THE GEOMETRY OF UNLEARNING IN SEMI-BANDITS: CATEGORIES, REGRET, AND EXACT REPLAY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study exact unlearning in combinatorial semi-bandits, where a learner selects several items each round and observes their individual rewards. After exploration, it chooses a final action and stops collecting feedback. Removing an observation can change that stopping decision. How can learning remain stable, and how can an erased run be reconstructed from feedback already collected? We assign each protected action category a budget for changes in public event probabilities. For product actions with a unique optimum and comparable reward gaps, we prove matching regret bounds up to logarithmic factors. The cost separates finding the best action from keeping decisions stable. Category arrangement controls the second term: categories with equal entropy can leave different sampling opportunities. Our learner attains the bound without knowing the gaps. Its completion envelope uses the largest stopping probability over possible missing rewards. Shared random draws ensure that erased exploration ends no later than factual exploration, so retained feedback suffices for exact replay of this missing-data rule. Synthetic and preference-derived experiments measure regret, feedback collection, and reconstruction together.

1 INTRODUCTION

A recommendation service learns from responses to the items it displays. When one response is erased, how should it rebuild the learning run? Its record contains feedback for displayed items. Responses to other items remain unknown. This partial record shapes both the cost of learning and the possibilities for erasure.

We study this problem for a learner that explores, chooses a final set of items, and then stops collecting feedback. It keeps serving that set for the rest of the run. We call the final choice *commitment*. Erasure can change when the learner would commit. If a rerun continues collecting feedback past the original stopping point, it asks for observations absent from the record. A successful missing-feedback rule must make the remaining record sufficient.

Where should stability hold? An action can display several items together; we call this set a *slate*. A service may group slates by the topics or product categories served at each position. We give each such protected category a budget for how much erasure can change the probability of any public event. Public outcomes comprise the served slates and decisions to continue or commit. Categories are fixed before learning.

Separate budgets matter when categories have different frequencies. Consider a topic that appears in 1% of interactions. Changing all later behavior after those interactions contributes at most 1% to an average probability change, when the topic frequency and behavior elsewhere stay fixed. A category-specific budget gives that topic the same protection as a frequent one. Gathering evidence for every such promise has a cost.

The arrangement of categories matters. Figure 1 uses two display positions with four candidates each. One rule labels a slate by the offset between its numbered choices, wrapping after four. Another labels it by the first choice. Under independent uniform sampling, both give four equally likely categories with four slates each. They have equal label entropy, meaning equal uncertainty about the label, and leave equal uncertainty about the slate. Yet an offset keeps every candidate

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.

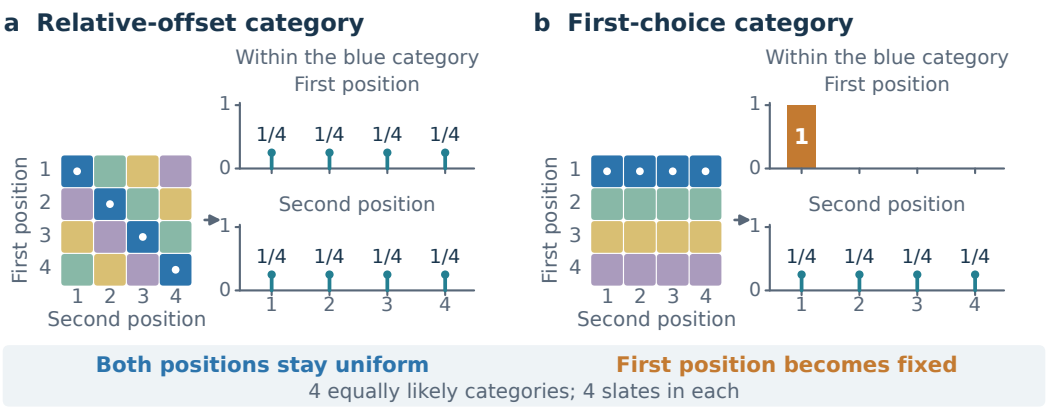

Figure 1: Equal entropy, different sampling opportunities. The same sixteen slates are grouped in two ways. Conditioning on one offset preserves each position's uniform distribution; conditioning on one first choice fixes that position. The resulting stability costs differ by the number of candidates per position, up to logarithmic factors (Corollary 1).

equally likely at either position, while a first-choice label fixes the first position. The erased response therefore has different influence on the evidence available within a category.

