed at a build.
g, w, m_{rep}	Trigger scale, interval radius, and number of uniform-repair cover cycles.
$L, K, M_i(T), \Xi_\mu$	Number of mean levels, total builds, nonexceptional selections of arm i, and an integrable random remainder in the stabilization proof.

In the routines below, $\mathrm{Observe}(A, \lambda)$ plays the certified action A at round t and returns $X = (X_i)_{i\in A}$, where $X_i = X_{t,i}$. It performs the following updates exactly once:

$$(N_i, S_i) \leftarrow (N_i + 1, S_i + X_i) \quad (i \in A), \qquad t \leftarrow t + 1.$$

If $\lambda = 1$, it also updates $(n_i, s_i) \leftarrow (n_i + 1, s_i + X_i)$ for $i \in A$ and $\ell_{\mathrm{ep}} \leftarrow \ell_{\mathrm{ep}} + 1$. If $\lambda = 0$, the local state is unchanged. Unselected coordinates are unchanged in either case. This routine makes no oracle calls and does not change ℓ.

For the initial cover, call $\mathrm{Observe}(A, 1)$ once for each action of $\mathcal{C}_\sigma$, in the stated order. This leaves $\ell_{\mathrm{ep}} = v$ and every $n_i \geq 1$. The first test runs only after this cycle is complete. Each later learning round also uses $\lambda = 1$ and increments ℓ once. Repair observations use $\lambda = 0$, and no test runs during a repair. The build that follows a repair resets $(n, s), \ell_{\mathrm{ep}}, \ell$, but leaves $(N, S), t$ untouched.

A.2 BUILDS AND ACTIONS

Let $\mathcal{O}(C) = \mathbf{1}\{C \in \mathcal{I}\}$ denote a counted independence call. In Algorithm 2, B is an ordered list; set expressions use its underlying set. The prefix P_k means its first k elements, reconstructed when needed. The prefixes and cover actions are not stored as separate copies.

Algorithm 2 $\mathrm{Build}(\sigma)$

Require: Priority permutation σ, rank r, and independence oracle $\mathcal{O}$.
Ensure: Ordered basis B, partners j_σ, groups $\mathcal{G}_\sigma$, and ordered edges $\mathcal{E}_\sigma$.

$B \leftarrow ()$; $p \leftarrow 1$.
while $|B| < r$ **do**
  $i \leftarrow \sigma_p$; $p \leftarrow p + 1$.
  if $\mathcal{O}(B \cup \{i\}) = 1$ **then**
    Append i to B.
  end if
end while
Write $B = (b_1, \ldots, b_r)$; set $G_k \leftarrow \{b_k\}$ for each $k \in [r]$.
for $i \in E \setminus B$ in increasing arm order **do**
  $\mathrm{lo} \leftarrow 1$; $\mathrm{hi} \leftarrow r$.
  while $\mathrm{lo} < \mathrm{hi}$ **do**
    $k \leftarrow \lfloor (\mathrm{lo} + \mathrm{hi})/2 \rfloor$.
    if $\mathcal{O}(P_k \cup \{i\}) = 1$ **then**
      $\mathrm{lo} \leftarrow k + 1$.
    else
      $\mathrm{hi} \leftarrow k$.
    end if
  end while
  $j_\sigma(i) \leftarrow b_{\mathrm{lo}}$; $G_{\mathrm{lo}} \leftarrow G_{\mathrm{lo}} \cup \{i\}$.
end for
$\mathcal{G}_\sigma \leftarrow (G_1, \ldots, G_r)$.
$\mathcal{E}_\sigma \leftarrow ((b_{k+1}, b_k))_{k=1}^{r-1}$, followed by $((i, j_\sigma(i)))_{i \in E \setminus B}$ in increasing i.
return $(B, j_\sigma, \mathcal{G}_\sigma, \mathcal{E}_\sigma)$.

The binary-search invariant is $k_i \in [\mathrm{lo}, \mathrm{hi}]$: $P_{\mathrm{lo}-1} + i$ is independent and $P_{\mathrm{hi}} + i$ is dependent. Initially these facts follow from looplessness and rank. At $r = 1$ the search makes no calls. The greedy scan uses at most d calls, and each search at most $\lceil \log_2 r \rceil$. The routine reads no rewards and does not change the reward clock or statistical arrays.

Algorithm 3 $\mathrm{OptimisticAction}(\mathcal{G}_\sigma, N, S, t)$

Require: Certified groups and global statistics before round t.
Ensure: An action in $\mathcal{B}_\sigma$; no state changes and no oracle calls.

for $i = 1, \ldots, d$ **do**
  Compute $U_i(t)$ by Equation 3; use $U_i(t) = 1$ if $N_i = 0$.
end for
for $k = 1, \ldots, r$ **do**
  $a_k \leftarrow \min \arg\max_{i \in G_k} U_i(t)$.
end for
return $\{a_1, \ldots, a_r\}$.

An outside arm may be selected when its index equals its representative's index; the smaller label is selected. This differs from the strictly positive-gain requirement in the serial variant, but satisfies the weak index inequality used in the regret proof.

A.3 TESTS AND REPAIR SAMPLES

The edge list passed to Trig has the order returned by Build. Its first $r-1$ entries test the basis chain from b_1 to b_r; the rest test outside arms in increasing label order. At rank one the chain part is empty.

Algorithm 4 $\mathrm{Trig}(\mathcal{E}_\sigma, n, s, \ell_{\mathrm{ep}})$

Require: Completed initial cover: $\ell_{\mathrm{ep}} \geq v$ and $n_i \geq 1$ for every i; $\alpha = 1/(8d)$.
Ensure: First edge satisfying the trigger test, or $\varnothing$; no state changes.

For each i, set $\widehat{\mu}_i \leftarrow s_i/n_i$ and $\rho_i \leftarrow \sqrt{\log(4d\ell_{\mathrm{ep}}^6/\alpha)/(2n_i)}$.
for (a, b) in the ordered list $\mathcal{E}_\sigma$ **do**
  if $\widehat{\mu}_a - \widehat{\mu}_b > 2(\rho_a + \rho_b)$ **then**
    return (a, b).
  end if
end for
return $\varnothing$.

The budgets $\alpha = \eta = 1/(8d)$ give a bad-refinement probability of at most $1/(4d)$ per reached build (Lemma D.2). With $w = g/(4d)$, an overlap path spans at most $2dw = g/2$, too little to connect the triggering pair on the confidence event. The choice of m_{rep} bounds the batch error by η. Separately, the coefficient 6 in Equation 3 makes index underestimation summable (Lemma D.4); these constants are not optimized.

At a trigger, the local estimates determine $g = (s_a/n_a - s_b/n_b)/2 > 0$. The following routine keeps the current basis and partners fixed throughout the batch. It calls Observe with local updates disabled. The temporary sums z_i contain only designated observations and are distinct from both global sums S_i and episode-local sums s_i.

Algorithm 5 $\mathrm{Repair}(B, j_\sigma, g)$: uniform batch

Require: Current certified basis and partners, $g > 0$, and $\eta = 1/(8d)$.
Ensure: Intervals $J_i = [x_i - w, x_i + w]$; global statistics and t include every batch observation.

$w \leftarrow g/(4d)$; $m_{\mathrm{rep}} \leftarrow \lceil 8d^2 g^{-2} \log(2d/\eta) \rceil$; $z_i \leftarrow 0$ for every i.
for $h = 1, \ldots, m_{\mathrm{rep}}$ **do**
  $X \leftarrow \mathrm{Observe}(B, 0)$; $z_j \leftarrow z_j + X_j$ for every $j \in B$.
  for $i \in E \setminus B$ in increasing arm order **do**
    $X \leftarrow \mathrm{Observe}(B - j_\sigma(i) + i, 0)$; $z_i \leftarrow z_i + X_i$.
  end for
end for
For each i, set $x_i \leftarrow z_i/m_{\mathrm{rep}}$ and $J_i \leftarrow [x_i - w, x_i + w]$.
return $(J_i)_{i \in E}$.

Each cycle adds one designated observation to every z_i and advances the reward clock by exactly v. Incidental basis observations during outside exchanges still enter (N, S), but not z. Thus the duration is exactly vm_{rep}, including after a false trigger, and the routine makes no oracle queries. Algorithm 1 passes these intervals to Refine before the next build.

A.4 MEAN REFINEMENT

The interval-overlap components can be found by a sweep, without storing a graph. Write $J_i = [l_i, u_i]$. Sort each parent block by l_i, breaking equal lower endpoints by the old priority. Keep the largest upper endpoint in the current component. An interval starts a new component exactly when its lower endpoint exceeds that upper endpoint; touching closed intervals remain connected.

Algorithm 6 $\mathrm{Refine}(\mathcal{P}, \sigma, a, b, (J_i)_{i \in E})$

Require: Ordered blocks, a compatible priority, the triggering edge, and batch intervals.
Ensure: Refined ordered blocks $\mathcal{P}'$ and a compatible priority σ'.

if a, b belong to different blocks of $\mathcal{P}$ **then**
  $\mathcal{P} \leftarrow (E)$.
end if
$\mathcal{P}' \leftarrow ()$.
for each parent block C of $\mathcal{P}$, in its existing order **do**
  Sort C by $(l_i, \mathrm{pos}_\sigma(i))$; set $\mathcal{D} \leftarrow ()$, $D \leftarrow \varnothing$, $u \leftarrow -\infty$.
  for i in this sorted order **do**
    if $D \neq \varnothing$ and $l_i > u$ **then**
      Append D to $\mathcal{D}$; set $D \leftarrow \varnothing$, $u \leftarrow -\infty$.
    end if
    $D \leftarrow D \cup \{i\}$; $u \leftarrow \max\{u, u_i\}$.
  end for
  Append D to $\mathcal{D}$.
  Append the blocks of $\mathcal{D}$ to $\mathcal{P}'$ in reverse order.
end for
Assign each arm its block index in $\mathcal{P}'$.
$\sigma' \leftarrow$ a stable sort of σ by these block indices.
return $(\mathcal{P}', \sigma')$.

The sweep produces components from low to high, and reversing them orders their disjoint hulls from high to low. Parent blocks keep their order, and the final stable sort preserves the old priority within each component, exactly as the refinement rule requires. For example, starting from $\mathcal{P} = (E)$ and $\sigma = (1, 2, 3, 4)$, the intervals $[.2, .3]$, $[.25, .35]$, $[.6, .7]$, and $[.65, .75]$ give $\mathcal{P}' = (\{3, 4\}, \{1, 2\})$ and $\sigma' = (3, 4, 1, 2)$.

The routine returns even when no block splits. It does not use the true means or test certificate correctness directly. Sorting costs $O(d \log d)$ comparisons in total, and the interval endpoints, block labels, and workspace all use $O(d)$ words. Only the next build can issue new structural queries.

Optional variants. Serial ECR-KL changes only $\mathrm{OptimisticAction}$ (Appendix G.1). Full-Order sets $\alpha = \eta = 1/8$ and changes the batch stopping rule (Appendix F); targeted refinement changes its prescribed counts and adds a finite fallback (Appendix G.1). Packed sampling changes how the prescribed observations are scheduled, not how interval components are formed (Appendix G.1). A direct rebuild bypasses the batch entirely and resets to one unresolved block; its credit and global-test rules appear in Appendix G.2. Direct rebuilds and finite prefixes have the additional query budgets stated in Appendix G.2; changes to the sampling schedule alone retain the core query bound.

B PRELIMINARIES

All expectations concern a fixed instance unless a prior on matroids is specified. For $S \subseteq E$, write $S+i = S \cup \{i\}$ and $r(S) = \max\{|I| : I \subseteq S,\ I \in \mathcal{I}\}$. The closure is $\mathrm{cl}(S) = \{i : r(S+i) = r(S)\}$; a flat is a set equal to its closure.

A circuit is an inclusion-wise minimal dependent set. For a basis B and $i \notin B$, the fundamental circuit $C_B(i)$ is the unique circuit in $B + i$, and $K_B(i) = C_B(i) \setminus \{i\}$ (Talebi & Proutière, 2016, Section 3.1).

Lemma B.1 (Basis-exchange matching; Talebi & Proutière, 2016, Lemma 1)**.** *For any bases A, B, there is a bijection $\phi : A \to B$ such that $\phi(i) = i$ for $i \in A \cap B$ and $B - \phi(i) + i \in \mathcal{B}$ for $i \in A \setminus B$.*

Proof. Put $I = A \cap B$, and form a bipartite graph on $A \setminus B$ and $B \setminus A$, joining i to j exactly when $j \in K_B(i)$.

For $S \subseteq A \setminus B$, let $N(S)$ be its neighborhood. Suppose $|N(S)| < |S|$. Then $I \cup N(S)$ and $I \cup S$ are independent, with the first strictly smaller. Augmentation gives an $i \in S$ such that $I \cup N(S) + i$ is independent. But $C_B(i) \setminus \{i\} \subseteq I \cup N(S)$, since every element outside I is adjacent to i. The enlarged set contains $C_B(i)$, a contradiction.

Thus $|N(S)| \geq |S|$ for every S. Hall's theorem gives a matching that covers $A \setminus B$. Since $|A \setminus B| = |B \setminus A|$, this matching is a bijection. Extending it by $\phi(i) = i$ on I gives the required map. □

Lemma B.2 (Change of measure; Kaufmann et al., 2016, Lemmas 1 and 18)**.** *Fix a matroid and a predictable policy. Let $P_\mu^T, P_{\mu'}^T$ be its laws of the full interaction history at a deterministic horizon T, for Bernoulli mean vectors $\mu, \mu' \in (0,1)^d$. Histories include oracle queries and answers and the policy randomness revealed by T. For coordinate i's nth observation $X_{i,n}$, set*

$$\ell_T = \sum_{i=1}^{d} \sum_{n=1}^{N_i(T)} \log \frac{\mu_i^{X_{i,n}}(1-\mu_i)^{1-X_{i,n}}}{(\mu_i')^{X_{i,n}}(1-\mu_i')^{1-X_{i,n}}}.$$

For every history-measurable event E,

$$\mathbb{P}_\mu(E) = \mathbb{E}_{\mu'}[\mathbf{1}_E e^{\ell_T}],$$

$$\mathrm{KL}(P_\mu^T, P_{\mu'}^T) = \sum_{i=1}^{d} \mathbb{E}_\mu N_i(T)\, \mathrm{kl}(\mu_i, \mu_i') \geq \mathrm{kl}\big(\mathbb{P}_\mu(E), \mathbb{P}_{\mu'}(E)\big).$$

Proof. The two laws use the same matroid and the same policy. Thus oracle answers, conditional action and query rules, and policy randomness contribute identical factors to the two likelihoods. Only the selected rewards contribute to their ratio.

Reveal the selected rewards in a fixed coordinate order within each round. Multiplying their Bernoulli likelihood ratios gives

$$\frac{dP_\mu^T}{dP_{\mu'}^T} = e^{\ell_T}.$$

This ratio is positive and finite: there are at most rT reward factors and all means belong to $(0,1)$. The number of queries may be random, but it is finite at a finite horizon and contributes no additional likelihood factor. Integrating over E proves the first identity.

Since A_t is chosen before the round-t rewards, the conditional expectation of each reward log-likelihood factor is its Bernoulli KL divergence. Hence

$$\mathbb{E}_\mu \ell_T = \sum_{t=1}^{T} \mathbb{E}_\mu \sum_{i \in A_t} \mathrm{kl}(\mu_i, \mu_i') = \sum_{i=1}^{d} \mathbb{E}_\mu N_i(T)\, \mathrm{kl}(\mu_i, \mu_i').$$

All sums are finite. The left-hand side is $\mathrm{KL}(P_\mu^T, P_{\mu'}^T)$. Applying data processing to the indicator of E gives the binary-KL lower bound. □

Concentration tools. We use Hoeffding's inequality (Hoeffding, 1963, Theorem 2, Eq. (2.6)) in the following form. For independent X_j with deterministic bounds $a_j \le X_j \le b_j$ and $\sum_j (b_j - a_j)^2 > 0$,

$$\mathbb{P}\left(\sum_{j=1}^{n}(X_j - \mathbb{E}X_j) \ge u\right) \le \exp\left(-\frac{2u^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right), \qquad u > 0.$$

The lower tail has the same bound. In particular, independent $[0,1]$-valued observations with common mean μ satisfy

$$\mathbb{P}(|\widehat{\mu}_n - \mu| \ge \varepsilon) \le 2e^{-2n\varepsilon^2}, \qquad \varepsilon > 0.$$

We also use Ville's inequality: if $(M_n)_{n\ge 0}$ is a nonnegative supermartingale with $M_0 \le 1$ almost surely, then

$$\mathbb{P}\left(\sup_{n\ge 0} M_n \ge \delta^{-1}\right) \le \delta, \qquad 0 < \delta < 1.$$

See Ville (1939) or the restatement in Howard et al. (2020, Lemma 1, Eq. (2.11)).

C Certificates and insertion losses

Proof of Lemma 1. Fix a priority σ and its greedy basis $B = (b_1, \dots, b_r)$.

Exchange partners. If greedy rejects i, the accepted prefix together with i contains a circuit. That prefix is contained in B, and the only circuit in $B + i$ is $C_B(i)$. Hence every element of $K_B(i)$ was accepted before i. If greedy reaches rank r before scanning i, all of B precedes i instead. In either case, every feasible removal precedes i in σ.

Put $P_\ell = \{b_1, \dots, b_\ell\}$ and $k_i = \max\{\ell : b_\ell \in C_B(i)\}$. A prefix becomes dependent after inserting i precisely when it contains all of the fundamental circuit:

$$P_\ell + i \notin \mathcal{I} \iff C_B(i) \subseteq P_\ell + i \iff \ell \ge k_i.$$

Looplessness and rank supply the endpoints at $\ell = 0, r$. Binary search therefore finds $j_\sigma(i) = b_{k_i}$ with at most $\lceil \log_2 r \rceil$ calls. Together with greedy, a build costs at most $c_{d,r}$ calls and uses $O(d)$ words. Both chain edges and outside edges point backward in σ.

Feasible actions. Set $F_k = \mathrm{cl}(P_k)$; looplessness gives $F_0 = \varnothing$. For an outside element,

$$i \in F_k \quad \iff \quad K_B(i) \subseteq P_k \quad \iff \quad k_i \le k \qquad (i \notin B).$$

Since b_k first enters the chain at F_k, the structural groups are exactly $G_k = F_k \setminus F_{k-1}$.

Choose any $a_k \in G_k$ and write $A_k = \{a_1, \dots, a_k\}$. The empty set A_0 is independent and spans F_0. If A_{k-1} is independent and spans F_{k-1}, then $a_k \notin \mathrm{cl}(A_{k-1})$, so A_k is independent. The resulting set has size k and lies in the rank-k flat F_k, which it therefore spans. At $k = r$, A_r is a basis. This is the chain-of-flats construction of Bérczi & Schwarcz (2020, Lemma 10); it does not use the means.

Optimality. If the chain inequalities hold, the last feasible removal also has the smallest mean:

$$\mu_{b_1} \ge \cdots \ge \mu_{b_r}, \qquad \mu_{j_\sigma(i)} = \min_{j \in K_B(i)} \mu_j.$$

If the outside inequalities also hold, take any basis A and its exchange matching from Lemma B.1. Each matched removal has mean at least that of the stored partner, so

$$\mu(A) - \mu(B) = \sum_{i \in A \setminus B} (\mu_i - \mu_{\phi(i)}) \le \sum_{i \in A \setminus B} (\mu_i - \mu_{j_\sigma(i)}) \le 0.$$

Thus B is optimal and every stored partner minimizes the removal mean. If the certificate fails, either a chain edge is inverted or the chain is ordered and an outside edge has $\mu_i > \mu_{j_\sigma(i)}$. The optimality argument uses only weak inequalities and therefore includes ties. □

Lemma C.1 (Insertion losses)**.** *For every optimal basis B and $i \notin B$,*

$$\min_{j \in K_B(i)} \mu_j = m_i = \mu_i + \Delta_i.$$

For every basis A,

$$\mu^\star - \mu(A) \ge \sum_{i \in A} \Delta_i, \qquad R_\mu(T) \ge \sum_{i \in E_+} \Delta_i \mathbb{E}_\mu N_i(T). \tag{10}$$

Proof. Fix an optimal basis B and an outside arm i. Write $m = \min_{j \in K_B(i)} \mu_j$, and choose a minimizing j. The basis $B - j + i$ contains i and has mean $\mu^\star - m + \mu_i$. The definition of insertion loss gives $\Delta_i \le m - \mu_i$.

For the reverse inequality, consider any basis A containing i. Match $A \setminus B$ into $B \setminus A$ using Lemma B.1. Optimality of B makes every matched exchange gain nonpositive. Dropping all terms except the one for i gives

$$\mu(A) - \mu^\star = \sum_{k \in A \setminus B} (\mu_k - \mu_{\phi(k)}) \le \mu_i - \mu_{\phi(i)} \le \mu_i - m.$$

Maximizing over bases containing i gives the reverse inequality, so $\Delta_i = m - \mu_i$ and $m_i = m$. In particular, $m_i \in (0, 1)$ for $i \in E_+$.

For an arbitrary basis A, use this matching again: each $i \in A \setminus B$ is matched to a removal with $\mu_{\phi(i)} - \mu_i \ge \Delta_i$. Arms in $A \cap B$ have zero insertion loss. Summing proves the first inequality in Equation 10. Applying it to each action, summing over rounds, and taking expectations proves the second. □

The chain inequality cannot be omitted. For $U_{2,3}$, take priority $(1, 2, 3)$ and means $(0.4, 0.9, 0.6)$. The build returns $B = \{1, 2\}$ and $j_\sigma(3) = 2$, so the stored actions are $\{1, 2\}$ and $\{1, 3\}$, with means 1.3 and 1.0. The outside inequality $\mu_3 \le \mu_2$ holds, but the omitted action $\{2, 3\}$ is optimal, with mean 1.5. Restricting play to the stored actions would therefore incur at least $0.2T$ pseudo-regret. The chain inequality for $(2, 1)$ is violated here, since $0.9 > 0.4$. At rank one the chain is empty and the partner is known without binary search. When the rank equals d, a case excluded from our model, the only basis is E and no structural query is needed.

D Proof of Theorem 1

Throughout, $q = d - r$, $v = q + 1$, and $\alpha = \eta = 1/(8d)$. The finite instance-dependent constants C_μ, a_μ may grow between displays.

D.1 Build count

Let $\mathcal{A}_k$ be the event that build k is reached, τ_k its starting time in reward rounds, and $\mathcal{G}_k$ the sigma-field of the interaction history immediately after the build, including its oracle answers. Thus $\tau_1 = 0$. Statements conditional on $\mathcal{G}_k$ apply on $\mathcal{A}_k$; variables for episodes not reached are set to zero in unconditional sums. The learning duration is $D_k \in \{v, v + 1, \dots\} \cup \{\infty\}$, with $D_k = \infty$ when no trigger occurs.

Lemma D.1 (Fresh observations at adaptive starts)**.** *At every reached build, trigger, or batch start, the unused coordinate streams are conditionally independent Bernoulli streams with means μ.*

Proof. Represent coordinate i by an independent stream $(X_{i,n})_{n \ge 1}$ of Bernoulli(μ_i) variables. Selecting i reveals its next unused entry. Predictability of the actions makes this representation equivalent to the round-indexed model.

Let $\mathcal{H}$ be the history at a finite reached start and C_i the number of entries already revealed from stream i. For a fixed count vector c, every event $A \in \mathcal{H}$ contained in $\{C = c\}$ is determined by the policy seed and the prefixes $(X_{i,n} : n \le c_i)_{i \in [d]}$. Thus, for finite lengths h_i and sets $D_i \subseteq \{0, 1\}^{h_i}$,

$$\mathbb{P}\left(A \cap \bigcap_i \{(X_{i,c_i+1}, \dots, X_{i,c_i+h_i}) \in D_i\} \right) = \mathbb{P}(A) \prod_i \text{Bernoulli}(\mu_i)^{\otimes h_i}(D_i).$$

Summing over c gives the same identity with the random counts C_i. This is the asserted product law conditional on $\mathcal{H}$. □

Lemma D.2 (Episode error bounds)**.** *Let $\mathcal{F}_{k,\ell_{\text{ep}}}$ be the event that episode k is tested at age ℓ_{ep} and some local estimation error exceeds its confidence radius. Let E_k indicate that the local confidence*

bounds fail at the trigger or some repair interval misses its mean; set $E_k = 0$ if the episode never triggers. Then

$$\mathbb{P}(\mathcal{F}_{k,\ell_{\mathrm{ep}}} \mid \mathcal{G}_k) \le \frac{\alpha}{2\ell_{\mathrm{ep}}^5}, \tag{11}$$

$$\mathbb{P}(E_k = 1 \mid \mathcal{G}_k) \le p := \alpha + \eta = \frac{1}{4d}. \tag{12}$$

If the certificate at $\mathcal{G}_k$ is correct, then for $0 \le s < 4$,

$$\mathbb{E}[D_k^s \mathbf{1}\{D_k < \infty\} \mid \mathcal{G}_k] \le \frac{\alpha}{2} \sum_{\ell_{\mathrm{ep}} \ge 1} \ell_{\mathrm{ep}}^{s-5} < \infty, \qquad \mathbb{E}[D_k \mathbf{1}\{D_k < \infty\} \mid \mathcal{G}_k] < \alpha. \tag{13}$$

Products involving $D_k = \infty$ are defined to be zero.

Proof. Let $\widehat{\mu}_{k,i,n}$ be the mean of the first n unused observations of coordinate i at build k. For deterministic $1 \le n \le \ell_{\mathrm{ep}}$, freshness and Hoeffding's inequality give

$$\mathbb{P}(|\widehat{\mu}_{k,i,n} - \mu_i| > \rho(n, \ell_{\mathrm{ep}}) \mid \mathcal{G}_k) \le 2e^{-2n\rho(n,\ell_{\mathrm{ep}})^2} = \frac{\alpha}{2d\ell_{\mathrm{ep}}^6}.$$

At age ℓ_{ep}, every local count lies in $\{1, \dots, \ell_{\mathrm{ep}}\}$. Union-bounding over these counts and the d coordinates proves Equation 11 at the adaptive counts.

Let $\mathcal{T}_k$ be the history at a reached trigger. The values g, w, m_{rep} are $\mathcal{T}_k$-measurable, and the designated observations are fresh, so

$$\mathbb{P}(\exists i : |x_i - \mu_i| > w \mid \mathcal{T}_k) \le 2de^{-2m_{\mathrm{rep}}w^2} \le \eta.$$

A bad refinement requires either a local error at a tested age or a batch error following a finite trigger. Conditioning the batch bound back on $\mathcal{G}_k$ gives

$$\begin{aligned}\mathbb{P}(E_k = 1 \mid \mathcal{G}_k) &\le \sum_{\ell_{\mathrm{ep}} \ge 1} \mathbb{P}(\mathcal{F}_{k,\ell_{\mathrm{ep}}} \mid \mathcal{G}_k) + \eta\, \mathbb{P}(D_k < \infty \mid \mathcal{G}_k) \\ &\le (\alpha/2) \sum_{\ell_{\mathrm{ep}} \ge 1} \ell_{\mathrm{ep}}^{-5} + \eta \le p.\end{aligned}$$

Here the conditioning remains at the build: $\mathcal{G}_k$ does not include the event $D_k < \infty$.

For a correct certificate, if all local errors are within their radii, an edge's empirical gain is at most the sum of its two radii and cannot pass the trigger threshold. Therefore $\{D_k = \ell_{\mathrm{ep}}\} \subseteq \mathcal{F}_{k,\ell_{\mathrm{ep}}}$. Multiply the probability bound by ℓ_{ep}^s and sum over ℓ_{ep} to obtain Equation 13. The series converges for $s < 4$; at $s = 1$ it is less than α. Episodes that never trigger contribute zero to these truncated moments. □

Proof of Lemma 2. Call a partition *ordered by mean* when every element in an earlier block has mean at least that of every element in a later block. The initial one-block partition is ordered by mean. At a good trigger, $\widehat{\mu}_a - \widehat{\mu}_b = 2g$ and the local bounds imply

$$\mu_a - \mu_b \ge 2g - \rho(n_a, D_k) - \rho(n_b, D_k) > g > 0. \tag{14}$$

Every batch interval contains its true mean. Equal means in the same parent block therefore have intersecting intervals and remain together. Distinct overlap components have disjoint interval hulls, so ordering those hulls orders their true means within the parent block.

The triggering pair cannot remain in one component. Otherwise an overlap path with at most $d - 1$ edges would imply

$$|x_a - x_b| \le 2w(d-1), \qquad \mu_a - \mu_b \le |x_a - x_b| + 2w \le 2wd = g/2,$$

contradicting Equation 14. Thus a good refinement without a reset adds at least one block, and a good refinement after a reset leaves at least two.

Certificate edges point backward in the priority. A good cross-block trigger has $\mu_a > \mu_b$ even though b lies in an earlier block, so the current block order cannot be correct. A good nonreset refinement preserves a correct block order, and a good reset restores it. Hence each good reset is preceded by a

bad refinement since initialization or the previous good reset. Assigning a reset to the last such bad refinement is injective.

Among R refinements, let G, C, H count good nonreset refinements, good resets, and bad refinements. We have $C \le H$. A good nonreset refinement adds at least one block; a good reset loses at most $d-2$; a bad refinement loses at most $d-1$. Starting from one block gives

$$R = G + C + H, \qquad C \le H, \qquad b_R \ge 1 + G - (d-2)C - (d-1)H.$$

Eliminating G yields

$$R \le b_R - 1 + (d-1)C + dH \le d - 1 + (2d-1)H. \tag{15}$$

If $H = 0$, no tie has been split and $b_R \le L$. If $H \ge 1$, the general bound $b_R \le d$ gives $b_R \le L + (d-L)H$. Substituting this and $C \le H$ into the first inequality yields

$$K_R = R + 1 \le L + (3d - L - 1)H.$$

□

Lemma D.3 (Number of builds). *Let $K = \sum_{k\ge 1} \mathbf{1}_{\mathcal{A}_k}$. Then*

$$\mathbb{E}Q_\infty \le c_{d,r}\mathbb{E}K \le \frac{4dL}{d+L+1} c_{d,r} < 2dc_{d,r}, \tag{16}$$

and

$$\mathbb{P}(K \ge M) \le \min\{1, 2^{d/(2d-1)} e^{-\lambda_d M}\}, \qquad \lambda_d = \frac{\log 2}{2d-1} - \log(1+p) > 0. \tag{17}$$

In particular, $\mathbb{E}K^s < \infty$ for every $s > 0$.

Proof. Allow $K = \infty$, and for $M \ge 1$ set $K_M = K \wedge M$ and $H_M = \sum_{k<M} \mathbf{1}_{\mathcal{A}_k} E_k$. Every triggered batch finishes and is followed by a build. Thus H_M counts exactly the bad refinements before build K_M. The conditional error bound and Lemma 2 give

$$\mathbb{E}H_M \le p \sum_{k<M} \mathbb{P}(\mathcal{A}_k) \le p\mathbb{E}K_M, \qquad \mathbb{E}K_M \le L + (3d - L - 1)p\mathbb{E}K_M.$$

Since $1 - (3d - L - 1)p = (d + L + 1)/(4d) > 0$, $\mathbb{E}K_M \le 4dL/(d+L+1)$ uniformly in M. Monotone convergence gives the same bound for $\mathbb{E}K$. Multiplying by the per-build cost $c_{d,r}$ proves Equation 16.

For the tail bound, on $\mathcal{A}_M$, the pathwise bound in Equation 15 requires at least $(M-d)/(2d-1)$ bad refinements before build M. Define $V_m = \mathbf{1}_{\mathcal{A}_m} 2^{\sum_{k<m} E_k}$, with indicators for unreached builds set to zero. On $\mathcal{A}_k$, $\mathbf{1}_{\mathcal{A}_{k+1}} 2^{E_k} \le 1 + E_k$, and hence

$$\mathbb{E}[\mathbf{1}_{\mathcal{A}_{k+1}} 2^{E_k} \mid \mathcal{G}_k] \le 1 + p.$$

Previous errors are measurable at $\mathcal{G}_k$. Multiplication by V_k and the tower property give $\mathbb{E}V_{k+1} \le (1+p)\mathbb{E}V_k$. Since $V_1 = 1$, we obtain $\mathbb{E}V_M \le (1+p)^{M-1}$. Markov's inequality now gives

$$\mathbb{P}(K \ge M) \le 2^{-(M-d)/(2d-1)} (1+p)^{M-1} \le 2^{d/(2d-1)} e^{-\lambda_d M}.$$

Since $\log(1+p) < 1/(4d) < \log 2/(2d-1)$, we have $\lambda_d > 0$. Taking the minimum with one proves Equation 17, and summing the tail gives $\mathbb{E}K^s < \infty$ for every $s > 0$. □

All stored arrays and the interval workspace have length at most d; cover actions are reconstructed from the basis and partner array. This also proves the $O(d)$ memory bound.

D.2 Stabilization

Set $\Lambda(\ell_{\mathrm{ep}}) = \log(4d\ell_{\mathrm{ep}}^6/\alpha)$. At a trigger of age $D_k < \infty$, the trigger rule and $n_a, n_b \le D_k$ imply

$$g > \rho(n_a, D_k) + \rho(n_b, D_k) \ge \sqrt{2\Lambda(D_k)/D_k}, \qquad m_{\mathrm{rep}} \le 1 + \frac{4d^2 D_k \log(2d/\eta)}{\Lambda(D_k)}.$$

Thus the repair duration $S_k = vm_{\text{rep}}$ satisfies, on every sample path,

$$S_k \le \kappa_d D_k, \qquad \kappa_d = v\left(1 + \frac{4d^2 \log(2d/\eta)}{\log(4d/\alpha)}\right). \tag{18}$$

Let W_T, S_T count learning rounds with incorrect certificates and repair rounds through T, respectively, and let

$$Y = \sum_{k\ge 1} \mathbf{1}_{\mathcal{A}_k} \mathbf{1}\{k \text{ correct}\} D_k \mathbf{1}\{D_k < \infty\}.$$

Only completed correct episodes enter Y. Since correctness and reachability are measurable at $\mathcal{G}_k$, summing Equation 13 over reached builds bounds its expectation. Each repair, even one still unfinished at T, can be charged to its completed triggering episode. Thus

$$\mathbb{E}Y \le \alpha \mathbb{E}K < \infty, \qquad S_T \le \kappa_d(W_T + Y). \tag{19}$$

For $\varepsilon > 0$, let $\bar{X}_{i,n}$ be the mean of the first n rewards in coordinate i's stream and define

$$H_\varepsilon = 1 + \sup\{n \ge 1 : \exists i,\ |\bar{X}_{i,n} - \mu_i| > \varepsilon\}, \qquad \sup \varnothing = 0.$$

For every $n \ge H_\varepsilon$, all d stream means are within ε of their expectations. This is a random threshold in observation count, not a stopping rule. A union bound gives

$$\mathbb{P}(H_\varepsilon > h) \le 2d \sum_{n\ge h} e^{-2n\varepsilon^2} = \frac{2de^{-2h\varepsilon^2}}{1 - e^{-2\varepsilon^2}}, \qquad h \in \mathbb{N},\ h \ge 1. \tag{20}$$

Define $H_{k,\varepsilon}$ from the unused streams at build k. Lemma D.1 gives the same conditional tail bound and hence

$$\mathbb{E}H_\varepsilon^s < \infty, \qquad \mathbb{E}[H_{k,\varepsilon}^s \mid \mathcal{G}_k] \le C_{\varepsilon,s} < \infty \quad (s > 0).$$

The moment bounds follow by summing $((h+1)^s - h^s)\mathbb{P}(H_\varepsilon > h)$ over integers $h \ge 0$. No independence between the delays is needed.

For this analysis, define $U_i(t)$ by Equation 3 at every round, including covers and repairs, whether or not the action rule uses it.

Lemma D.4 (Summable index underestimation)**.** *For* $\mathcal{U}_t = \{\exists i : U_i(t) < \mu_i\}$,

$$Z := \sum_{t\ge 1} \mathbf{1}_{\mathcal{U}_t}, \qquad \mathbb{E}Z < \infty. \tag{21}$$

Proof. For a Bernoulli stream of mean μ, write $V_n = \sum_{s\le n} X_s$ and $\widehat{\mu}_n = V_n/n$. Fix $x \ge 1$ and $t \ge 2$. Cover $[1, t]$ by geometric intervals $[u, v]$, where $u = (1 + 1/x)^j$ and $v = \min\{(1 + 1/x)u, t\}$. Since $\log(1 + 1/x) \ge 1/(x + 1)$, at most $(x + 1)\log t + 2$ intervals are needed.

For $n \in [u, v]$, the event

$$\widehat{\mu}_n < \mu, \qquad n\,\mathrm{kl}(\widehat{\mu}_n, \mu) > x$$

forces $\widehat{\mu}_n \le \xi$, where $\mathrm{kl}(\xi, \mu) = x/v$ and $0 \le \xi < \mu$. Indeed, $n \le v$ and KL divergence decreases on $[0, \mu]$. If $x/v > \mathrm{kl}(0, \mu)$, the event is empty. Otherwise, when $0 < \xi < \mu$, use the exponential martingale

$$\lambda = \log \frac{\xi(1-\mu)}{\mu(1-\xi)} < 0, \qquad M_n = \exp\{\lambda V_n - n \log(1 - \mu + \mu e^\lambda)\}.$$

Since $\lambda < 0$, this event implies

$$\log M_n \ge n\,\mathrm{kl}(\xi, \mu) \ge ux/v \ge x/(1 + 1/x) \ge x - 1.$$

Ville's inequality bounds the probability of this event anywhere in the interval by e^{1-x}; for $x = 1$ the same bound is trivial. At the endpoint $\xi = 0$, use $M_n = \mathbf{1}\{V_n = 0\}(1 - \mu)^{-n}$. Summing over intervals yields

$$\mathbb{P}(\exists n \le t : \widehat{\mu}_n < \mu,\ n\,\mathrm{kl}(\widehat{\mu}_n, \mu) > x) \le \mathrm{e}\{(x + 1)\log t + 2\}e^{-x}.$$

An unobserved arm has index one. For every other arm,

$$\{U_i(t) < \mu_i\} \subseteq \bigcup_{1 \le n \le t-1} \{\widehat{\mu}_{i,n} < \mu_i,\ n\,\mathrm{kl}(\widehat{\mu}_{i,n}, \mu_i) > f(t)\}.$$

The bound above therefore applies at the adaptive count $N_i(t-1)$ without conditioning on that count. Summing over arms gives

$$\mathbb{P}(\mathcal{U}_t) \le d e\{(f(t)+1)\log t + 2\} e^{-f(t)} = O\left(\frac{1}{t \log^4 t}\right).$$

Summing this bound and including the finitely many initial times gives, by Tonelli's theorem, $\mathbb{E}Z = \sum_t \mathbb{P}(\mathcal{U}_t) < \infty$, proving Equation 21. □

If the means are not all equal, set

$$\gamma = \min_{\mu_i \ne \mu_j} |\mu_i - \mu_j| > 0, \qquad \varepsilon = \gamma/32, \qquad H = H_\varepsilon, \quad H_k = H_{k,\varepsilon},$$

and $Z_k(\ell_{\mathrm{ep}}) = \sum_{t=\tau_k+1}^{\tau_k+\ell_{\mathrm{ep}}} \mathbf{1}_{\mathcal{U}_t}$.

Lemma D.5 (Detection of an incorrect certificate)**.** *Suppose episode k has an incorrect certificate and no test has triggered through age $\ell_{\mathrm{ep}} \ge v$. With $t_* = \tau_k + \ell_{\mathrm{ep}} + 1$,*

$$\ell_{\mathrm{ep}} \le C_\mu\{f(t_*) + H + H_k + Z_k(\ell_{\mathrm{ep}}) + 1\}, \tag{22}$$

where $C_\mu < \infty$ is independent of k, ℓ_{ep}.