What does protection cost? We measure cost by regret, the expected reward lost through suboptimal choices. For slates that choose an item or a fixed-size bundle at each position, with one best slate and comparable reward differences, Theorem 1 separates the cost of identifying good choices from the cost of stable commitment. Matching bounds show how the latter grows with the number of groups per position. The geometry also guides sampling: balancing group exposure while preserving candidate coverage reduces the evidence needed for protection.

How can the same learner replay an erasure? We consider one requested interaction, chosen independently of rewards and learner randomness. Its feedback becomes a missing marker; its slot and served action remain. Our *completion envelope* defines the learner's response to that marker. It takes the largest probability of commitment over all possible missing rewards and tests it with the original random draw. Whenever erased learning continues, the original run also continues. The required observations are therefore already recorded. Theorem 2 combines this replay rule with group-aware sampling and doubling exploration phases to attain the cost law without supplied reward gaps.

2 Related Work

Exact unlearning and sequential feedback. Certified removal (Guo et al., 2020) and stability-based exact unlearning (Ullah et al., 2021; Ullah and Arora, 2023) connect forgetting to changes in output laws. Our closest sequential predecessor, Nguyen-Tang and Arora (2026), establishes exact unlearning with regret and computational guarantees for tabular reinforcement learning. Their Gaussian binary-tree prefix sums and maximal coupling use an environment oracle during reconstruction. Our product semi-bandit results characterize the cost of protected action categories and make retained observations sufficient for exact replay of the specified erased execution. Appendix A.6 compares the feedback and output contracts. Offline bandit unlearning studies final decisions from a fixed dataset (Ye et al., 2026).

Privacy and the cost of learning. Classical and private semi-bandit analyses quantify observation and privacy costs (Kveton et al., 2015; Chen et al., 2020; Azize and Basu, 2022). Pure differential privacy controls likelihood ratios and, for our pre-feedback categories, protects every category simultaneously. Total variation controls changes in event probabilities. Placing this latter budget within chosen categories yields the cost dependence we characterize. Section 4 instantiates the private bounds on the same product actions. Private stopping rules also connect privacy calibration to bandit data acquisition (Sajed and Sheffet, 2019).

Certificates, shared randomness, and requests. Our commitment rule uses evidence of a stable winner, following the broader idea of certificate-based release (Dwork and Lei, 2009). Its calibration depends on conditional sampling probabilities, connecting to privacy amplification by subsampling (Balle et al., 2018). Replicable bandits use shared randomness across independently sampled reward tables (Esfandiari et al., 2023); our coupling pairs a table with its erased version. Adaptive unlearning studies how requests are selected (Gupta et al., 2021; Chourasia and Shah, 2023). Appendix F gives our causal registration extension to the main fixed-row guarantee.

3 SETTING AND GUARANTEES

Learning from served items. There are m items, called *arms*, and a known number of rounds T, the *horizon*. An action S_u is a slate from a feasible family $\mathcal{F}$. The row $X_u \in [0,1]^m$ lists the potential rewards at round u. Serving S_u makes only the slice X_{u,S_u} observable. This partial feedback defines a semi-bandit. The learner chooses each action from earlier collected feedback and private random draws. Our learner collects feedback during exploration. After commitment it serves its chosen slate through T and stops collecting rewards.

A missing interaction. Write $X = (X_1, \dots, X_T)$ for the reward table. A request names one row t, independently of this table and the learner's randomness. Its feedback is replaced by the marker $\perp$:

$$X^{(-t)} = (X_1, \dots, X_{t-1}, \perp, X_{t+1}, \dots, X_T). \tag{1}$$

The calendar slot and served action remain. The learner's rule for a missing row defines its *erased execution*. We call its original complete-data execution the *factual execution*.

What replay receives. The ordinary retained log contains the factual run's acquired reward slices other than the erased one, their exploration-action metadata, the marker, and the saved private random draws. These records include feedback collected after t. Replay must reconstruct the erased execution using this log alone. Definition A.3 formalizes the access contract.