Proof. Fix the coordinate streams and write $F_* = f(t_*)$. By Lemma 1, either a basis edge is inverted or the ordered chain has an improving outside exchange.

The Bernoulli inequality $\mathrm{kl}(p,u) \ge 2(p-u)^2$ implies

$$\bar{\mu}_i(t-1) \le U_i(t) \le \bar{\mu}_i(t-1) + \sqrt{\frac{f(t)}{2N_i(t-1)}} \quad \text{when } N_i(t-1) > 0. \tag{23}$$

For $u \in (0,1)$, the function $h(p) = \mathrm{kl}(p,u) - 2(p-u)^2$ satisfies $h(u) = h'(u) = 0$ and $h''(p) = 1/[p(1-p)] - 4 \ge 0$. Continuity handles $p = 0, 1$; at $u = 0, 1$, the divergence is infinite unless $p = u$. Forced plays also give

$$n_j(v+u) \ge 1 + \lceil u/2 \rceil \qquad (j \in B,\ u \ge 0). \tag{24}$$

Here $n_j(v+u)$ is the local count after the cover and u additional rounds.

Case 1: an inverted basis edge. Suppose

$$\ell_{\mathrm{ep}} \ge 2v + 4, \qquad \ell_{\mathrm{ep}} \ge 4H_k, \qquad \ell_{\mathrm{ep}} \ge 128\Lambda(\ell_{\mathrm{ep}})/\gamma^2.$$

Then $n_j \ge \ell_{\mathrm{ep}}/4 \ge H_k$ and $\rho(n_j, \ell_{\mathrm{ep}}) \le \gamma/8$ for every $j \in B$. An inverted edge (a,b) satisfies

$$\widehat{\mu}_a - \widehat{\mu}_b \ge \gamma - 2\varepsilon = 15\gamma/16 > \gamma/2 \ge 2[\rho(n_a, \ell_{\mathrm{ep}}) + \rho(n_b, \ell_{\mathrm{ep}})],$$

contradicting the absence of a trigger.

Case 2: an ordered chain and an improving outside exchange. There is an outside arm a with $\mu_a - \mu_{j(a)} \ge \gamma$. Call $i \notin B$ *improving* when $\mu_i > \mu_{j(i)}$, and *nonimproving* otherwise. Suppose

$$\ell_{\mathrm{ep}} \ge 8(v+3), \qquad \ell_{\mathrm{ep}} \ge 8H, \qquad \ell_{\mathrm{ep}} \ge 4096F_*/\gamma^2. \tag{25}$$

Consider local rounds $\lfloor \ell_{\mathrm{ep}}/2 \rfloor + 1, \ldots, \ell_{\mathrm{ep}}$. All follow the cover. Before the first of these rounds, every basis coordinate has at least $1 + \lceil (\lfloor \ell_{\mathrm{ep}}/2 \rfloor - v)/2 \rceil \ge \ell_{\mathrm{ep}}/8$ local observations and at least that many global observations. Alternation also supplies

$$\#\{\text{optimistic rounds in the second half}\} \ge \lfloor \lceil \ell_{\mathrm{ep}}/2 \rceil / 2 \rfloor \ge \ell_{\mathrm{ep}}/8. \tag{26}$$

By Equation 23, throughout this half every basis index satisfies

$$\mu_j - \varepsilon \le U_j(t) \le \mu_j + \varepsilon + \sqrt{4F_*/\ell_{\mathrm{ep}}} \le \mu_j + \gamma/16.$$

On an optimistic round outside $\mathcal{U}_t$, the selected maximum a^+ in a's group satisfies

$$U_{a^+}(t) - U_{j(a)}(t) \geq U_a(t) - U_{j(a)}(t) \geq 15\gamma/16.$$

Thus $a^+ \notin B$. Among selected outside arms, let c_t maximize the index gain, breaking ties by label. Charge the round to c_t; then

$$c_t \in A_t \setminus B, \qquad U_{c_t}(t) - U_{j(c_t)}(t) > \gamma/2. \tag{27}$$

Each charge increases both the global and local count of c_t.

A nonimproving outside arm i cannot receive such a charge once $N_i(t-1) \geq \max\{H, 8F_*/\gamma^2\}$, because then

$$U_i(t) - U_{j(i)}(t) \leq 2\varepsilon + \sqrt{F_*/(2N_i(t-1))} \leq 5\gamma/16 < \gamma/2.$$

A given pre-observation global count can occur on at most one charge to that arm. Thus the total number of charges to nonimproving arms in the second half is at most

$$B_q = q\{H + \lceil 8F_*/\gamma^2 \rceil + 1\}. \tag{28}$$

At most $Z_k(\ell_{\mathrm{ep}})$ of the optimistic rounds belong to $\mathcal{U}_t$. If

$$\ell_{\mathrm{ep}} \geq 16\{Z_k(\ell_{\mathrm{ep}}) + B_q + 1\},$$

then at least $\ell_{\mathrm{ep}}/8 - Z_k(\ell_{\mathrm{ep}}) - B_q \geq \ell_{\mathrm{ep}}/16$ rounds are charged to improving arms. There are at most q improving arms, so one of them, say i, has $n_i \geq \ell_{\mathrm{ep}}/(16q)$ local observations by age ℓ_{ep}. Its partner has $n_{j(i)} \geq \ell_{\mathrm{ep}}/8$. If also

$$\ell_{\mathrm{ep}} \geq 32qH_k, \qquad \ell_{\mathrm{ep}} \geq 512q\Lambda(\ell_{\mathrm{ep}})/\gamma^2,$$

then

$$\min\{n_i, n_{j(i)}\} \geq H_k, \qquad \max\{\rho(n_i, \ell_{\mathrm{ep}}), \rho(n_{j(i)}, \ell_{\mathrm{ep}})\} \leq \sqrt{\frac{8q\Lambda(\ell_{\mathrm{ep}})}{\ell_{\mathrm{ep}}}} \leq \frac{\gamma}{8}. \tag{29}$$

The empirical gain is at least $\gamma - 2\varepsilon = 15\gamma/16$, whereas the trigger threshold is at most $\gamma/2$, a contradiction.

In either case, absence of a trigger requires at least one threshold to fail. Each threshold is bounded by the right-hand side below; hence

$$\begin{aligned} \ell_{\mathrm{ep}} \leq\ & 16(v+3) + (8+16q)H + 32qH_k + 16Z_k(\ell_{\mathrm{ep}}) + 32q + 16 \\ & + \gamma^{-2}\{(4096 + 128q)F_* + 512q\Lambda(\ell_{\mathrm{ep}})\}. \end{aligned} \tag{30}$$

Finally, $\ell_{\mathrm{ep}} \leq t_*$ gives $\Lambda(\ell_{\mathrm{ep}}) \leq \log(4d/\alpha) + 6f(t_*)$. Absorbing the fixed instance-dependent coefficients into C_μ proves Equation 22. □

Lemma D.6 (Finite transient duration)**.** *The total number of incorrect learning rounds and the total repair duration satisfy*

$$\mathbb{E}W_\infty < \infty, \qquad \mathbb{E}S_\infty \leq \kappa_d \mathbb{E}(W_\infty + Y) < \infty. \tag{31}$$

The expected number of reward rounds before the last build is finite, and its certificate is correct almost surely.

Proof. If all means coincide, every certificate is correct. Then $W_\infty = 0$, and Equation 19 gives $\mathbb{E}S_\infty < \infty$. The last-build argument below still applies. For the rest of the proof, assume at least two mean levels.

Apply Lemma D.5 to each incorrect episode's portion observed by T. If it triggers at age $D_k > v$, apply the lemma at $D_k - 1$ and add the triggering round. An episode triggering at its first test, or a prefix shorter than v, contributes at most v rounds per build.

Let s_T be the last incorrect learning round by T, taking $s_T = 0$ when $W_T = 0$. Every application above has $t_* \leq s_T + 1$, so $f(t_*) \leq C\log(\mathrm{e} + s_T)$ for an absolute constant C. Episode portions are disjoint, and hence their underestimation counts sum to at most Z. We obtain

$$W_T \leq C_\mu K \log(\mathrm{e} + s_T) + \Xi_{\mu,0}, \qquad \Xi_{\mu,0} = C_\mu \left(KH + \sum_{k \leq K} H_k + Z + K \right). \tag{32}$$

The tail bounds in Equation 17 and Equation 20, together with freshness at reached builds, give

$$\mathbb{E}(KH) \le (\mathbb{E}K^2\,\mathbb{E}H^2)^{1/2} < \infty, \qquad \mathbb{E}\sum_{k\le K} H_k = \sum_k \mathbb{E}[\mathbf{1}_{\mathcal{A}_k}\mathbb{E}(H_k \mid \mathcal{G}_k)] \le C_\mu \mathbb{E}K < \infty.$$

Together with $\mathbb{E}Z < \infty$, these bounds give $\mathbb{E}\Xi_{\mu,0} < \infty$.

All correct episodes before s_T have ended, so their durations are included in Y. The remaining rounds are incorrect learning or repair rounds. Therefore

$$s_T \le W_T + Y + S_{s_T} + 1 \le (1+\kappa_d)(W_T + Y) + 1.$$

Using $\mathrm{e} + s_T \le (2+\kappa_d)(\mathrm{e}+W_T)(1+Y)$ in Equation 32 gives

$$W_T \le a_\mu K \log(\mathrm{e}+W_T) + \Xi_\mu, \qquad \Xi_\mu = \Xi_{\mu,0} + a_\mu K\{\log(2+\kappa_d) + \log(1+Y)\}. \tag{33}$$

The term involving Y is integrable because $\log(1+y) \le \sqrt{y}$ for $y \ge 0$ and

$$\mathbb{E}[K\log(1+Y)] \le (\mathbb{E}K^2\,\mathbb{E}Y)^{1/2} < \infty, \qquad \mathbb{E}\Xi_\mu < \infty.$$

For $a, w \ge 0$, concavity gives the tangent bound

$$a\log(\mathrm{e}+w) \le a\log(\mathrm{e}+2a) + \frac{a(w-2a)}{\mathrm{e}+2a} \le a\log(\mathrm{e}+2a) + w/2.$$

Thus, for $a, b, w \ge 0$,

$$w \le a\log(\mathrm{e}+w) + b \implies w \le 2a\log(\mathrm{e}+2a) + 2b.$$

With $a = a_\mu K$, this bounds every W_T by $2a_\mu K\log(\mathrm{e}+2a_\mu K) + 2\Xi_\mu$. Its expectation is finite by Equation 17, and it does not depend on T. Monotone convergence proves $\mathbb{E}W_\infty < \infty$; Equation 19 then gives the repair-time bound in Equation 31.

Since $K < \infty$ almost surely and every episode before the last build is completed,

$$\tau_{\mathrm{last}} \le W_\infty + Y + S_\infty, \qquad \mathbb{E}\tau_{\mathrm{last}} < \infty.$$

An incorrect final certificate would create infinitely many incorrect learning rounds, contradicting $\mathbb{E}W_\infty < \infty$. This pathwise inequality does not require τ_{last} to be a stopping time. □

More generally, suppose Equation 32 holds with $\mathbb{E}K^2 < \infty$, $\mathbb{E}\Xi_{\mu,0} < \infty$, and

$$s_T \le c(h_0 + W_T + Y + V + 1), \qquad \mathbb{E}(Y+V) < \infty, \tag{34}$$

where $h_0 \ge 0$, $c \ge 1$ are deterministic and $V \ge 0$ is additional sampling time. The inequality

$$\mathrm{e} + s_T \le (\mathrm{e}+c)(1+h_0)(\mathrm{e}+W_T)(1+Y+V)$$

gives $W_T \le a_\mu K\log(\mathrm{e}+W_T) + \widetilde{\Xi}_\mu$, where

$$\widetilde{\Xi}_\mu = \Xi_{\mu,0} + a_\mu K\{\log(\mathrm{e}+c) + \log(1+h_0) + \log(1+Y+V)\}.$$

The remainder is integrable because

$$\mathbb{E}[K\log(1+Y+V)] \le (\mathbb{E}K^2\,\mathbb{E}(Y+V))^{1/2} < \infty.$$

Also, $K\log(\mathrm{e}+2a_\mu K) \le C_\mu(1+K^2)$. The same tangent bound therefore yields an integrable upper bound,

$$W_T \le 2a_\mu K\log(\mathrm{e}+2a_\mu K) + 2\widetilde{\Xi}_\mu, \qquad \mathbb{E}W_\infty < \infty. \tag{35}$$

For constant means, $W_\infty = 0$ without a detection estimate.

D.3 REGRET

Lemma D.7 (Counting observations across episodes)**.** *Fix a coordinate i and $m \in (\mu_i, 1)$. Let*

$$L_i(m,T) = \sum_{t=1}^{T} \mathbf{1}\{i \in A_t,\ U_i(t) \geq m\}, \qquad I = \mathrm{kl}(\mu_i, m), \quad a = \log \frac{m(1-\mu_i)}{\mu_i(1-m)} > 0.$$

For all $T \geq 1$,

$$\mathbb{E}L_i(m,T) \leq \frac{f(T)}{I} + C_{\mu_i,m}(\sqrt{f(T)} + 1), \tag{36}$$

where $C_{\mu_i,m} < \infty$ is independent of the certificate sequence.

Proof. Write $V_n = \sum_{s \leq n} X_{i,s}$ and $\widehat{\mu}_n = V_n/n$. Convexity of $p \mapsto \mathrm{kl}(p,m)$, whose derivative at μ_i is $-a$, gives

$$\mathrm{kl}(p,m) \geq I - a(p - \mu_i), \qquad p \in [0,1]. \tag{37}$$

Suppose i is selected at $t \leq T$, with $n = N_i(t-1) > f(T)/I$ and $U_i(t) \geq m$. If $\widehat{\mu}_n < m$, then

$$nI - an(\widehat{\mu}_n - \mu_i) \leq n\,\mathrm{kl}(\widehat{\mu}_n, m) \leq f(t) \leq f(T).$$

If $\widehat{\mu}_n \geq m$, Equation 37 at $p = m$ gives $an(\widehat{\mu}_n - \mu_i) \geq an(m - \mu_i) \geq nI$. Thus, in either case,

$$V_n - n\mu_i \geq \frac{nI - f(T)}{a}. \tag{38}$$

Every selection increments the global count N_i, so each pre-selection count appears at most once, even across episodes. Put $h = f(T)/I$. There are at most $\lfloor h \rfloor + 1$ possible counts from zero through h, and all later qualifying counts satisfy Equation 38. Hence

$$L_i(m,T) \leq \lfloor h \rfloor + 1 + \sum_{n > h} \mathbf{1}\left\{ V_n - n\mu_i \geq \frac{I(n-h)}{a} \right\}.$$

Taking expectations and applying Hoeffding at each deterministic count gives

$$\mathbb{E}L_i(m,T) \leq h + 1 + \sum_{n > h} \exp\left(-\frac{2I^2(n-h)^2}{a^2 n} \right).$$

For $h < n \leq 2h$, replacing the denominator n by $2h$ gives a decreasing function of $n - h$. Its sum is at most $1 + \int_0^\infty e^{-I^2u^2/(a^2h)}\,du = 1 + a\sqrt{\pi h}/(2I)$; the extra one covers noninteger h. For $n > 2h$, use $n - h > n/2$ to bound the summand by $e^{-I^2 n/(2a^2)}$, whose sum is geometric. The total remainder above $f(T)/I$ is thus at most

$$2 + \frac{a\sqrt{\pi}}{2I^{3/2}}\sqrt{f(T)} + \frac{1}{1 - e^{-I^2/(2a^2)}}.$$

This proves Equation 36 with a constant depending only on μ_i and m, not on the certificates. □

Lemma D.8 (Known-matroid lower bound)**.** *For any feasible, uniformly efficient policy and every mean vector,*

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\mu N_i(T)}{\log T} \geq I_i^{-1} \quad (i \in E_+), \tag{39}$$

and $\liminf_{T\to\infty} R_\mu(T)/\log T \geq C^\star(\mu)$*. The policy may know the matroid.*

Proof. Fix $i \in E_+$ and $u \in (m_i, 1)$. Keep the matroid fixed and change only μ_i to u. Call the new vector μ'. Lemma C.1 gives a basis containing i with new mean $\mu^\star + u - m_i$, while every basis omitting i still has mean at most $\mu^\star$. Uniform efficiency and Equation 10 therefore imply, for every $a > 0$,

$$\mathbb{E}_\mu N_i(T) \leq R_\mu(T)/\Delta_i = o(T^a), \qquad \mathbb{E}_{\mu'}[T - N_i(T)] \leq R_{\mu'}(T)/(u - m_i) = o(T^a).$$

Set $E_T = \{N_i(T) \le T/2\}$, $p_T = \mathbb{P}_\mu(E_T)$, and $q_T = \mathbb{P}_{\mu'}(E_T)$. Markov's inequality gives, for $0 < a < 1$,

$$1 - p_T \le \frac{2\mathbb{E}_\mu N_i(T)}{T} \longrightarrow 0, \qquad q_T \le \frac{2\mathbb{E}_{\mu'}[T - N_i(T)]}{T} = o(T^{a-1}).$$

Since only coordinate i changes, Lemma B.2 yields

$$\mathbb{E}_\mu N_i(T)\,\mathrm{kl}(\mu_i, u) \ge \mathrm{kl}(p_T, q_T) \ge p_T \log(1/q_T) - \log 2.$$

The last inequality follows by bounding binary entropy by $\log 2$ and dropping the nonnegative term $(1 - p_T)\log(1/(1 - q_T))$. Since $p_T \to 1$ and $\liminf_T \log(1/q_T)/\log T \ge 1 - a$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\mu N_i(T)\,\mathrm{kl}(\mu_i, u)}{\log T} \ge 1 - a.$$

First let $a \downarrow 0$, then $u \downarrow m_i$. Continuity of KL gives Equation 39. Summing the coordinate bounds in Equation 10 proves the regret lower bound. Neither step assumes distinct means or an unknown matroid. □

Proof of Theorem 1. Call a round exceptional if it is in an incorrect episode, a repair batch, or an initial cover, or if $\mathcal{U}_t$ occurs. A round incurs regret at most r, so Lemmas D.3, D.4, and D.6 bound the total expected exceptional regret by

$$A_0 := r\mathbb{E}\{W_\infty + S_\infty + vK + Z\} < \infty. \tag{40}$$

The sum is an upper bound; the categories need not be disjoint.

On any remaining round, the stored basis B is optimal. Lemma C.1 gives $\mu_{j(i)} = m_i$ for every outside arm. Because the action chooses one member per group,

$$\mu^\star - \mu(A_t) = \sum_{i \in A_t \setminus B} (\mu_{j(i)} - \mu_i) = \sum_{i \in A_t \cap E_+} \Delta_i. \tag{41}$$

A selected $i \in E_+$ cannot belong to B, and its index satisfies

$$U_i(t) \ge U_{j(i)}(t) \ge m_i. \tag{42}$$

Thus its number $M_i(T)$ of nonexceptional selections is at most $L_i(m_i, T)$. The insertion threshold m_i is the same for every optimal basis. Applying Lemma D.7 once per arm therefore gives

$$R_\mu(T) \le A_0 + \sum_{i \in E_+} \Delta_i \mathbb{E} L_i(m_i, T) \le C^\star(\mu) f(T) + A_\mu(\sqrt{f(T)} + 1).$$

The constant is finite because there are finitely many arms and $I_i > 0$ on E_+. This logarithmic bound proves uniform efficiency at every mean vector. If $E_+ = \varnothing$, all nonexceptional losses vanish and $R_\mu(T) \le A_0$ for every T. This does not require every basis to be optimal: on $U_{2,3}$ with means $(.4, .4, .8)$, each arm belongs to an optimal basis, but $\{1, 2\}$ is suboptimal.

Combining this upper bound with Lemma D.8 and $f(T)/\log T \to 1$ gives the exact asymptotic coefficient. Lemmas D.3 and D.6 supply the query, memory, and last-rebuild claims. For later use, the same calculation shows that any exceptional set $\mathcal{X}$ satisfying $\mathbb{E}|\mathcal{X}| < \infty$ and Equation 41–Equation 42 outside $\mathcal{X}$ gives

$$R_\mu(T) \le r\mathbb{E}|\mathcal{X}| + \sum_{i \in E_+} \Delta_i \mathbb{E} L_i(m_i, T).$$

□

E QUERY LOWER BOUNDS

The family below has a common mean vector and optimal basis, but structure-dependent optimal sampling rates. These rates identify the hidden group labels.

E.1 SAMPLING RATES

Let $q = d - r$. Fix anchors $b_1, \dots, b_r$, outside elements $e_1, \dots, e_q$, and means

$$0 < p_1 < \cdots < p_q < \nu_1 < \cdots < \nu_r < 1.$$

For $\theta \in [r]^q$, let $\mathcal{M}_\theta$ be the capacity-one partition matroid with groups

$$G_j(\theta) = \{b_j\} \cup \{e_i : \theta_i = j\}, \qquad j \in [r].$$

These loopless rank-r instances share the means $\mu_{b_j} = \nu_j$, $\mu_{e_i} = p_i$, and the unique optimal basis $B^\star = \{b_1, \dots, b_r\}$. Moreover, $K_{B^\star}(e_i) = \{b_{\theta_i}\}$. Write

$$I_{\theta,i} = \mathrm{kl}(p_i, \nu_{\theta_i}), \qquad v_{\theta,i} = I_{\theta,i}^{-1}, \qquad \Delta_{\theta,i} = \nu_{\theta_i} - p_i > 0.$$

A feasible basis selects exactly one coordinate per group, so

$$R_\theta(T) = \sum_{i=1}^{q} \Delta_{\theta,i} \mathbb{E}_\theta N_{e_i}(T), \qquad C_\theta^\star = \sum_{i=1}^{q} \Delta_{\theta,i} v_{\theta,i}. \tag{43}$$

For $T \geq 2$, write $X_i(T) = N_{e_i}(T)/\log T$.

Lemma E.1 (Sampling rates under optimal regret)**.** *For a universally feasible, uniformly efficient policy, every θ, i and $\epsilon > 0$ satisfy*

$$\mathbb{P}_\theta\left(\frac{N_{e_i}(T)}{\log T} < v_{\theta,i} - \epsilon\right) \longrightarrow 0. \tag{44}$$

If, in addition, $R_\theta(T)/\log T \to C_\theta^\star$, then for every i,

$$\mathbb{E}_\theta\left|\frac{N_{e_i}(T)}{\log T} - v_{\theta,i}\right| \longrightarrow 0. \tag{45}$$

Proof. Lower sampling rate. Fix $b \in (0, v_{\theta,i})$. At $u = \nu_{\theta_i}$, $b\,\mathrm{kl}(p_i, u) < 1$. By continuity, choose $u \in (\nu_{\theta_i}, 1)$ close enough that this remains true, avoiding all other coordinate means. Keep $\mathcal{M}_\theta$ fixed and change only the mean of e_i to u. Its unique optimal basis now contains e_i, and omitting e_i costs at least $u - \nu_{\theta_i}$. Uniform efficiency implies

$$\mathbb{E}_{\theta,u}[T - N_{e_i}(T)] = o(T^a) \qquad (a > 0). \tag{46}$$

For the reward stream $(Y_n)_n$ of e_i, let

$$\ell_n = \sum_{s=1}^{n} \log \frac{p_i^{Y_s}(1-p_i)^{1-Y_s}}{u^{Y_s}(1-u)^{1-Y_s}}, \qquad \ell_0 = 0.$$

Under $\mathbb{P}_\theta$, its increments have mean $k = \mathrm{kl}(p_i, u)$ and range length $A = \log\{u(1-p_i)/[p_i(1-u)]\} > 0$. Choose $c \in (bk, 1)$. For $1 \leq n \leq b\log T$, the threshold $c\log T$ exceeds the mean nk by at least $(c-bk)\log T$. Hoeffding's inequality bounds $\mathbb{P}_\theta(\ell_n > c\log T)$ by $\exp\{-2(c-bk)^2\log^2 T/(nA^2)\}$. A union bound over the possible positive counts gives

$$\mathbb{P}_\theta\left(\max_{0 \leq n \leq \lfloor b\log T \rfloor} \ell_n > c\log T\right) \leq b\log T\; T^{-2(c-bk)^2/(bA^2)} = o(1). \tag{47}$$

For $E_T = \{N_{e_i}(T) \leq b\log T,\ \ell_{N_{e_i}(T)} \leq c\log T\}$, change of measure gives

$$\begin{aligned}
\mathbb{P}_\theta(E_T) &= \mathbb{E}_{\theta,u}[\mathbf{1}_{E_T} e^{\ell_{N_{e_i}(T)}}] \\
&\leq T^c \mathbb{P}_{\theta,u}(N_{e_i}(T) \leq b\log T) \\
&\leq \frac{T^c \mathbb{E}_{\theta,u}[T - N_{e_i}(T)]}{T - b\log T} = o(1).
\end{aligned}$$

For the last step choose $0 < a < 1 - c$ in Equation 46; the denominator is positive for large T. Since

$$\{N_{e_i}(T) \leq b\log T\} \subseteq E_T \cup \left\{\max_{0 \leq n \leq \lfloor b\log T \rfloor} \ell_n > c\log T\right\},$$

Equation 47 proves $\mathbb{P}_\theta(N_{e_i}(T) \le b \log T) \to 0$. Set $b = v_{\theta,i} - \epsilon$ when this is positive; otherwise the event in Equation 44 is empty.

L^1 convergence. Since $X_i(T) = N_{e_i}(T)/\log T \ge 0$, for every $\epsilon > 0$ we have

$$\mathbb{E}_\theta(v_{\theta,i} - X_i(T))_+ \le \epsilon + v_{\theta,i}\mathbb{P}_\theta(X_i(T) < v_{\theta,i} - \epsilon).$$

By Equation 44, taking $\limsup_{T\to\infty}$ and then $\epsilon \downarrow 0$ gives $\mathbb{E}_\theta(v_{\theta,i} - X_i(T))_+ \to 0$.

Writing the regret identity in Equation 43 in terms of positive and negative parts gives

$$\sum_i \Delta_{\theta,i}\mathbb{E}_\theta(X_i(T) - v_{\theta,i})_+ = \frac{R_\theta(T)}{\log T} - C^\star_\theta + \sum_i \Delta_{\theta,i}\mathbb{E}_\theta(v_{\theta,i} - X_i(T))_+ \longrightarrow 0.$$

Every term on the left is nonnegative and every weight $\Delta_{\theta,i}$ is positive. Hence each expected excess also vanishes. The expected absolute deviation is the sum of these two parts, proving Equation 45. □

E.2 Oracle information

Define $v_{i,j} = 1/\operatorname{kl}(p_i, \nu_j)$ and $\delta_i = \min_{j\neq k} |v_{i,j} - v_{i,k}| > 0$; positivity follows from

$$\partial_u \operatorname{kl}(p_i, u) = \frac{u - p_i}{u(1-u)} > 0 \qquad (u > p_i).$$

Let $\widehat{\theta}_{T,i}$ be the index whose target $v_{i,j}$ is closest to $X_i(T)$, with ties broken by index. An error on coordinate i requires $|X_i(T) - v_{\theta,i}| \ge \delta_i/2$. Markov's inequality, a union bound, and Equation 45 give

$$\mathbb{P}_\theta(\widehat{\theta}_T \neq \theta) \le \sum_{i=1}^{q} \frac{2}{\delta_i}\mathbb{E}_\theta|X_i(T) - v_{\theta,i}| \longrightarrow 0. \tag{48}$$

The decoder, unlike the policy, is given the common mean vector.

Lemma E.2 (Information in the oracle answer string)**.** *Let Θ be uniform on $[r]^q$, and let Ω, independent of Θ, contain the common reward streams and policy randomness. For the oracle answer string W_T at a deterministic horizon T,*

$$I(\Theta; \Omega, \mathsf{W}_T) \le \mathbb{E}Q_T. \tag{49}$$

Information and entropy are measured in bits.

Proof. Fix $\Omega = \omega$. Given the answers, the policy's actions, queries, query times, and observed counts are deterministic. For almost every ω, all attainable terminal strings are finite: the family has only r^q members, and each makes finitely many queries by T.

For this fixed ω, suppose both w and a proper extension wv were terminal answer strings. The two runs agree through the answers in w. Thereafter their actions and rewards still agree until a further query is requested. If one run reaches T without another query, the other must do the same, contradicting the extension. The strings are therefore prefix-free conditional on Ω. Query times are also determined by the answers and ω, so they require no additional symbols.

Write p_w for the conditional probability of string w. Prefix-freeness gives $Z_\omega = \sum_w 2^{-|w|} \le 1$ by Kraft's inequality. Nonnegativity of relative entropy between p_w and $2^{-|w|}/Z_\omega$ yields

$$H(\mathsf{W}_T \mid \Omega = \omega) \le \mathbb{E}[|\mathsf{W}_T| \mid \Omega = \omega] + \log_2 Z_\omega \le \mathbb{E}[Q_T \mid \Omega = \omega].$$

Averaging over Ω and using its independence from Θ gives

$$I(\Theta; \Omega, \mathsf{W}_T) = I(\Theta; \mathsf{W}_T \mid \Omega) \le H(\mathsf{W}_T \mid \Omega) \le \mathbb{E}Q_T.$$

□

Proof of Theorem 2. The decoder is measurable with respect to (Ω, W_T). Let $\eta_T = \mathbb{P}(\widehat{\theta}_T \neq \Theta)$; Equation 48 and finiteness of the family imply $\eta_T \to 0$. Set $D_T = \mathbf{1}\{\widehat{\theta}_T \neq \Theta\}$. Because D_T is determined by Θ and the decoder's information, the entropy chain rule gives

$$H(\Theta \mid \Omega, \mathsf{W}_T) = H(D_T \mid \Omega, \mathsf{W}_T) + H(\Theta \mid D_T, \Omega, \mathsf{W}_T).$$

The first entropy is at most $h_2(\eta_T)$. Conditional on the decoder's information, the second is zero when $D_T = 0$ and at most $\log_2(r^q - 1)$ when $D_T = 1$. Therefore

$$H(\Theta \mid \Omega, \mathsf{W}_T) \le h_2(\eta_T) + \eta_T \log_2(r^q - 1),$$

where h_2 is binary entropy. Since the uniform prior has $H(\Theta) = q \log_2 r$, Lemma E.2 gives

$$\frac{1}{r^q} \sum_{\theta \in [r]^q} \mathbb{E}_\theta Q_T \ge q \log_2 r - h_2(\eta_T) - \eta_T \log_2(r^q - 1).$$

As $T \to \infty$, the error terms vanish and monotone convergence applies to $Q_T \uparrow Q_\infty$:

$$\frac{1}{r^q} \sum_{\theta \in [r]^q} \mathbb{E}_\theta Q_\infty \ge q \log_2 r.$$

At least one instance therefore has expected query count at least $q \log_2 r$. The family and the one-coordinate alternatives used above all have distinct means, so the argument also holds when the model is restricted to that submodel. □

E.3 Observable completion

Let $\mathfrak{M}_{d,r}$ be the finite class of labeled loopless rank-r matroids. For a finite structural transcript H, define

$$\mathfrak{V}(H) = \{\mathcal{M} \in \mathfrak{M}_{d,r} : \mathcal{M} \text{ agrees with every answer in } H\},$$

and, when this version space is nonempty,

$$\mathcal{B}_\cap(H) = \bigcap_{\mathcal{M} \in \mathfrak{V}(H)} \mathcal{B}(\mathcal{M}), \qquad \mathcal{B}_\cup(H) = \bigcup_{\mathcal{M} \in \mathfrak{V}(H)} \mathcal{B}(\mathcal{M}).$$

A universal query-free continuation must be feasible and consistent ($R_\mu(T) = o(T)$) for every structure in $\mathfrak{V}(H)$ and every vector of independent Bernoulli rewards with means in $(0, 1)$. This is a zero-error requirement across instances, not a fixed-confidence identification guarantee.

Proposition 1 (Completion requires identification)**.** *A universal query-free continuation exists after a finite transcript H with nonempty version space if and only if $|\mathfrak{V}(H)| = 1$. Consequently, a universally feasible, consistent policy that observably disables all future queries must identify the entire matroid almost surely on the event of finite completion. On a hidden uniform matroid $U_{r,d}$, with $2 \le r < d$, every such completion requires at least $\binom{d}{r}$ queries.*

Proof. A fixed transcript. A matroid is determined by its bases, so

$$|\mathfrak{V}(H)| = 1 \iff \mathcal{B}_\cap(H) = \mathcal{B}_\cup(H).$$

If the version space is a singleton, enumerate its $M = |\mathcal{B}|$ bases. In epoch $[2^k, 2^{k+1})$, the policy explores each basis k times, then exploits the basis with largest empirical reward over all its designated exploration plays so far. Fix k_0 with $Mk \le 2^k$ for $k \ge k_0$; earlier rounds use any known basis.

For $k \ge 2k_0$, each basis has at least $k^2/4$ designated samples. If a suboptimal basis A with gap $\delta_A > 0$ beats an optimal basis empirically, one of their estimates deviates by at least $\delta_A/2$. Since basis rewards lie in $[0, r]$, Hoeffding bounds this probability by $2e^{-k^2\delta_A^2/(8r^2)}$. Exploration therefore costs $O(Mr \log^2 T)$, while the total expected exploitation loss is at most

$$\sum_{k \ge 2k_0} \sum_{A:\, \delta_A > 0} 2r\, 2^k e^{-k^2\delta_A^2/(8r^2)} < \infty.$$

The finitely many earlier epochs contribute a constant. Thus the continuation is feasible, query-free, and has $R_\mu(T) = o(T)$; if every basis is optimal, its regret is zero.

Conversely, suppose the version space contains two different structures. Choose $S \in \mathcal{B}_\cup(H) \setminus \mathcal{B}_\cap(H)$ and compatible matroids $\mathcal{M}_+, \mathcal{M}_-$ in which S is, respectively, a basis and not a basis. Give elements of S mean $3/4$ and all other elements mean $1/4$. In $\mathcal{M}_+$, S is uniquely optimal and every other

basis loses at least $1/2$. Couple the two continuations using identical reward streams and randomness. Without further queries their actions agree, so feasibility gives

$$\mathbb{P}_{\mathcal{M}_+,\lambda}(A_t = S) = \mathbb{P}_{\mathcal{M}_-,\lambda}(A_t = S) = 0 \qquad (t \geq 1),$$

where $\lambda_i = 3/4$ on S and $1/4$ off S. Consequently $R_{\mathcal{M}_+,\lambda}(T) \geq T/2$, which rules out consistency.

An observable completion time. Let τ_{comp} denote the interaction stopping time at which all future queries are disabled. Suppose some instance has

$$\mathbb{P}(\tau_{\mathrm{comp}} < \infty,\ |\mathfrak{V}(H_{\tau_{\mathrm{comp}}})| > 1) > 0.$$

Fix such an instance $(\mathcal{M}_0, \mu_0)$. There are countably many finite structural transcripts and integer horizons, so some transcript h with $|\mathfrak{V}(h)| > 1$ and some integer n make

$$E = \{\tau_{\mathrm{comp}} \text{ occurs by reward round } n,\ H_{\tau_{\mathrm{comp}}} = h\}$$

have positive probability. Choose compatible $\mathcal{M}_+, \mathcal{M}_-$ and an r-set S that is a basis only in the first. Put $\lambda_i = 3/4$ on S and $1/4$ elsewhere. On the fixed structure $\mathcal{M}_0$, the finite-horizon likelihood ratio has at most rn positive reward factors. Lemma B.2 therefore shows that E also has positive probability at means λ.

Couple all three structures using the same λ-reward streams and policy seed. Along a path in E, their answers agree with h. Induction over queries and reward rounds gives identical actions, observations, and completion decisions up to τ_{comp}. Universal feasibility lets us discard the null set on which any of these three runs plays an infeasible action. It follows that E has the same positive probability ϵ_0 under $\mathcal{M}_+$ and $\mathcal{M}_-$.

After τ_{comp}, actions continue to agree because no queries remain. Feasibility under $\mathcal{M}_-$ forbids S on E, whereas every other basis in $\mathcal{M}_+$ loses at least $1/2$. Hence

$$R_{\mathcal{M}_+,\lambda}(T) \geq \tfrac{1}{2}(T-n)\epsilon_0, \qquad T > n,$$

contradicting consistency. Observable finite completion must therefore have a singleton version space almost surely. The positive-probability event E is available because completion is observable; an unobserved last-query time does not give such an event.

Queries on the uniform matroid. For each $S \in \binom{[d]}{r}$, consider $\mathcal{B}_S = \binom{[d]}{r} \setminus \{S\}$. To verify the basis-exchange axiom, take $A, B \in \mathcal{B}_S$ and $x \in A \setminus B$. If $|B \setminus A| = 1$, the exchange returns B. Otherwise at least two replacement elements are available and at most one produces the excluded set S.

Every $(r-1)$-set has at least two rank-r extensions, so deleting S leaves all smaller sets independent. For $r \geq 2$ the resulting matroid is loopless. Its independence oracle differs from that of $U_{r,d}$ only on S (Jensen & Korte, 1982; Huang et al., 2024). Until S is queried, both structures remain compatible with the transcript. Identifying $U_{r,d}$ therefore requires all $\binom{d}{r}$ rank-r queries. Conversely, those queries determine all bases and hence the entire matroid. □

F Proof of Theorem 3

Table 4: Query bounds for the binary independence oracle. Both algorithms use $O(d)$ words. ECR-KL attains $C^\star(\mu)$ at every mean vector; Full-Order requires distinct means. Balanced rank means $r, d-r = \Theta(d)$; the lower bound holds on some instance with distinct means.

Method / regime	Expected total queries	Balanced-rank scaling
ECR-KL (ties allowed)	$\leq \dfrac{4dL}{d+L+1}\, c_{d,r}$	$O(d^2 \log d)$
Full-Order ECR-KL (distinct means)	$\leq \dfrac{113}{49}\, c_{d,r}$	$O(d \log d)$
Lower bound (distinct-means family)	$\geq (d-r)\log_2 r$	$\Omega(d \log d)$

Full-Order ECR-KL retains the cover, alternating scheduler, global KL indices, and local trigger, with $\alpha = \eta = 1/8$. After a trigger, it collects fresh cover cycles. Let $x_{i,n}$ be the mean of coordinate i's n

designated observations and set

$$b_n = \sqrt{\frac{\log(4dn^2/\eta)}{2n}}, \qquad J_{i,n} = [x_{i,n} - b_n, x_{i,n} + b_n].$$

The phase ends at

$$N = \inf\{n \geq 1 : J_{i,n} \cap J_{j,n} = \varnothing \text{ for every } i \neq j\}.$$

The intervals are then sorted from high to low and a new certificate is built. All observations remain in the global statistics; the next episode starts with zero local counts.

Lemma F.1 (One ordering phase)**.** *Let $\mathcal{H}$ be the history at the start of an ordering phase, N the number of cover cycles until it stops, and $\delta = \min_{i \neq j} |\mu_i - \mu_j| > 0$. The phase returns the true order with conditional probability at least $1 - \eta$. For n_0 such that $b_n \leq \delta/8$ whenever $n \geq n_0$,*

$$\mathbb{P}(N > n \mid \mathcal{H}) \leq 2de^{-n\delta^2/32}, \qquad n \geq n_0. \tag{50}$$

Consequently, $\mathbb{E}[N^s \mid \mathcal{H}] \leq C_{\mu,s} < \infty$ for every $s > 0$, uniformly over reached histories.

Proof. Conditional on $\mathcal{H}$, the cover is fixed and its designated observations are fixed subsequences of fresh independent Bernoulli streams. Define $x_{i,n}$ for all n by continuing the same cover after the stopping rule, without changing any observations made before stopping. Hoeffding and a union bound give

$$\mathbb{P}(\exists i, n : |x_{i,n} - \mu_i| > b_n \mid \mathcal{H}) \leq \sum_{i,n} 2e^{-2nb_n^2} = \frac{\eta}{2} \sum_{n \geq 1} n^{-2} < \eta.$$

On the complementary event, all intervals contain their means. If they are pairwise disjoint at return, their order must be the true mean order.