Exactness. The public transcript $A(X)$ consists of the served actions and phase decisions to continue or commit, including the selected slate. Rewards, scores, and random draws are private. Write $\mathcal{L}$ for a probability law and Z_t^{unl} for the reconstructed transcript. Exact replay requires

$$\mathcal{L}(Z_t^{\mathrm{unl}}) = \mathcal{L}(A(X^{(-t)})). \tag{2}$$

Our construction matches the reference on each saved sequence of random draws. Both exactness and stability below hold for every bounded deterministic table X.

Protection within a category. A fixed rule ϕ labels each action; $V_t = \phi(S_t)$ is the category served at the requested row. Total variation measures the largest change in an event probability: $d_{\mathrm{TV}}(P,Q) = \sup_E |P(E) - Q(E)|$. For a budget $\rho \in (0,1]$, category-wise stability requires

$$d_{\mathrm{TV}}\Big(\mathcal{L}(A(X) \mid V_t = v), \mathcal{L}(A(X^{(-t)}) \mid V_t = v)\Big) \le \rho \tag{3}$$

for every table, requested row, and supported label v. The label is chosen before the deleted reward arrives, so it has the same probability $p(v)$ in both executions. With $P = \mathcal{L}(A(X))$ and $Q = \mathcal{L}(A(X^{(-t)}))$, the unconditional distance is

$$d_{\mathrm{TV}}(P,Q) = \sum_v p(v)\, d_{\mathrm{TV}}(P(\cdot \mid V_t = v), Q(\cdot \mid V_t = v))\,. \tag{4}$$

A constant label gives *calendar stability*, the average protection

$$d_{\mathrm{TV}}\Big(\mathcal{L}(A(X)), \mathcal{L}(A(X^{(-t)}))\Big) \le \rho. \tag{5}$$

At the other endpoint, $\phi(S) = S$ gives each complete slate its own budget.

4 THE COST OF CATEGORY-WISE STABILITY

Products and protected groups. A product action chooses one of $L \geq 2$ candidate blocks at each of d positions. The blocks are disjoint, each with w arms, so a slate serves $K = dw$ arms. The example in Figure 1 has $w = 1$ and $K = d$. Write $A_{u,k} \in [L]$ for the candidate chosen at position k in round u. We assume coverage $KL = \Theta(m)$: these blocks contain a constant fraction of the m arms. At each position the candidates are partitioned into G nonempty protected groups. Their sizes may differ. The label ϕ records the tuple of served groups. Thus G counts groups per position, while there may be G^d labels. Here $G = 1$ gives calendar protection and $G = L$ protects each full action.

Rewards and learning cost. For the regret analysis, the reward table is random and independent of all learner and requester randomness. Its coordinate means μ_i satisfy

$$\mathbb{E}[X_{u,i} \mid X_1, \ldots, X_{u-1}] = \mu_i \quad \text{almost surely.} \tag{6}$$

Rewards may be dependent within a row or across time; the conditional means stay fixed. Let $r(S) = \sum_{i \in S} \mu_i$ and let $S^\star$ maximize r. The gap $\Delta(S) = r(S^\star) - r(S)$ measures the loss from serving S. We measure learning cost by accumulated gaps, called *pseudo-regret*:

$$\widehat{R}_T = \sum_{u=1}^{T} \Delta(S_u), \qquad R_T = \mathbb{E}[\widehat{R}_T]. \tag{7}$$

Our product class $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ has one best slate. At every position its winning block's average mean reward exceeds each alternative by at least $\Delta \in (0, 1]$. The largest slate gap obeys $\Delta_{\max} \leq C_\Delta K \Delta$, for a fixed $C_\Delta \geq 1$. This keeps the losses of exploration comparable to the gap used for identification.

The cost has two sources. Finding good choices requires enough observations to separate their mean rewards. Keeping the final decision stable within each protected group requires further evidence. The next result quantifies both. The notation $\widetilde{\Theta}$ means matching bounds up to logarithmic factors; here the logarithm appears only in the upper bound.