To bound the duration without conditioning on that confidence event, take $n \geq n_0$. If every sample mean is within $\delta/8$ of its true mean, any two interval centers are separated by at least $3\delta/4$, while their radii sum to at most $\delta/4$. The intervals are therefore disjoint and the phase has stopped by n. Thus

$$\{N > n\} \subseteq \bigcup_i \{|x_{i,n} - \mu_i| > \delta/8\}.$$

Hoeffding gives Equation 50. Its right-hand side tends to zero, proving almost-sure termination. For $s > 0$, the tail-sum identity yields

$$\mathbb{E}[N^s \mid \mathcal{H}] = \sum_{n \geq 0} ((n+1)^s - n^s)\mathbb{P}(N > n \mid \mathcal{H}) < \infty.$$

The first n_0 terms sum to at most n_0^s; the exponential tail bounds the rest uniformly over $\mathcal{H}$. □

Proof of Theorem 3. Build count. Let J count reached ordering phases and let $K = 1 + J$ count builds. By Lemma F.1, every reached phase terminates almost surely; a countable union handles all phases. At a phase start $\mathcal{H}$, let $p_{\mathrm{err}}(\mathcal{H}) \leq \eta$ be the probability of returning an incorrect order. A correct returned order gives a correct certificate at the next build. Conditional on that build history, freshness and the local error bound give a later-trigger probability at most $(\alpha/2)\sum_{\ell_{\mathrm{ep}} \geq 1} \ell_{\mathrm{ep}}^{-5} < \alpha$, now with $\alpha = 1/8$.

Conditioning on whether the returned order is correct therefore gives

$$\mathbb{P}(\text{another ordering} \mid \mathcal{H}) \leq p_{\mathrm{err}}(\mathcal{H}) + (1 - p_{\mathrm{err}}(\mathcal{H}))\alpha \leq p_{\mathrm{next}} := 15/64.$$

Conditioning successively at the reached phase starts gives

$$\mathbb{P}(J \geq j) \leq p_{\mathrm{next}}^{j-1}, \qquad \mathbb{E}K \leq 1 + \frac{1}{1 - p_{\mathrm{next}}} = \frac{113}{49}.$$

Including the initial build, $\mathbb{E}Q_\infty \leq (113/49)c_{d,r}$, and K has every finite moment.

Transient duration. Let $\mathcal{O}_j$ be the event that phase j is reached, $\mathcal{H}_j$ its starting history, and vN_j its duration. For total ordering time V, the uniform conditional duration bound gives

$$\mathbb{E}V = v \sum_j \mathbb{E}[\mathbf{1}_{\mathcal{O}_j} \mathbb{E}(N_j \mid \mathcal{H}_j)] \leq vC_{\mu,1}\mathbb{E}J < \infty.$$

For a correct episode, the local confidence calculation gives $\mathbb{E}Y \le \alpha \mathbb{E}K$, now with $\alpha = 1/8$. Lemma D.5 uses α only through Λ, so its detection estimate holds with this choice as well. At the last incorrect round,

$$s_T \le W_T + Y + V + 1.$$

Moreover,

$$\mathbb{E}\left(KH + \sum_{k \le K} H_k + Z + K \right) \le (\mathbb{E}K^2\, \mathbb{E}H^2)^{1/2} + C_\mu \mathbb{E}K + \mathbb{E}Z < \infty.$$

Thus Equation 35 gives $\mathbb{E}W_\infty < \infty$. Since

$$\tau_{\text{last}} \le W_\infty + Y + V,$$

its expectation is finite. The final certificate must be correct: otherwise it would produce infinitely many incorrect learning rounds.

Regret. Distinct means imply a unique optimal basis. Indeed, if B is optimal and $A \neq B$, exchange matching gives

$$\mu(A) - \mu(B) = \sum_{i \in A \setminus B} (\mu_i - \mu_{\phi(i)}) < 0.$$

Optimality makes every summand nonpositive; distinct coordinate means make each strictly negative. With $A_0^{\text{ord}} = r\mathbb{E}(W_\infty + V + vK + Z) < \infty$, the exceptional-round decomposition and Lemma D.7 give

$$R_\mu(T) \le A_0^{\text{ord}} + \sum_{i \notin B^\star} \Delta_i \left\{ \frac{f(T)}{I_i} + C_{\mu_i, m_i}(\sqrt{f(T)} + 1) \right\}.$$

This proves uniform efficiency on the distinct-means submodel. The alternatives in Lemma D.8 can also be chosen with distinct means, so the matching lower bound proves Equation 8. Stored arrays and ordering workspace use $O(d)$ words. At balanced rank, $c_{d,r} = \Theta(d \log d)$, matching the query lower bound in Theorem 2. □

Known number of mean levels. If L is supplied, stop the ordering phase when the interval-overlap graph has at least L components, rather than requiring pairwise-disjoint intervals. On the simultaneous confidence event, all coordinates with the same mean belong to one component. There can therefore be at most L components. At stopping the number of components is both at least and at most L. Since no level is split, the L components correspond one-to-one to the L levels, and their disjoint hulls determine the order.

For $L > 1$, let δ be the minimum gap between distinct levels and take $n \ge n_0$ with $b_n \le \delta/8$. On $\max_i |x_{i,n} - \mu_i| \le \delta/8$, intervals from different levels are disjoint, so the graph has at least L components, even if intervals within a level have split. Hence

$$\{N > n\} \subseteq \{\max_i |x_{i,n} - \mu_i| > \delta/8\}, \qquad \mathbb{P}(N > n \mid \mathcal{H}) \le 2de^{-n\delta^2/32}.$$

The simultaneous confidence event ensures the returned level order is correct, and any priority within a level is valid. Thus the next-phase probability is still at most $15/64$, and the preceding build-count and duration bounds apply.

For $L = 1$, every certificate is correct and every basis has the same reward. Keep the initial certificate and ignore repair triggers: no ordering samples or later queries are needed, and regret is zero. Thus the regret upper bound and finite expected stabilization hold with known L. Supplying L changes the information model; no matching regret lower bound for this stronger model is asserted here.

G Algorithm variants

G.1 Sampling

The following changes to learning and repair preserve Theorem 1, including its query bound. Each repair schedule below can be paired with serial actions or either groupwise scheduler, with a possibly different finite regret remainder. Write $q = d - r$ and $v = q + 1$.

Serial actions. On an optimistic round, let

$$i_t = \min \arg\max_{i \notin B} \{U_i(t) - U_{j_\sigma(i)}(t)\}, \qquad A_t = \begin{cases} B - j_\sigma(i_t) + i_t, & U_{i_t}(t) > U_{j_\sigma(i_t)}(t), \\ B, & \text{otherwise.} \end{cases}$$

The initial cover and forced-basis rounds are unchanged. In the ordered-chain case of Lemma D.5, the improving arm a satisfies

$$U_{i_t}(t) - U_{j(i_t)}(t) \geq U_a(t) - U_{j(a)}(t) \geq 15\gamma/16 > \gamma/2$$

on each late optimistic round outside $\mathcal{U}_t$. Hence $i_t \in A_t \setminus B$ satisfies Equation 27. Each charge increases its count, and Equation 28 bounds the nonimproving charges. The basis quota Equation 24 is unchanged, so both cases of the detection lemma hold.

When the certificate is correct, a nontrivial exchange has loss Δ_{i_t} and, outside $\mathcal{U}_t$, satisfies $U_{i_t}(t) > U_{j(i_t)}(t) \geq m_{i_t}$. Thus Equation 41–Equation 42 hold for this action. The local error and refinement bounds are independent of the number of simultaneous exchanges. Lemmas D.3, D.6, and D.7 therefore give the asserted guarantees. With Full-Order repair, the build and ordering-time bounds in Appendix F instead give Theorem 3. The experiments call these policies Serial ECR-KL and Serial Full-Order.

Representative quotas. For parallel actions, whole-basis alternation can be replaced by a separate quota in each group. Let $N_k^{\mathrm{rep}}(u)$ count selections of b_k in the first u post-cover learning rounds, with $N_k^{\mathrm{rep}}(0) = 0$, and put $q_{\mathrm{rep}}(u) = \lceil u/2 \rceil$. On round $u+1$, select b_k if $N_k^{\mathrm{rep}}(u) < q_{\mathrm{rep}}(u+1)$; otherwise select a maximum-index member of G_k. Reset the counters at each build.

Induction gives $N_k^{\mathrm{rep}}(u) \geq q_{\mathrm{rep}}(u)$. A forced selection therefore requires $q_{\mathrm{rep}}(u+1) > q_{\mathrm{rep}}(u)$ and can occur only on an odd post-cover round. If $M_k(u, w)$ is the number of maximization rounds for group k in $(u, w]$, then

$$M_k(u, w) \geq (w - u) - [q_{\mathrm{rep}}(w) - q_{\mathrm{rep}}(u)] \geq \frac{w-u}{2} - 1. \tag{51}$$

For episode age $\ell_{\mathrm{ep}} \geq 8(v+3)$, put $u_0 = \lfloor \ell_{\mathrm{ep}}/2 \rfloor - v$. Before its second half, every representative has at least $1 + q_{\mathrm{rep}}(u_0) \geq \ell_{\mathrm{ep}}/8$ observations, and in that half each group has at least

$$M_k(u_0, \ell_{\mathrm{ep}} - v) \geq \frac{\lceil \ell_{\mathrm{ep}}/2 \rceil}{2} - 1 \geq \ell_{\mathrm{ep}}/8$$

maximization rounds. Under Equation 25, the basis-index bounds of Lemma D.5 follow. In the group containing its improving arm a, each maximizer outside $\mathcal{U}_t$ has gain at least $15\gamma/16$ and is observed. Charging these selections gives Equation 27 and the budget B_q in Equation 28. If $\ell_{\mathrm{ep}} \geq 16[Z_k(\ell_{\mathrm{ep}}) + B_q + 1]$, then

$$\#\{\text{improving charges}\} \geq \ell_{\mathrm{ep}}/8 - Z_k(\ell_{\mathrm{ep}}) - B_q \geq \ell_{\mathrm{ep}}/16.$$

Some improving arm therefore has at least $\ell_{\mathrm{ep}}/(16q)$ local observations. Under the additional thresholds $\ell_{\mathrm{ep}} \geq 32qH_k$ and $\ell_{\mathrm{ep}} \geq 512q\Lambda(\ell_{\mathrm{ep}})/\gamma^2$, Equation 29 forces a trigger. Inverted basis edges use only the representative counts. Thus both cases of Lemma D.5 apply to the quota rule.

Every action is a group transversal, so Equation 41 holds whenever the certificate is correct. On such post-cover rounds outside $\mathcal{U}_t$, each selected $i \in E_+$ is an unforced group maximizer and satisfies $U_i(t) \geq U_{j(i)}(t) \geq m_i$. Forced representatives have zero loss. The retained-count bound therefore applies, with the cover and underestimation rounds included in the finite-mean exceptional set. If the initial cover has length $h_\sigma \leq v$, replace v by h_σ in u_0; both observation and maximization lower bounds can only increase.

Targeted repair. At a trigger (a, b) with scale g, put $w = g/(4d)$ and

$$m_0 = \left\lceil \frac{\log(4d/\eta)}{2w^2} \right\rceil.$$

This is the uniform batch size at error level $\eta/2$, not the smaller core batch in Equation 6. All phase comparisons below use this same m_0. Let $\mathcal{H}$ be the candidate-start history and $\widetilde{\mu}_i$ any finite, $\mathcal{H}$-measurable estimates. For the closed interval J between $\widetilde{\mu}_a$ and $\widetilde{\mu}_b$, define

$$w_i = \max\{w, \mathrm{dist}(\widetilde{\mu}_i, J)/4\}, \qquad h_i = \left\lceil \frac{\log(4d/\eta)}{2w_i^2} \right\rceil \leq m_0.$$

The candidate plays $B - j(i) + i$ exactly h_i times for each outside arm, in label order, followed by m_B plays of B, where

$$c_j = \sum_{i \notin B} h_i \mathbf{1}\{j \neq j(i)\}, \qquad m_B = \max_{j \in B} (h_j - c_j)_+ \leq m_0.$$

The resulting fresh counts are exactly h_i for $i \notin B$ and $c_j + m_B \geq h_j$ for $j \in B$. The first h_i observations of each coordinate form x_i and its interval $[x_i - w_i, x_i + w_i]$. All observations also enter the global statistics; the episode-local counts n_i are not updated during repair.

The parent partition is the current partition when a, b share a block, and $([d])$ otherwise. Within each parent, the candidate consists of the interval-overlap components. It is accepted exactly when it separates a, b. On rejection, a fresh uniform batch of m_0 cover cycles replaces the candidate, using radius w and the same parents; this fallback is accepted regardless of separation. In either case the component hulls are ordered from high to low, the parent order is retained, and the old priority is preserved within each component. At most two finite batches are taken, including at a false trigger involving equal means.

Let F be the fallback indicator and $\mathcal{E}_c, \mathcal{E}_f$ the respective confidence failures, with $\mathcal{E}_f = \varnothing$ when $F = 0$. Lemma D.1 and Hoeffding give

$$\mathbb{P}(\mathcal{E}_c \mid \mathcal{H}) \leq \sum_i 2e^{-2h_i w_i^2} \leq \eta/2, \qquad \mathbb{P}(\mathcal{E}_f \mid \mathcal{H}) \leq (\eta/2)\mathbb{P}(F = 1 \mid \mathcal{H}).$$

The second inequality conditions first on the reached fallback history. Hence $\mathbb{P}(\mathcal{E}_c \cup \mathcal{E}_f \mid \mathcal{H}) \leq \eta$, without an accuracy assumption on $\widetilde{\mu}$.

On the confidence event, equal means within a parent lie in intersecting intervals, and disjoint component hulls order the means. A good trigger has $\mu_a - \mu_b > g$ by Equation 14. An accepted candidate separates this pair by definition. If the fallback did not separate it, an overlap path would give

$$\mu_a - \mu_b \leq |x_a - x_b| + 2w \leq 2wd = g/2,$$

a contradiction. Thus Lemma 2 applies, with conditional bad-refinement probability at most $\alpha + \eta$.

For a phase starting from a fixed history, write n_A^c for candidate play counts and $\Delta(A) = \mu^\star - \mu(A) \geq 0$. Since $n_A^c \leq m_0$ for every $A \in \mathcal{C}_\sigma$,

$$S_c = \sum_A n_A^c \leq v m_0 = S_u, \qquad \mathcal{R}_c = \sum_A n_A^c \Delta(A) \leq m_0 \sum_A \Delta(A) = \mathcal{R}_u.$$

A fallback adds one uniform batch, so

$$S_{\text{target}} \leq (1 + F) S_u \leq 2S_u, \qquad \mathcal{R}_{\text{target}} \leq (1 + F)\mathcal{R}_u \leq 2\mathcal{R}_u. \tag{52}$$

These compare phases from the same history, not complete policies; the stored basis need not be optimal.

The trigger inequality $g^2 > 2\Lambda(D_k)/D_k$ also yields

$$S_k^{\text{target}} \leq \kappa_d^{\text{target}} D_k, \qquad \kappa_d^{\text{target}} = 2v \left(1 + \frac{4d^2 \log(4d/\eta)}{\log(4d/\alpha)}\right). \tag{53}$$

Hence $\mathbb{E}K \leq 4dL/(d + L + 1)$ and Equation 17 still bounds the tail of K. The local test gives $\mathbb{E}Y \leq \alpha \mathbb{E}K$. Moreover,

$$S_T \leq \kappa_d^{\text{target}}(W_T + Y), \qquad s_T \leq (1 + \kappa_d^{\text{target}})(W_T + Y) + 1.$$

By Equation 35, $\mathbb{E}(W_\infty + S_\infty) < \infty$, and $\tau_{\text{last}} \leq W_\infty + Y + S_\infty$ has finite expectation. After also excluding initial covers and index-underestimation rounds, Equation 41–Equation 42 hold on every remaining round. The expected number of excluded rounds is at most $\mathbb{E}(W_\infty + S_\infty + vK + Z) < \infty$, so Lemma D.7 gives the regret bound with all repair observations retained.

Packed sampling. Fix the groups $G_1, \dots, G_r$ and integer demands $h_i \geq 1$ chosen at the sampling start. Set

$$H_\times = \max_{k \leq r} \sum_{i \in G_k} h_i.$$

Within each group, select the outside elements consecutively in label order, h_i times each, and its representative in all remaining slots. Combining the group schedules gives $H_\times$ simultaneous actions. Every outside element is observed exactly h_i times, while b_k is observed $H_\times - \sum_{i \in G_k \setminus \{b_k\}} h_i \geq h_{b_k}$ times. Designate the first h_i observations of each coordinate and retain all observations globally.

Proposition 2 (Packed sampling)**.** *The schedule is feasible and uses the minimum number of rounds among schedules in $\mathcal{B}_\sigma$ meeting the prescribed counts. Let a serial schedule start from the same history, meet all prescribed counts, take exactly h_i observations of each outside arm, and otherwise play the stored basis. If the serial duration is H_{ser}, then*

$$H_\times \leq H_{\mathrm{ser}}, \qquad R_{\mathrm{ser}} - R_\times = (H_{\mathrm{ser}} - H_\times)[\mu^\star - \mu(B)] \geq 0.$$

Here $R_{\mathrm{ser}}, R_\times$ are the cumulative pseudo-regrets of these two phases, not expected regrets of complete policies.

Proof. A transversal observes one member of each group per round, so its duration is at least $\sum_{i \in G_k} h_i$ for every k. The construction attains their maximum and is feasible by Lemma 1. For either schedule of length H, summing the replacement losses gives

$$R = H[\mu^\star - \mu(B)] + \sum_{i \notin B} h_i(\mu_{j(i)} - \mu_i).$$

The second term cancels on subtraction. The sign follows from $H_{\mathrm{ser}} \geq H_\times$ and $\mu^\star \geq \mu(B)$, including when some exchanges have positive gain. □

For a uniform cover, list each group with its representative first, align corresponding positions, and fill missing positions with the representative. This uses $h_\sigma = \max_k |G_k| \leq v$ rounds. Designate the first-slot representative observations and each outside arm's unique observation. A uniform repair then takes $m_{\mathrm{rep}} h_\sigma$ rather than $m_{\mathrm{rep}} v$ rounds. For $d = 32, r = 16$ and groups of size two, this reduces the duration from $17 m_{\mathrm{rep}}$ to $2 m_{\mathrm{rep}}$ rounds. If B is optimal, however, the phase regrets are equal by Proposition 2.

Packed targeted repair uses the demands and radii above and, on rejection, a fresh packed uniform fallback. Both schedules are fixed at their respective starts, so the two conditional error bounds still sum to at most η. The interval rule is unchanged. Proposition 2 bounds the duration of each reached batch. Thus

$$S_k^{\mathrm{packed}} \leq \begin{cases} v m_{\mathrm{rep}}, & \text{uniform repair,} \\ 2 v m_0, & \text{targeted repair,} \end{cases} \leq \kappa D_k, \qquad \kappa \in \{\kappa_d, \kappa_d^{\mathrm{target}}\}.$$

After a packed cover, every coordinate has at least one local observation. Its length $h_\sigma \leq v$ can only increase the post-cover opportunities available by a fixed episode age. Thus the count bounds in Lemma D.5 hold for each of the learning rules above. Together with $S_T \leq \kappa(W_T + Y)$, they give Equation 35 and finite expected last-build time. On correct post-cover learning rounds outside $\mathcal{U}_t$, the loss and index inequalities give the same regret bound; the expected number of excluded rounds is again at most $\mathbb{E}(W_\infty + S_\infty + vK + Z)$.

G.2 Rebuilds

Build reuse. Let $a_k = \arg\min_{i \in G_k} \mathrm{pos}_\pi(i)$ be the first member of G_k in a new priority π. If $\mathrm{pos}_\pi(a_1) < \cdots < \mathrm{pos}_\pi(a_r)$, the build returns

$$B' = (a_1, \dots, a_r), \qquad j'(i) = a_k \quad (i \in G_k \setminus \{a_k\}). \tag{54}$$

Indeed, for $F_k = \mathrm{cl}(\{b_1, \dots, b_k\})$, the transversal induction in Appendix C gives

$$\mathrm{cl}(\{a_1, \dots, a_k\}) = F_k, \qquad a_k \in F_k \setminus F_{k-1}.$$

Greedy therefore accepts a_k at the first occurrence of G_k and rejects subsequent members of represented groups. Every $i \in G_k \setminus \{a_k\}$ first enters a prefix closure at F_k, so its binary-search partner is a_k. Two scans recover Equation 54 in $O(d)$ time and words, without queries.

This sufficient reuse test need not detect every query-free build. A successful reuse still resets local statistics and counts as a logical build; it also spends a direct-rebuild credit when applicable. Thus the logical-build bounds remain valid. Avoided requests can change later eviction-cache contents, so external-call savings are measured separately.

Greedy with a structural cache. The exact greedy baseline stores the partition reduction from its latest queried build. In a new decreasing-index order, it accepts a group's first occurrence. A subsequent member of G_k is rejected without a query only when all groups $G_1, \dots, G_k$ have appeared. The scan stops at r acceptances; a failed rejection test causes a full queried greedy build. Capacity-one recognition is attempted once. After success, every repeated-group element can be rejected regardless of group order.

Every set of accepted first occurrences extends to a transversal in $\mathcal{B}_\sigma$ and is therefore independent. When a repeated element $i \in G_k$ passes the rejection test, the accepted set A contains representatives $a_1, \dots, a_k$, so

$$i \in G_k \subseteq F_k = \mathrm{cl}(\{a_1, \dots, a_k\}) \subseteq \mathrm{cl}(A).$$

Thus $A + i$ is dependent. Both query-free decisions agree with greedy; a failed verification is resolved by the full build.

Warm starts.

Proposition 3 (Initialization with a finite prefix)**.** *Fix deterministic finite constants h_0, Q_0. Let a universally feasible prefix policy take h_0 reward rounds and at most Q_0 independence queries. Then initialize ECR-KL with a history-measurable priority, one unresolved block, and all prefix reward counts and sums retained. Its total expected query count satisfies*

$$\mathbb{E}Q_\infty \le Q_0 + \frac{4dL}{d+L+1} c_{d,r}.$$

The expected time of its last rebuild is finite. It is uniformly efficient and satisfies Equation 8 at every mean vector, with a possibly different finite constant A_μ.

Proof. The reward clock continues at $t = h_0 + 1$. Let $\mathcal{H}_0$ be the history after the first post-prefix build, and let K count that build and its successors. Freshness and the one-block initialization give, by Lemma D.3,

$$\mathbb{E}[K \mid \mathcal{H}_0] \le \frac{4dL}{d+L+1}, \qquad Q_\infty \le Q_0 + c_{d,r} K,$$

with a uniform conditional exponential tail for K.

For nonconstant means, define H from the full reward streams, including the prefix. Its unconditional tail is Equation 20, even though prefix estimates may be biased. Consequently,

$$\mathbb{E}(KH) \le (\mathbb{E}K^2\, \mathbb{E}H^2)^{1/2} < \infty, \qquad \mathbb{E} \sum_{k \le K} H_k \le C_\mu \mathbb{E}K < \infty.$$

The global underestimation count satisfies $\mathbb{E}Z < \infty$, so the remainder in Equation 32 is integrable. Correct episodes still satisfy $\mathbb{E}Y \le \alpha \mathbb{E}K$, and

$$s_T \le h_0 + (1 + \kappa_d)(W_T + Y) + 1.$$

Equations Equation 34–Equation 35 and Equation 19 give $\mathbb{E}(W_\infty + S_\infty) < \infty$. For constant means, the same conclusion follows from $W_\infty = 0$ and $S_\infty \le \kappa_d Y$. Thus

$$\mathbb{E}\tau_{\mathrm{last}} \le h_0 + \mathbb{E}(W_\infty + Y + S_\infty) < \infty.$$

The prefix, incorrect episodes, repairs, covers, and index-underestimation rounds cost at most $r[h_0 + \mathbb{E}(W_\infty + S_\infty + vK + Z)] < \infty$ in expected regret. On all other rounds, Equation 41–Equation 42 hold, and the retained-count bound includes the prefix observations. This proves the regret bound and uniform efficiency. For targeted repair, κ_d is replaced by $\kappa_d^{\mathrm{target}}$ in the same inequalities. □

The experimental prefixes and their charged query costs are specified in Appendix H.3.

Direct rebuilds. Fix a budget of $b \geq 0$ credits. While a credit remains, a local trigger replaces the repair batch by an empirical-mean-order build: sort the global means, reset the partition to one block, and spend one credit. The build or its exact reuse test certifies feasibility, without an accuracy assumption on the ordering. After the credits are exhausted, ordinary refinement is used.

A global test can also spend a credit. With counts after round t, define

$$\mathcal{L}_i(u;t) = u^{S_i(t)}(1-u)^{N_i(t)-S_i(t)}, \qquad \mathcal{L}_i^{\mathrm{mix}}(t) = \frac{\mathrm{B}(S_i(t)+1/2, N_i(t)-S_i(t)+1/2)}{\mathrm{B}(1/2,1/2)},$$

where B is the beta function and $0^0 = 1$. An edge (a, b) with positive counts rejects when

$$E_{ab}(t) = \frac{\mathcal{L}_a^{\mathrm{mix}}(t)\mathcal{L}_b^{\mathrm{mix}}(t)}{\sup_{0\leq u\leq v\leq 1}\mathcal{L}_a(u;t)\mathcal{L}_b(v;t)} > 8d(d-1)(t+1)^3. \tag{55}$$

For $\bar{\mu}_a \leq \bar{\mu}_b$, the constrained maximum is the unconstrained one and $E_{ab}(t) \leq 1$. Otherwise, concavity places the maximum at $u = v = (S_a(t) + S_b(t))/(N_a(t) + N_b(t))$. After the cover and each learning round, global edges are scanned in certificate order before the local test. With no credit, the global test is disabled. This beta-mixture construction is standard in sequential inference (Howard et al., 2020; 2021).

Proposition 4 (Direct rebuilds with global tests)**.** *With b fixed credits, either the parallel or serial action rule, combined with uniform or targeted refinement, retains uniform efficiency, Equation 8, and finite expected last-build time. The resulting policy has expected query count*

$$\mathbb{E}Q_\infty \leq \left(b + \frac{4dL}{d+L+1}\right) c_{d,r}.$$

A prefix as in Proposition 3 adds at most Q_0 queries and a finite regret term.

Proof. For $\mu_a \leq \mu_b$, the true parameters are feasible in the denominator of Equation 55; hence

$$E_{ab}(t) \leq W_{ab}(t) := \frac{\mathcal{L}_a^{\mathrm{mix}}(t)\mathcal{L}_b^{\mathrm{mix}}(t)}{\mathcal{L}_a(\mu_a;t)\mathcal{L}_b(\mu_b;t)}.$$

For fixed $u, v \in (0,1)$, define

$$L_t^{u,v} = \frac{\mathcal{L}_a(u;t)\mathcal{L}_b(v;t)}{\mathcal{L}_a(\mu_a;t)\mathcal{L}_b(\mu_b;t)}.$$

Conditional on the action-selection history $\mathcal{H}_t$, each observed reward contributes a likelihood-ratio factor of expectation one. The two factors are independent when both arms are selected, so $\mathbb{E}[L_t^{u,v} \mid \mathcal{H}_t] = L_{t-1}^{u,v}$ and $\mathbb{E}L_t^{u,v} = 1$. For the beta$(1/2, 1/2)$ law π, Tonelli gives

$$W_{ab}(t) = \iint L_t^{u,v}\,\pi(du)\pi(dv), \qquad \mathbb{E}W_{ab}(t) = \iint 1\,\pi(du)\pi(dv) = 1.$$

Let F_t be the event of any false global rejection at t. Markov's inequality and a union bound over all true ordered-pair nulls give

$$\mathbb{P}(F_t) \leq \frac{1}{8(t+1)^3}, \qquad \mathbb{E}\sum_{t\geq 1} t\mathbf{1}_{F_t} \leq \frac{1}{8}\sum_{t\geq 1}\frac{t}{(t+1)^3} = \frac{\zeta(2)-\zeta(3)}{8} < \infty. \tag{56}$$

This union includes the adaptively selected certificate edges.

Every trigger spends a credit before any ordinary refinement occurs. On the event $\mathcal{C}$ that all credits are used, let K' count the final credited build and its successors, and let $\mathcal{H}_c$ be that build's history. Its partition is one block, so on $\mathcal{C}$,

$$\mathbb{E}[K' \mid \mathcal{H}_c] \leq \frac{4dL}{d+L+1},$$

and the conditional tail satisfies Equation 17. If credits are never exhausted, set $K' = 0$; then $K \leq b$. For $b = 0$, set $K' = K$. In all cases,

$$K \leq b + K', \qquad \mathbb{E}K \leq b + \frac{4dL}{d+L+1}, \qquad \mathbb{E}K^s < \infty \quad (s > 0).$$

This proves the query bound.

Let Y_{loc} and Y_{glob} sum the durations of correct episodes ending at local and global triggers. Ending locally at age ℓ_{ep} requires the local confidence failure at that age. A correct episode ending globally at round t has duration at most t and implies F_t. Therefore

$$\mathbb{E}Y_{\mathrm{loc}} \le \alpha \mathbb{E}K, \qquad Y_{\mathrm{glob}} \le \sum_{t\ge 1} t\mathbf{1}_{F_t}, \qquad \mathbb{E}Y < \infty, \quad Y := Y_{\mathrm{loc}} + Y_{\mathrm{glob}}.$$

A global trigger only truncates the learning trajectory. Applying Lemma D.5 before its terminal round and adding that round gives Equation 32. Its remainder is integrable by the moments of K, Equation 20, freshness, and $\mathbb{E}Z < \infty$. Direct builds take no reward rounds; other repairs satisfy

$$S_T \le \kappa(W_T + Y), \qquad s_T \le (1+\kappa)(W_T + Y) + 1, \qquad \kappa \in \{\kappa_d, \kappa_d^{\mathrm{target}}\}.$$

Equations Equation 34–Equation 35 imply $\mathbb{E}(W_\infty + S_\infty) < \infty$ for nonconstant means. For constant means, $W_\infty = 0$ and $S_\infty \le \kappa Y$. In either case $\mathbb{E}\tau_{\mathrm{last}} \le \mathbb{E}(W_\infty + Y + S_\infty) < \infty$.

A prefix adds h_0 to the elapsed-time bound, Q_0 to queries, and at most rh_0 to regret. Correct post-cover learning rounds outside $\mathcal{U}_t$ satisfy Equation 41–Equation 42. The other rounds have finite expected count, bounded by $h_0 + \mathbb{E}(W_\infty + S_\infty + vK + Z)$, so Lemma D.7 gives the regret bound. The time-weighted estimate Equation 56, not merely the finite credit count, controls the last-build time. □

A fixed finite list of deterministic refresh times is also allowed. Each refresh retains global observations, discards any unfinished refinement decision, resets to one block, and builds without spending an adaptive credit. Let h_* be the final refresh time and m the number of scheduled refreshes. Through h_*, ordinary builds are at most h_*, because each follows a new reward round; scheduled and credited builds are at most $m + b$. Including one initial build gives the deterministic prefix budget

$$Q_{\mathrm{prefix}} \le (h_* + m + b + 1)c_{d,r}.$$

A separate initialization adds its own query budget. Proposition 4 then applies after h_* with at most b remaining credits. A branch that already knows the full structure may skip refreshes. Appendix H.2 specifies the experimental schedule.

G.3 Recognition

The stored groups always certify a partition reduction. The following tests determine when additional queries identify the entire matroid.

Proposition 5 (Capacity-one partitions)**.** *After an ordinary build, query $\{i, j_\sigma(i)\}$ for every $i \notin B$. If all pairs are dependent, the unknown matroid is exactly the capacity-one partition matroid with groups $G_1, \dots, G_r$. The test succeeds on every such partition matroid and has no false positives. Its total cost is at most $c_{d,r} + d - r$ queries and $O(d)$ words.*

Proof. In a loopless matroid, a dependent pair consists of parallel elements. On success, every element of G_k is parallel to b_k, so an independent set takes at most one element per group. Conversely, every such set extends to a transversal, which is a basis by Lemma 1. Thus

$$\mathcal{I} = \{A \subseteq E : |A \cap G_k| \le 1 \text{ for all } k\}.$$

In a capacity-one partition matroid, the only feasible removal for an outside arm is its group's representative; all queried pairs are therefore dependent. The test uses $d - r$ calls after a build of cost at most $c_{d,r}$, and stores only the groups and a temporary pair. □

After success, a finite group cover followed by a maximum-KL-index choice in each group requires no further queries. On $\mathcal{U}_t^c$, each selected suboptimal i satisfies $U_i(t) \ge \max_{j\in G(i)} \mu_j = m_i$, and the action loss is $\sum_{i\in A_t\cap E_+} \Delta_i$. Lemmas D.7 and D.4 give Equation 8, including at ties, with the fixed-structure lower bound of Lemma D.8. This is the scalar KL-UCB argument in each group (Garivier & Cappé, 2011).

A matroid component is an equivalence class generated by elements on a common circuit; a coloop forms a singleton. For a basis B, the graph joining each $i \notin B$ to every member of $K_B(i)$ has these components (Huang et al., 2023, Lemma 8). Identifying the restrictions also requires their bases.

Proposition 6 (Bounded-size components). *Fix $2 \le s \le d$. A deterministic oracle procedure returns the exact matroid as a direct sum of restrictions of size at most s, or returns failure. It succeeds exactly when every component has size at most s. On every input its cost is at most*

$$Q_{\mathrm{rec}} \le c_{d,r} + (d-r)(s-1)(1+\lceil \log_2 r \rceil) + \frac{d2^s}{s}. \tag{57}$$

For fixed s, its storage is $O_s(d)$ words and its query count is $O_s(d \log d)$. It uses neither rewards nor a supplied component partition.

Proof. The procedure starts with an ordinary build and the lists $J_i = \{j_\sigma(i)\}$ for $i \notin B$. For $J_i \subseteq K_B(i)$, uniqueness of the fundamental circuit gives

$$J_i + i \notin \mathcal{I} \quad \Longleftrightarrow \quad J_i = K_B(i).$$

Thus one query tests whether the partner list is complete. If $J_i + i$ is independent and $|J_i| = s-1$, the full circuit has more than s elements and the procedure returns failure.

Otherwise, if the last discovered partner is b_h, all undiscovered partners precede it. For $0 \le k \le h-1$,

$$(J_i \cup P_k) + i \notin \mathcal{I} \quad \Longleftrightarrow \quad K_B(i) \setminus J_i \subseteq P_k.$$

The endpoint answers are known, so binary search finds the last undiscovered partner in at most $\lceil \log_2 r \rceil$ calls. The procedure adds this partner to J_i and repeats the completeness test. Each outside arm requires at most $s-1$ completeness tests and $s-2$ additional searches. A disjoint-set structure joins each discovered pair (i,j); a joined set exceeding s also certifies failure, since its elements lie in one true component.

If all lists are complete, let C range over the resulting graph components and put $B_C = B \cap C$. Every outside $i \in C$ lies in $\mathrm{cl}(B_C)$, so

$$C \subseteq \mathrm{cl}(B_C), \qquad r(C) = |B_C| =: r_C, \qquad \sum_C r_C = r.$$

For restriction bases A_C, put $A = \bigcup_C A_C$. Since $C \subseteq \mathrm{cl}(A_C) \subseteq \mathrm{cl}(A)$ for every C, we have $r(A) = r(E) = r$ and $|A| = \sum_C r_C = r$. Thus A is a basis. Conversely, a basis meets each C in at most r_C elements, and equality holds because its total size is r. Therefore

$$\mathcal{B}(\mathcal{M}) = \left\{ \bigcup_C A_C : A_C \in \mathcal{B}(\mathcal{M}|C) \right\}.$$

Graph edges lie within true components, and the displayed direct-sum decomposition prevents a true component from crossing graph components. The partitions coincide.

The procedure returns failure if a recovered component exceeds s. Otherwise, it queries every r_C-subset of C and stores the independent ones. These basis tables determine all restrictions. Since $2^n/n$ is nondecreasing for integer $n \ge 1$,

$$\sum_C \binom{|C|}{r_C} \le \sum_C 2^{|C|} \le \frac{2^s}{s} \sum_C |C| = \frac{d2^s}{s}.$$

Adding the build and partner-search costs proves Equation 57; early failure only reduces the cost. If every component has size at most s, neither failure condition occurs. A larger component is detected either during the searches or in the completed graph. For fixed s, the labels, partner lists, and local basis tables occupy $O_s(d)$ words; a temporary query uses $O(d)$ words. □

Corollary 1 (Query complexity for bounded components). *For fixed $s \ge 2$, unknown matroids with component sizes at most s admit $O_s(d \log d)$ total external queries and $O_s(d)$ words while attaining $C^\star$ at every mean vector. At $d = 2r$, the worst-case query order is $\Theta(d \log d)$, even allowing arbitrary ties and without knowledge of L.*

Proof. Proposition 6 supplies local basis tables. They answer every subsequent ECR-KL request exactly, so Theorem 1 applies with no further external queries. Without the size promise, failure can instead be followed by ordinary ECR-KL at the additional bounded recognition cost.

For the lower bound, restrict the family in Appendix E.1 to $d = 2r$ and bijections from its r outside elements to its r anchors. There are $r!$ structures, each with components of size two. The sampling limits in Lemma E.1 identify the bijection. With a uniform prior Π on these permutations, the answer-code lemma and the decoding bound give

$$\frac{1}{r!}\sum_{\pi}\mathbb{E}_{\pi}Q_{\infty} \geq H(\Pi) = \log_2(r!) = \Theta(d\log d).$$

All one-coordinate reward alternatives keep the same components. Distinct-mean instances are included in the arbitrary-ties model, so this is also a lower bound for that model. □

Here s is fixed. Taking $s = d$ restores exponential enumeration and does not settle the arbitrary-ties query order for unrestricted matroids.

H Experimental details

H.1 Setup

All performance studies use independent Bernoulli rewards. Learners receive only d, r, oracle answers, and selected rewards; the means and full matroid are used only for evaluation. Within each instance and seed, policies share coordinate reward streams indexed by observation count. Every evaluation run is retained, including those with unfinished repairs.

Unless stated otherwise, intervals are unadjusted 95% Student's t-intervals. For a statistic $z_{f,s}$ on F fixed instances and m seeds, suite summaries use

$$\bar{z}_s = \frac{1}{F}\sum_{f=1}^{F} z_{f,s}, \qquad \bar{z} = \frac{1}{m}\sum_{s=1}^{m}\bar{z}_s, \qquad \bar{z} \pm t_{.975,m-1}\sqrt{\frac{\sum_{s=1}^{m}(\bar{z}_s-\bar{z})^2}{m(m-1)}}.$$

For paired contrasts, $z_{f,s}$ is the within-instance, within-seed policy difference. These intervals describe reward variability on the specified instances, not variation across a population of matroids. Certificate correctness and query plateaus are checkpoint diagnostics, not evidence of permanent stabilization. The supplement contains the instance specifications and random seeds.

Small instances. The serial-policy studies in Appendices H.3 and H.4 use the single-exchange rule of Appendix G.1. Their four illustrative instances have $d = 8$, $r = 4$, unique optimal bases, and zero-based labels:

Instance	Mean vector
Uniform, distinct	$(.35, .85, .55, .75, .30, .65, .25, .45)$
Partition, distinct	$(.30, .80, .35, .75, .25, .70, .45, .65)$
Graphic, distinct	$(.35, .80, .75, .30, .70, .25, .65, .40)$
Uniform, two levels	$(.52, .50, .50, .52, .50, .52, .52, .50)$

The uniform instances use $U_{4,8}$. Partition groups are $\{0,1\}$, $\{2,3\}$, $\{4,5\}$, $\{6,7\}$. Graphic edges, in coordinate order, are $(0,1)$, $(1,2)$, $(2,3)$, $(3,4)$, $(0,4)$, $(0,2)$, $(1,3)$, $(2,4)$. These are illustrative instances, not a sample from an instance distribution.