Theorem 1 (Category-wise stability cost)**.** *On $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$, with $0 < \Delta \leq 1$, coverage $KL = \Theta(m)$, and $0 < \rho \leq 1$, the minimax expected pseudo-regret over all causal learners satisfying category-wise stability* (3) *for G nonempty groups at every position is*

$$\widetilde{\Theta}\left(\min\left\{ \underbrace{K\Delta T}_{\textit{horizon}}, \quad \underbrace{\frac{m}{\Delta}}_{\textit{identification}} + \underbrace{\frac{KG}{\rho}}_{\textit{stability}} \right\} \right). \tag{8}$$

The constants absorb the fixed gap-regularity and coverage constants. The lower bound allows reward-dependent exploration. The Envelope learner attains the upper scale from the ordinary retained log without receiving Δ, including unequal group sizes.

Why these terms appear. At one position, finding a winner among L candidates takes order L/Δ^2 observations on a hard instance. Multiplying by the exploration loss across positions gives m/Δ. For $G \geq 2$, compare instances whose winners lie in different groups. Category stability keeps their action laws close for order $G/(\rho\Delta)$ rounds, forcing a constant fraction of suboptimal choices. Calendar stability gives the $G = 1$ endpoint. These losses supply KG/ρ, subject to the horizon cap. Appendices C and D prove the lower bounds; Section 5 gives the attaining learner.

Corollary 1 (Equal entropy, different stability costs)**.** *Take a covered two-position product with $L \geq 2$ candidates per position, in the class of Theorem 1. Under a uniform product reference distribution, the relative-offset label $A_{t,2} - A_{t,1} \pmod L$ and first-choice label $A_{t,1}$ each have L equally likely values, L slates per value, label entropy $\log L$, and residual action entropy $\log L$. Their minimax stability terms are, respectively,*

$$\widetilde{\Theta}(K/\rho) \qquad \textit{and} \qquad \widetilde{\Theta}(KL/\rho) = \widetilde{\Theta}(m/\rho). \tag{9}$$

Both share the identification term. When K/ρ dominates identification and the horizon cap is inactive, the minimax rates differ by order L, up to logarithmic factors. Uniform product probes attain the upper bounds; the lower bounds apply to causal adaptive exploration.

The separation follows the sampling opportunities inside a category. An offset leaves every candidate available at each position. A first-choice label concentrates all first-position observations on one candidate. Proposition C.4 extends this pair to the best and worst category arrangements at any feasible equal-size category count. The same principle suggests sampling groups as well as individual candidates.

Connection to private learning. Pure differential privacy bounds likelihood ratios of the full public transcript by e^{ϵ}, where ϵ is its budget. Under present-row versus missing-row adjacency, it gives category-wise TV at $\rho = \tanh(\epsilon/2)$ for every label $\phi(S_t)$. For $0 < \epsilon \leq 1$ on i.i.d.-row products, positionwise AdaP-UCB (Azize and Basu, 2022) gives $\widetilde{O}(m/\Delta + md/\epsilon)$ regret; block-aggregated CUCB-DP (Chen et al., 2020) gives $\widetilde{O}(md/\Delta + md/\epsilon)$. Appendix A.7 derives these bounds and the budget calibration used in our experiments. Appendix C.5 develops probe design and overlapping categories; Appendix E treats heterogeneous singleton gaps.

5 Exact Replay with a Completion Envelope

From evidence to a decision. With complete feedback, Envelope follows ordinary phased training; the envelope specifies its missing-row rule. It explores in consecutive phases I_p of lengths $1, 2, 4, \ldots$, truncated at T. Each reached phase is acquired in full. Its scores propose a slate and determine a commitment probability, called a *gate*. A saved independent draw $U_p \sim \mathrm{Unif}[0, 1)$ tests that gate. Continuing resets the scores before the next phase.

What happens at a missing row. Imagine filling the missing served rewards in every possible way. A vector $v \in [0, 1]^d$ gives one block-average reward per position. Let $g_{\mathrm{comp}}(v)$ be the commitment probability from that completion, after recomputing both the proposed slate and its evidence. The *completion envelope* uses the largest such probability:

$$g_{\mathrm{env}} = \max_{v \in [0,1]^d} g_{\mathrm{comp}}(v), \qquad \{U_p \geq g_{\mathrm{env}}\} \subseteq \{U_p \geq g_{\mathrm{comp}}(v)\} \quad \text{for every } v. \tag{10}$$

The inclusion is the access guarantee. If erased replay continues, every compatible complete-data execution continues, including the factual one. Hence each phase replay needs is already in the log. Figure 2 shows how replay can instead commit earlier. The envelope is the smallest local gate that gives this shared-draw containment (Proposition B.1).