Implementation checks. Structural checks enumerate 216 labeled loopless matroids with $2 \leq d \leq 5$, $1 \leq r < d$, including weak mean orders and nonunique optima. They compare insertion thresholds, exchange charging, group transversals, additive losses, build reuse, greedy verification, and component recognition against independent definitions. Numerical checks use the experimental families, fixed demand schedules, deterministic-reward repair cases, reuse replays, and a 60-step bisection reference for the KL solver, separately from performance samples.

Every performance record is checked for $\sum_i N_i(T) = rT$, $0 \leq S_i(T) \leq N_i(T)$, and

$$R_{\mathrm{path}}(T) = T\mu^{\star} - \sum_i N_i(T)\mu_i.$$

Label tie-breaking is deterministic and learner code receives no evaluator information. The implementation, not the theory, is limited to $d \leq 32$.

The checks verify 295244 insertion identities, 585932 charging inequalities, 11085882 additive-loss identities, and 90254 component-recognition cases. All 10000 demand schedules meet their counts; deterministic tests complete 500 targeted candidates and 500 false-tie fallbacks. Supplied triggers do not estimate false-trigger probabilities.

Independent oracles agree on 453216 subset queries. Replays of 66 trajectories and 5520000 rounds match counts, sums, and action/reward hashes. All 192 structural-cache greedy runs match ordinary greedy. Replaying all 1152 cache-controlled trajectories matches every saved output, with regret-identity discrepancy below $5 \cdot 10^{-9}$. Replays add no independent samples and do not verify stochastic convergence proofs.

H.2 Parallel exploration and refreshes

Serial and parallel actions. Suite A uses the eight partition/graphic instances of Appendix H.3 with fresh reward seeds. Suite B has four instances per family: binary linear, direct sums of $U_{2,4}$, and direct sums of a rank-four sparse-paving matroid on eight elements. Each family uses $d \in \{16, 32\}$, $r = d/2$, and nominal minimum exchange gaps $.02, .05$. Suite C has four exact-tie, nonunique-optimum instances: two uniform, one capacity-one partition, and one graphic.

Binary-linear instances contain the r standard basis vectors of $\mathbb{F}_2^r$ and r independent uniform draws from the nonzero vectors; the drawn vectors need not be linearly independent. The sparse-paving restriction excludes four-element sets with zero-based binary masks $15, 51, 195, 60, 204$; their pairwise intersections have size at most two. In Suite B, the label-order greedy basis receives means uniform in $[.57, .83]$. Each outside mean equals its minimum feasible-removal mean minus δu_i, with u_i uniform in $[1, 2]$ and the first multiplier fixed at one. Gap pairs share topology, basis means, multipliers, and label permutation.

The six policies in Table 5 use exactly the exploration function $f(t)$ in Equation 3, eight seeds per instance, and $T = 10^6$. Checkpoints are 2048, 10^4, $5 \cdot 10^4$, 10^5, $2 \cdot 10^5$, $5 \cdot 10^5$, and 10^6. Appendix G.2 specifies the structural-cache greedy rule.

The four ECR-KL variants attempt capacity-one recognition after the initial build, costing at most $d - r$ extra calls. Success gives a group cover followed by scalar KL selection in each exact group. Failure leads to an $h_0 = 2048$ prefix, eight bootstrap builds, $b = d$ direct credits, the global test, and targeted refinement after the credits are exhausted. Builds occur at $\lfloor kh_0/8 \rfloor$, $k = 0, \ldots, 7$, and at h_0. Variants use their own action and cover rules without prefix tests, so their states at h_0 can differ.

The prefix was selected in Appendix H.3; 30000-round development variants on one reused graphic instance are excluded here. Suite A is not a structural holdout. Suite B families, seeds, and policies were fixed before outcomes were observed; the structural-cache baseline was specified during the run but before aggregate results were inspected.

All policies use the same 16-entry inference cache. An independent superset implies independence, a dependent subset implies dependence, and looplessness and rank supply further answers. Hits move entries to the front. Misses insert the answer, remove subsumed same-answer entries, and evict least-recently-used entries. Storage is $O(d)$ words. Raw requests and external calls include prefixes, failed recognition, and rebuilds; no action depends on the policy's query count.

For the separate post-hoc comparison, common cap-eight component recognition costs q_f on instance f and gives

$$Q^{\mathrm{rec}}(f) = q_f + \mathbf{1}\{\text{recognition fails on } f\}\, Q_{16}(f).$$

Recognition calls are uncached. After success, later requests are answered internally; after failure, recognition state is discarded and the original cache starts empty. The cap was chosen after the main results. This accounting preserves reward trajectories and adds no performance samples. The study has 1152 trajectories, including 192 structural-cache greedy runs. Suite intervals use the eight seed-level means defined in Appendix H.1.

Parallel exploration reduces regret relative to serial by 11.0% and 14.2% on Suites A and B, but remains 13.8% and 19.0% above matched KL-greedy (Table 5, Figure 3). Paired parallel-minus-serial differences are -376.2 ± 348.4 and -629.8 ± 259.4. Adding packing/reuse/quotas to parallel

gives -29.1 ± 107.0 and 0.5 ± 151.4; adding packing/reuse to serial gives -277.5 ± 246.4 and 150.2 ± 273.0. These additions do not consistently reduce regret.

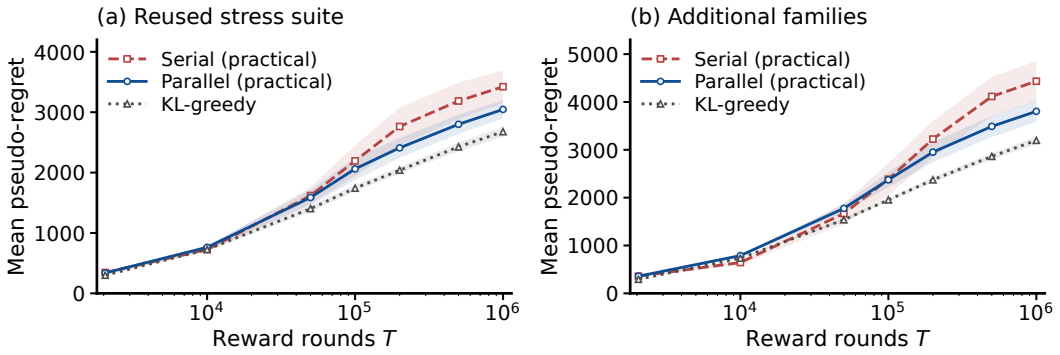


Figure 3: Serial/parallel ablation on Suite A (left: partition and graphic) and Suite B (right: binary linear, capacity two, and sparse paving). Shading shows unadjusted 95% t-intervals over eight seed-level instance averages. All reward costs are included.

Table 5: Matched-schedule results at $T = 10^6$. Intervals use eight seed-level means over eight, twelve, or four instances. Q_{raw} counts all requests; Q_{16} counts external calls through the 16-entry cache. All ECR-KL runs are outside repair at the horizon.

Policy	Regret	Q_{raw}	Q_{16}
Suite A: partition and graphic			
Serial (practical)	$3\,423 \pm 273$	338	254 ± 6
Parallel (practical)	$3\,046 \pm 150$	339	248 ± 4
Serial + packing/reuse	$3\,145 \pm 289$	322	249 ± 7
Parallel + packing/reuse/quota	$3\,017 \pm 84$	320	245 ± 7
KL-greedy	$2\,677 \pm 83$	21 257 942	$118\,221 \pm 6\,503$
KL-greedy + structural cache	$2\,677 \pm 83$	5 547	$4\,243 \pm 207$
Suite B: binary linear, capacity two, and sparse paving			
Serial (practical)	$4\,434 \pm 421$	606	427 ± 5
Parallel (practical)	$3\,804 \pm 223$	608	420 ± 4
Serial + packing/reuse	$4\,584 \pm 456$	560	418 ± 8
Parallel + packing/reuse/quota	$3\,804 \pm 246$	561	406 ± 5
KL-greedy	$3\,197 \pm 79$	19 493 205	$46\,677 \pm 1\,102$
KL-greedy + structural cache	$3\,197 \pm 79$	19 717	$14\,705 \pm 343$
Suite C: exact ties			
Serial (practical)	220 ± 13	112	31 ± 2
Parallel (practical)	220 ± 13	111	31 ± 2
Serial + packing/reuse	220 ± 13	81	31 ± 2
Parallel + packing/reuse/quota	220 ± 13	81	31 ± 2
KL-greedy	221 ± 14	7 249 896	$2\,766 \pm 204$
KL-greedy + structural cache	221 ± 14	2 545	373 ± 19

In Suite A, every ECR-KL variant recognizes the structure in $32/64$ runs; parallel and the full variant have correct certificates in $25/32$ and $26/32$ remaining runs. In Suite B, capacity-one recognition always fails: the corresponding counts are $57/96$ and $61/96$, although all stored bases are optimal. Suite C has eight recognized runs and 24 correct remaining certificates. Recognized runs do not maintain mean certificates. No run enters ordinary repair or uses more than one direct credit. These counts concern the sufficient mean certificate, not a direct test of the asymptotic sampling rates.

Common recognition eliminates later calls on capacity-two and sparse-paving instances (Table 6). At $d = 16, 32$, recognition costs are $(89, 217)$ calls for capacity-two and $(247, 553)$ for sparse-paving instances. Recognition fails on all tested binary-linear and larger graphic instances. Without common

recognition, binary-linear greedy makes 18661 cached calls with structural caching versus 10878 without it. The structural cache adds $O(d)$ words but need not reduce calls.

Table 6: Common cap-eight recognition, expanding Table 2. Intervals use eight seed-level means over four instances per family. Entries are mean regret and external calls. Both greedy variants have the same regret. Recognition identifies $4/4$ instances in each partition and sparse-paving family, and $0/4$ in the other families. All recognition calls are charged; reward trajectories are unchanged. Per-instance values are in the supplement.

Family	R_{par}	R_{gr}	Q_{par}	Q_{gr}	$Q_{\mathrm{gr+str}}$
Capacity one	$2\,460 \pm 91$	$2\,460 \pm 91$	100 ± 0	100 ± 0	100 ± 0
Graphic	$3\,633 \pm 294$	$2\,895 \pm 110$	514 ± 7	$114\,101 \pm 3\,560$	$8\,504 \pm 415$
Binary linear	$4\,029 \pm 373$	$3\,314 \pm 155$	465 ± 7	$10\,957 \pm 472$	$18\,740 \pm 776$
Capacity two	$3\,310 \pm 372$	$2\,907 \pm 79$	153 ± 0	153 ± 0	153 ± 0
Sparse paving	$4\,073 \pm 399$	$3\,371 \pm 156$	400 ± 0	400 ± 0	400 ± 0

Scheduled rebuilds. The 16 connected graphic/binary-linear instances use $d \in \{16, 32\}$, $r = d/2$, two topologies per family and dimension, and exchange gaps $.01, .05$. Graphic instances are cycles on $r + 1$ vertices plus $r - 1$ distinct random chords. Binary-linear instances contain the standard basis, the all-ones column, and $r - 1$ independent uniform nonzero columns. Both have a single component. Means and gap pairing use the Suite B construction in Appendix H.2.

We compare practical parallel ECR-KL, its three-refresh variant, and KL-greedy with and without structural caching. Refreshes after 8192, 32768, 131072 rounds sort global empirical means, reset to one block, and rebuild before the next action. Global counts/sums are retained and local statistics reset. Refreshes consume no direct credit, are skipped after successful capacity-one recognition, and retain observations but discard any interrupted batch. Appendix G.2 covers this finite schedule.

Through the final refresh at $h_* = 131072$, the unrecognized branch has at most nine bootstrap builds including the build at h_0, three scheduled refreshes, d direct rebuilds, and h_* ordinary builds. Including the capacity-one test, a deterministic upper bound is

$$Q_{\mathrm{prefix}} \le (h_* + d + 12)c_{d,r} + d - r.$$

Proposition 4 applies after the final refresh with at most d remaining credits. The scheduled and unscheduled runs can have different later triggers, so their query difference need not equal three build costs.

All other settings match Appendix H.2: $h_0 = 2048$, eight bootstrap builds, $b = d$, the global test, and $f(t)$ from Equation 3, with packing/reuse/quotas disabled. Eight shared-stream seeds and $T = 10^6$ give 512 runs; checkpoints are 2048, 8192, 32768, 131072, 200000, 500000, 10^6.

The schedule was chosen after inspecting earlier results. Generator, seeds, horizon, policies, and analysis were then fixed before refresh-study rewards were observed. There was no external preregistration, outcome-driven schedule search, or run exclusion, extension, or replacement.

Every policy uses the 16-entry cache and common cap-eight recognition, which fails on all instances. Its cost is added without carrying state forward; reward trajectories are unchanged. The primary contrast is refreshed-minus-original parallel regret at 10^6, with paired intervals over eight seed-level means across 16 instances. Percentages are ratios of pooled means; family/instance analyses are secondary.

Table 7 reports endpoints; Figure 2 shows all seven checkpoints for both regret and queries. Refreshes reduce regret from 6509 to 4973 (23.6%), with paired difference -1536 ± 276 for 65 ± 3 extra calls. Regret remains 149 ± 85 (3.1%) above greedy. Without common recognition, original/refreshed query means are 410 ± 3, 475 ± 4; with it, 494 ± 3, 558 ± 4. Additional raw requests average 120 ± 4.

Paired reductions are 1560 ± 509 on graphic and 1512 ± 597 on binary-linear instances, leaving regret 2.2% and 4.0% above greedy. All 16 instance means favor refreshes, but some intervals include zero (Table 8). Greedy implementations share identical trajectories.

Table 7: Refresh study at $T = 10^6$: 128 runs per policy; intervals use eight seed-level means. Q includes common recognition and caching. *Cert.* counts correct active certificates; *Partners* counts optimal bases with minimum-mean feasible partners. Both are evaluator diagnostics.

Policy	Regret	Q	Cert.	Partners
Parallel (practical)	$6\,509 \pm 213$	494 ± 3	93/128	107/128
Parallel + three refreshes	$4\,973 \pm 145$	558 ± 4	95/128	111/128
KL-greedy	$4\,824 \pm 88$	$39\,224 \pm 1\,680$	—	—
KL-greedy + structural cache	$4\,824 \pm 88$	$19\,725 \pm 963$	—	—

Table 8: All refresh-study instances, with eight paired seeds per row. ΔR is refresh minus original parallel regret, with intervals from paired differences. Topology indices are local to each family/dimension.

Family	d	Top.	Gap	Parallel	Refresh	ΔR
Binary	16	0	0.01	$7\,414 \pm 1\,714$	$5\,362 \pm 1\,213$	$-2\,052 \pm 1\,852$
Binary	16	0	0.05	$2\,014 \pm 417$	$1\,699 \pm 195$	-315 ± 358
Binary	16	1	0.01	$6\,442 \pm 1\,234$	$4\,727 \pm 285$	$-1\,715 \pm 1\,378$
Binary	16	1	0.05	$1\,778 \pm 353$	$1\,576 \pm 57$	-202 ± 372
Binary	32	0	0.01	$12\,913 \pm 2\,730$	$9\,561 \pm 500$	$-3\,352 \pm 2\,515$
Binary	32	0	0.05	$3\,512 \pm 445$	$3\,237 \pm 106$	-275 ± 404
Binary	32	1	0.01	$14\,683 \pm 1\,833$	$10\,502 \pm 963$	$-4\,181 \pm 2\,098$
Binary	32	1	0.05	$3\,156 \pm 137$	$3\,153 \pm 137$	-3 ± 32
Graphic	16	0	0.01	$8\,449 \pm 1\,886$	$4\,964 \pm 622$	$-3\,485 \pm 2\,045$
Graphic	16	0	0.05	$1\,739 \pm 322$	$1\,609 \pm 132$	-130 ± 275
Graphic	16	1	0.01	$8\,231 \pm 2\,420$	$5\,145 \pm 410$	$-3\,086 \pm 2\,275$
Graphic	16	1	0.05	$1\,712 \pm 353$	$1\,593 \pm 190$	-119 ± 282
Graphic	32	0	0.01	$12\,829 \pm 2\,067$	$9\,905 \pm 377$	$-2\,925 \pm 2\,024$
Graphic	32	0	0.05	$3\,152 \pm 184$	$3\,138 \pm 192$	-15 ± 18
Graphic	32	1	0.01	$12\,617 \pm 2\,015$	$10\,232 \pm 1\,489$	$-2\,385 \pm 1\,886$
Graphic	32	1	0.05	$3\,509 \pm 390$	$3\,171 \pm 176$	-338 ± 334

All 128 refresh runs perform three rebuilds; neither parallel policy enters repair, so no batch is cancelled. Optimal bases increase from $123/128$ to $125/128$, correct certificates from $93/128$ to $95/128$, and correct insertion partners from $107/128$ to $111/128$. At $d = 32$, however, the correct-certificate count falls from $37/64$ to $34/64$. The regret improvement is therefore not accompanied by uniformly better certificate recovery.

For original parallel, refreshed parallel, and greedy, mean normalized suboptimal-arm counts $N_i(T)I_i/\log T$ are $1.928 \pm .061$, $1.602 \pm .049$, $1.584 \pm .034$; late-window ratios $(N_i(T) - N_i(T/2))I_i/\log 2$ are $4.644 \pm .750$, $4.214 \pm .406$, $4.025 \pm .312$. Arms are averaged within instances, then instances within seeds. These finite-horizon ratios do not establish asymptotic allocation. In particular, a limit after normalization by $\log T$ does not determine increments over $(T/2, T]$. The schedule has not been separately tested on exact ties or larger dimensions.

All 3584 checkpoints satisfy feasibility and count identities, with maximum regret-identity discrepancy $4.13 \cdot 10^{-9}$. The 128 greedy/cache pairs match throughout; original/refresh pairs match through round 8192. Independent oracle and KL-solver checks agree and are separate from the 512 performance trajectories.

H.3 Initialization and repair

Prefix selection. Development and held-out mean vectors use the three structures above, each with a wide and narrow vector. Eight evenly spaced values in $[.15, .85]$ or $[.30, .70]$ receive small independent perturbations and label permutations. The six held-out vectors have distinct coordinates and basis values and were fixed before policy comparisons (Table 9).

Table 9: Held-out mean vectors, rounded to four decimals.

Instance	Means in element order
Uniform, wide	$(0.3632, 0.4643, 0.7370, 0.8444, 0.5584, 0.6631, 0.2555, 0.1577)$
Uniform, narrow	$(0.4803, 0.6534, 0.3476, 0.6873, 0.4233, 0.2962, 0.5886, 0.5305)$
Partition, wide	$(0.7363, 0.6645, 0.1574, 0.2375, 0.3362, 0.4359, 0.5442, 0.8504)$
Partition, narrow	$(0.5417, 0.4618, 0.2951, 0.4182, 0.6426, 0.7063, 0.3493, 0.6005)$
Graphic, wide	$(0.8593, 0.7382, 0.2522, 0.4377, 0.6462, 0.3643, 0.5469, 0.1517)$
Graphic, narrow	$(0.4608, 0.6980, 0.5234, 0.6360, 0.5746, 0.2861, 0.4120, 0.3477)$

Development uses six instances, eight seeds, and $T = 2 \times 10^5$. The grid compares KL-greedy prefixes of length $h_0 \in \{128, 512, 2048\}$ and bootstrap prefixes at the same lengths with $K_{\mathrm{boot}} \in \{2, 4, 8\}$ builds. We select $h_0 = 2048, K_{\mathrm{boot}} = 8$ by minimizing mean normalized regret, breaking ties by fewer queries and then shorter duration.

The bootstrap prefix builds at $t_k = \lfloor k h_0 / K_{\mathrm{boot}} \rfloor$, $k = 0, \ldots, K_{\mathrm{boot}} - 1$, each followed by a cover and alternating basis/optimistic-exchange plays without tests. At h_0, a new empirical-order build retains (N, S) and resets local statistics. Bootstrap and KL-greedy prefix costs are bounded by $K_{\mathrm{boot}} c_{d,r}$ and $d h_0$, respectively, and are charged. Proposition 3 gives

$$\mathbb{E} Q_\infty \le K_{\mathrm{boot}} c_{d,r} + \frac{4dL}{d + L + 1} c_{d,r}$$

for the bootstrap continuation. Both h_0 and K_{boot} are fixed independently of the evaluation horizon.

Cold, warm, and bootstrap starts are crossed with uniform and targeted refinement, using Serial ECR-KL and $\alpha = \eta = 1/(8d)$. Baselines are KL-greedy with Equation 3, KL-OSM (Talebi & Proutière, 2016, Algorithm 2), and MAUB (Delage & Gaudel, 2025, Algorithms 2–3). KL-OSM uses $\log t + 3 \log \log t$ for $t \ge 3$, zero thresholds in the first two rounds, and index one for unobserved arms; MAUB uses $\bar{\mu}_i + \sqrt{2 \log \ell_B / N_i}$. All ties are broken by arm label.

Evaluation uses six held-out instances, 32 shared-stream seeds, and $T = 10^6$. Queries are uncached, including initialization and repeated requests; all reward costs and unfinished repairs are counted. Intervals use 32 seed-level suite averages.

Repairs dominate cold-start regret (Tables 10–13). Neither initialized policy triggers a held-out repair, so uniform and targeted refinement coincide in these runs. Bootstrap's conservative KL schedule also contributes to its regret relative to KL-OSM; matched KL-greedy separates this effect from structural access. Figure 4 reports the trajectories.

Table 10: Complete initialization–refinement ablation for Serial ECR-KL on the held-out set. All prefix costs are charged, and runs with unfinished repairs are retained.

Policy	Mean $R/(C^\star \log T)$	Mean Q_T	Active repair
Cold ECR-KL	$2\,194.9 \pm 18.5$	23.4	39/192
Cold + targeted	918.5 ± 20.5	25.7	32/192
Warm + uniform	2.02 ± 0.05	10 982	0/192
Warm + targeted	2.02 ± 0.05	10 982	0/192
Bootstrap + uniform	2.20 ± 0.05	115.4	0/192
Bootstrap + targeted	2.20 ± 0.05	115.4	0/192
KL-OSM	1.45 ± 0.04	5.33×10^6	–
MAUB	3.84 ± 0.05	374.8	–
KL-greedy	2.02 ± 0.05	5.33×10^6	–

Table 14 separately reports the three original instances, reusing their original 32 streams and excluding them from held-out averages. Their equal-valued suboptimal bases violate MAUB's distinct-basis-values query assumption. Bootstrap's partition regret is 4.94 versus MAUB's 4.05, and one graphic run remains in repair at 10^6: the absence of held-out post-prefix repairs does not generalize to every instance.

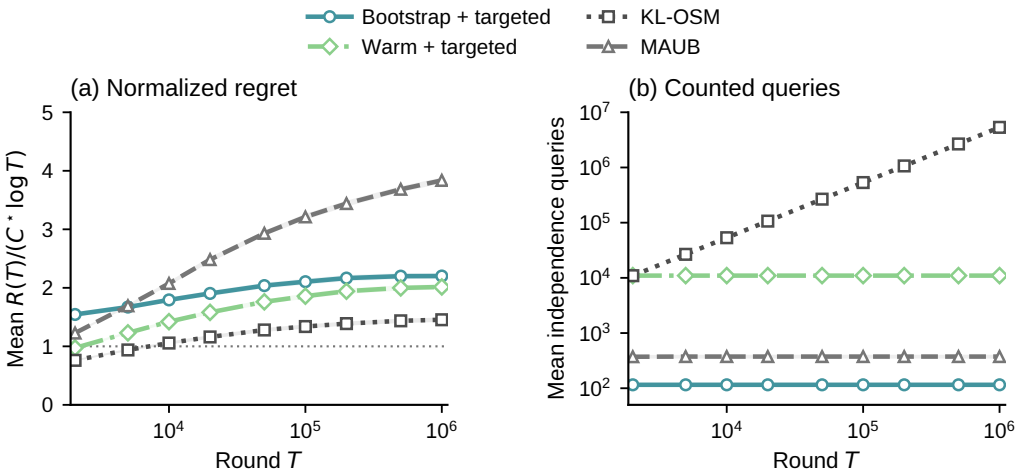


Figure 4: Finite-prefix policies on six held-out instances; oracle requests are uncached. (a) Normalized regret; the dotted line is the asymptotic target, not a finite-time prediction. (b) Counted independence queries, on a logarithmic scale. For each of 32 independent seeds, we first average over the six instances; shading shows unadjusted 95% Student's t-intervals. Every point includes the prefix cost.

Table 11: Paired differences at $T = 10^6$ over 32 seed-level averages of six instances. Negative values favor the first policy; intervals use paired differences.

Contrast	$\Delta[R/(C^\star \log T)]$	ΔQ_T
Cold + targeted − Cold ECR-KL	$-1\,276.4 \pm 21.8$	2.29 ± 0.30
Warm + targeted − Warm + uniform	0.00 ± 0.00	0.00 ± 0.00
Bootstrap + targeted − Bootstrap + uniform	0.00 ± 0.00	0.00 ± 0.00
Bootstrap + targeted − KL-OSM	0.75 ± 0.02	$-5.33 \times 10^6 \pm 27.4$
Bootstrap + targeted − MAUB	-1.63 ± 0.04	-259.4 ± 29.9
Bootstrap + targeted − KL-greedy	0.19 ± 0.02	$-5.33 \times 10^6 \pm 27.5$
Warm + targeted − KL-OSM	0.56 ± 0.02	$-5.32 \times 10^6 \pm 8.96$
Warm + targeted − MAUB	-1.82 ± 0.04	$10\,607 \pm 33.5$

Table 12: Held-out instances over 32 seeds. Bootstrap and warm starts use targeted refinement; both have correct certificates ($32/32$ per instance) and no active repair ($0/32$). The two panels report normalized regret and uncached queries.

Instance	Bootstrap	Warm	KL-OSM	MAUB
Normalized regret $R/(C^\star \log T)$				
Uniform, wide	2.39 ± 0.08	2.11 ± 0.08	1.51 ± 0.08	3.87 ± 0.13
Uniform, narrow	2.08 ± 0.10	2.00 ± 0.10	1.43 ± 0.09	3.55 ± 0.12
Partition, wide	2.22 ± 0.13	2.04 ± 0.13	1.45 ± 0.12	4.45 ± 0.13
Partition, narrow	2.04 ± 0.12	2.03 ± 0.12	1.52 ± 0.12	3.78 ± 0.14
Graphic, wide	2.46 ± 0.15	1.96 ± 0.11	1.40 ± 0.10	3.71 ± 0.17
Graphic, narrow	2.01 ± 0.10	1.96 ± 0.10	1.41 ± 0.08	3.65 ± 0.12
Uncached queries Q_T				
Uniform, wide	108.0	8 204.0	4.00×10^6	185.2
Uniform, narrow	108.0	8 204.0	4.00×10^6	379.5
Partition, wide	134.3	14 339	7.00×10^6	427.6
Partition, narrow	125.7	14 318	7.00×10^6	694.9
Graphic, wide	108.4	12 199	6.00×10^6	216.6
Graphic, narrow	108.1	8 629.7	4.00×10^6	344.8

Table 13: Empirical quantiles over 192 runs per policy, pooled across six fixed instances. Query and normalized-regret quantiles are descriptive, not confidence bounds.

Policy	$Q_{.10}$	$Q_{.50}$	$Q_{.90}$	$\rho_{R,.50}$	$\rho_{R,.90}$	$\rho_{R,.99}$
Cold ECR-KL	12.0	24.0	30.0	1 204.4	4 772.7	5 058.5
Cold + targeted	24.0	24.0	29.0	821.7	1 795.1	2 052.2
Warm + targeted	8 204.0	11 129	14 343	1.99	2.42	2.75
Bootstrap + targeted	108.0	108.0	134.0	2.16	2.67	3.14
KL-OSM	4.00×10^6	5.00×10^6	7.00×10^6	1.45	1.79	2.12
MAUB	140.0	295.0	705.6	3.80	4.49	4.95

Table 14: Exploratory results on the original instances and streams, excluded from held-out averages. Both initialized policies use targeted refinement: each has one unfinished graphic repair ($1/32$), and none on uniform or partition ($0/32$). MAUB's distinct-basis-values query guarantee is not invoked.

Instance	Bootstrap	Warm	KL-OSM	MAUB
Normalized regret $R/(C^\star \log T)$				
Uniform	2.37 ± 0.09	2.12 ± 0.09	1.53 ± 0.09	3.77 ± 0.11
Partition	4.94 ± 0.09	2.09 ± 0.10	1.50 ± 0.10	4.05 ± 0.15
Graphic	3.77 ± 3.38	3.61 ± 3.38	1.44 ± 0.08	3.81 ± 0.11
Uncached queries Q_T				
Uniform	108.0	8 204.0	4.00×10^6	157.6
Partition	111.0	14 113	7.00×10^6	417.6
Graphic	115.5	10 249	5.00×10^6	388.8

Larger instances. Eight partition/graphic instances use $d \in \{16, 32\}$, $r = d/2$, and nominal minimum exchange gaps .02, .05. Partition block sizes are $(1, 3, 2, 2)$ repeated $r/4$ times. Graphic instances start from a random tree on $r+1$ vertices and add r non-tree edges. Designated basis means are uniform in $[.55, .85]$; outside means equal the minimum feasible-removal mean minus δu_i, with u_i uniform in $[1, 3]$ and the first multiplier one. Gap pairs share topology, basis means, multipliers, and label permutation. The encoding

$$\mu_i = 2^{-52}\big(2^{34}q_i + 2^i\big), \qquad q_i = \lfloor 2^{18}\widetilde{\mu}_i \rfloor, \qquad i = 0, \ldots, d-1$$

ensures distinct basis values through distinct subset residues modulo 2^{34}. It changes means by less than $4 \cdot 10^{-6}$ and exchange gaps by less than $5 \cdot 10^{-6}$ relative to nominal values.

Bootstrap with targeted repair keeps $h_0 = 2048$, $K_{\mathrm{boot}} = 8$; it, KL-OSM, and MAUB use 16 shared-stream seeds and $T = 10^6$ (384 runs). Queries are uncached; the selected settings and evaluation horizon were not changed. Intervals use 16 seed-level suite averages and retain unfinished repairs. Optimal bases and $C^\star$ are independently checked by group maxima or maximum-weight spanning trees, with exhaustive checks for partition and $d = 16$ graphic instances.

Bootstrap has 8.02 times KL-OSM's mean regret and 3.09 times MAUB's, using about 557.8 queries—factors $38\,108$ and 152 fewer (Table 15). All 25 triggered repairs remain active at T. Of 128 bootstrap runs, 86 store an optimal basis but only seven have a correct active certificate (Tables 16 and 17). Because dimension, topology, and gap all change in this study, it does not isolate the effect of dimension.

Table 15: Fixed-prefix results at $T = 10^6$: eight instances and 16 seed-level means. All runs, including unfinished repairs, contribute.

Policy	Unnormalized regret	Independence queries	Active repair
Bootstrap + targeted	$15\,816 \pm 4\,658$	558 ± 2	25/128
KL-OSM	$1\,973 \pm 44$	$(21.257 \pm 0.002)\,10^6$	–
MAUB	$5\,119 \pm 63$	$84\,694 \pm 5\,037$	–

Table 16: Unnormalized regret over 16 seeds per stress instance. *Repair* counts bootstrap runs still in repair at $T = 10^6$; all runs are included.

Family / d / gap	Bootstrap	KL-OSM	MAUB	Repair
Partition / 16 / 0.02	$15\,072 \pm 9\,500$	$1\,765 \pm 102$	$4\,602 \pm 129$	2/16
Partition / 16 / 0.05	$21\,274 \pm 26\,381$	844 ± 59	$2\,234 \pm 74$	4/16
Partition / 32 / 0.02	$18\,864 \pm 8\,678$	$3\,185 \pm 106$	$8\,948 \pm 190$	3/16
Partition / 32 / 0.05	$19\,068 \pm 19\,297$	$1\,416 \pm 55$	$4\,131 \pm 109$	4/16
Graphic / 16 / 0.02	$10\,041 \pm 5\,995$	$2\,002 \pm 119$	$4\,895 \pm 240$	3/16
Graphic / 16 / 0.05	$4\,436 \pm 6\,546$	975 ± 60	$2\,460 \pm 78$	1/16
Graphic / 32 / 0.02	$22\,686 \pm 12\,235$	$3\,825 \pm 176$	$9\,199 \pm 264$	6/16
Graphic / 32 / 0.05	$15\,090 \pm 18\,300$	$1\,773 \pm 51$	$4\,484 \pm 84$	2/16

Unmodified ECR-KL. Six small instances cover both failure types, ties, and nonunique optima with zero or positive $C^\star$ (Table 18); partition groups are $\{1, 3\}, \{2, 4\}$. Algorithm 1 starts from zero observations and label order, without prefixes, recognition, direct credits, global tests, packing, or reuse. Seeds are 701–704, $T = 200000$, and checkpoints 1000, 10000, 50000, 100000, 200000 (24 runs).

Sixteen runs complete one repair each; eight start with correct certificates and never trigger. All finish with correct certificates and no active repair. The first instance has $C^\star = 0$ but substantial regret from correcting its initial basis. Across the runs, 2723 learning actions select at least two outside arms. These runs test the core routines; they are not a performance comparison.

Table 17: Bootstrap diagnostics over 16 runs per instance. Quantiles are empirical; *Optimal* counts optimal bases, and *Correct* counts true-mean certificates during learning or initial coverage.

Family / d / gap	$R_{.1}$	$R_{.5}$	$R_{.9}$	Optimal	Correct
Partition / 16 / 0.02	2 201	6 930	37 938	8/16	0/16
Partition / 16 / 0.05	962	1 186	63 482	13/16	1/16
Partition / 32 / 0.02	4 394	15 026	40 038	6/16	0/16
Partition / 32 / 0.05	1 809	1 994	87 552	14/16	0/16
Graphic / 16 / 0.02	2 180	2 665	29 693	10/16	2/16
Graphic / 16 / 0.05	1 219	1 381	1 520	15/16	4/16
Graphic / 32 / 0.02	4 850	13 766	53 316	6/16	0/16
Graphic / 32 / 0.05	2 432	2 573	49 719	14/16	0/16

Table 18: Core ECR-KL at $T = 200000$, four seeds per instance. Query counts are constant within each row. *Repaired* counts completed repairs; all 24 runs finish with correct certificates and no active repair. These are diagnostics, not a performance comparison.

Matroid	Means in arm order	Mean regret	Queries	Repaired
$U_{2,3}$	$(.2, .2, .8)$	3 040.6	6	4/4
$U_{2,4}$	$(.8, .5, .5, .1)$	27.5	4	0/4
$U_{2,4}$	$(.5, .5, .5, .5)$	0.0	4	0/4
$U_{2,3}$	$(.1, .8, .8)$	5 126.1	6	4/4
Partition	$(.1, .2, .9, .8)$	13 586.5	8	4/4
Partition	$(.2, .8, .1, .7)$	1 640.3	9	4/4

H.4 Full-Order ECR-KL

Ordering gaps. We use $U_{4,8}$ with means

$$\{.90, .80, .70, .60, .45, .30 + \delta/2, .30 - \delta/2, .15\}, \qquad \delta \in \{.08, .04, .02, .01\}.$$

Only two suboptimal means change; the optimal basis and minimum exchange gap 0.15 remain fixed. Cold Full-Order and the selected warm/bootstrap targeted ECR-KL policies use 16 paired seeds per gap and $T = 10^6$. First-ordering completion times are right-censored at T, without extensions or averages conditional on completion.

Full-Order's median completion time rises from 34744 at $\delta = .08$ to 669654 at $.02$; no run finishes at $.01$ (Table 19). Thus full ordering can be slow despite a fixed decision gap. These results give no uniform finite-time bound for either prefix policy.

Table 19: First Full-Order completion with decision gap 0.15. Unfinished times are right-censored at 10^6; means conditional on completion are not reported.

Ordering gap δ	Completed	Median completion	Active at 10^6
0.08	16/16	34 744	0/16
0.04	16/16	151 751	0/16
0.02	16/16	669 654	0/16
0.01	0/16	$> 1.00 \times 10^6$	16/16

Full-Order from a cold start. On the first three distinct-means instances of Appendix H.1, Full-Order starts from zero observations, priority $(0, \ldots, 7)$, and $\alpha = \eta = 1/8$, with the serial action rule and ordering procedure of Appendix F. Ordering observations remain in global counts.

Seeds 0–31 are shared with KL-greedy; eight pilot seeds are excluded. Queries are uncached. Checkpoints are 2×10^4, 5×10^4, 10^5, 2×10^5, 5×10^5, 10^6; no run activates the prespecified extension to 2×10^6. Settings are fixed before evaluation, and interrupted runs resume the same streams. Regret contrasts are paired; completion proportions use Wilson intervals (Table 21).

All 96 runs finish one ordering and perform two builds, with much higher regret than KL-greedy (Tables 20–21). Unnormalized regret and outside-arm counts are unchanged at saved checkpoints from 2×10^5 onward. The decreasing normalized regret therefore reflects division of transient costs by $\log T$, not evidence of asymptotic allocation.

Table 20: End-to-end results at $T = 10^6$, using 32 shared-stream seeds. Completion means a finished ordering and currently correct certificate. Query counts are constant within each instance.

	Full-Order ECR-KL			KL-greedy
Instance	Completed	Q_T	$R/(C^\star \log T)$	$R/(C^\star \log T)$
Uniform	32/32	24	447.7 ± 26.6	2.12 ± 0.09
Partition	32/32	27	$2\,163.0 \pm 81.7$	2.09 ± 0.10
Graphic	32/32	25	210.4 ± 6.6	1.94 ± 0.07

Table 21: Completion, last observed rebuild, and paired regret minus KL-greedy. Times do not certify permanent completion; completion intervals are Wilson 95%.

Instance	Completed (Wilson 95%)	Last rebuild, median	Paired regret difference
Uniform	32/32 (89.3–100%)	118198	$67\,908 \pm 4\,045$
Partition	32/32 (89.3–100%)	131112	$162\,416 \pm 6\,142$
Graphic	32/32 (89.3–100%)	129774	$60\,848 \pm 1\,930$

Query counts at balanced rank. Capacity-one partition matroids have $d \in \{8, 16, 24, 32\}$, $r = d/2$, anchor means $\nu_j = .55 + .025j$, and outside means $p_j = .45 - .025j$, $j = 0, \ldots, r-1$. Anchors have labels $0, \ldots, r-1$ and outside elements $r, \ldots, 2r-1$; a fixed permutation assigns one outside element to each anchor. Means lie in $(0, 1)$ and are separated by at least .025.

Serial Full-Order starts in label order with $\alpha = \eta = 1/8$ and 16 prespecified seeds per dimension. At 2×10^6, a prespecified rule extends every run at a dimension to 5×10^6 if fewer than 90% have finished ordering with a correct certificate, continuing the same streams.

All $d = 8$ runs finish by 2×10^6; all $d \geq 16$ runs are still ordering then. After extension, all 64 runs have correct certificates and two observed builds (Table 22). The nearly constant ratio $Q_T/(d \log_2 d)$ reflects the two observed builds on this family; the worst-case lower bound is proved separately in Theorem 2.

Table 22: Full-Order scaling over 16 seeds per dimension. The prespecified rule extends $d \geq 16$ to 5×10^6. Final query counts are constant within each dimension.

		Completed				
d	Final T	At 2×10^6	Final	Q_T	$Q_T/c_{d,r}$	$Q_T/(d \log_2 d)$
8	2×10^6	16/16	16/16	24	1.500	1.000
16	5×10^6	0/16	16/16	64	1.600	1.000
24	5×10^6	0/16	16/16	112	1.556	1.018
32	5×10^6	0/16	16/16	160	1.667	1.000