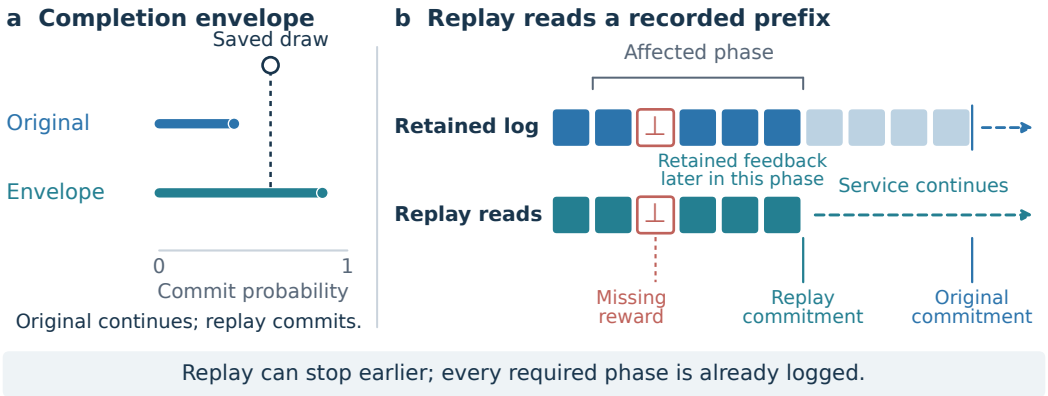

Figure 2: Replay uses a prefix of collected feedback. The envelope raises the missing phase's commitment probability. A shared draw can make replay commit while factual learning continues. Both runs use the full affected phase, including retained feedback after the missing row. An earlier replay commitment leaves a factual suffix unused; service continues after feedback collection ends.

Sampling for coverage and protection. We mix two probes with equal probability: draw a candidate uniformly, or draw a group uniformly and then a candidate within it. Let $h_{k,g}$ be group g's size at position k, and let $\mathrm{grp}_k(a)$ be candidate a's group. The candidate probability $q_{k,a}$ and group probability $Q_{k,g}$ are

$$q_{k,a} = \frac{1}{2L} + \frac{1}{2Gh_{k,\mathrm{grp}_k(a)}}, \qquad Q_{k,g} = \sum_{a:\mathrm{grp}_k(a)=g} q_{k,a} = \frac{h_{k,g}}{2L} + \frac{1}{2G}. \tag{11}$$

Every candidate receives probability at least $1/(2L)$; every group receives at least $1/(2G)$. Probes are independent across positions and rounds, and independent of rewards and other learner randomness. For candidate a in group g, conditioning on that group gives probability $q_{k,a}/Q_{k,g}$. Thus category conditioning magnifies its sampling probability by at most $1/Q_{k,g} \le 2G$, even when conditioning on the full group tuple. For example, with 16 candidates in groups of sizes 1 and 15, uniform probing visits the singleton with probability $1/16$. The mixture raises this to $9/32$ while still sampling every candidate in the large group. Protection within the singleton then uses more of the learner's observations.

Scores and the weakest advantage. Write $X_{u,k,a}$ for candidate block a's average reward at position k in round u. Dividing an observation by its sampling probability makes the candidate scores comparable:

$$z_{k,a} = \sum_{\substack{u \in I_p \\ X_u \neq \perp}} \frac{\mathbf{1}\{A_{u,k} = a\} X_{u,k,a}}{q_{k,a}}, \qquad \tau := \frac{1}{\min_{k,a} q_{k,a}} \le 2L. \tag{12}$$

The scale τ bounds one row's contribution to a score. Let $j_k(z)$ maximize the score at position k, with saved random draws breaking ties, and let $j(z) = (j_1(z), \dots, j_d(z))$ be the proposed slate. For a fixed proposal j, define its *certificate*

$$C_j(z) = -\tau \log\left(\sum_{k=1}^{d} \sum_{a \neq j_k} \exp\left(\frac{z_{k,a} - z_{k,j_k}}{\tau} \right) \right). \tag{13}$$

It smoothly approximates the smallest score advantage over all local alternatives. A close competitor lowers the certificate, so commitment waits for evidence at every position. The complete-data rule evaluates $C_{j(z)}(z)$.

Calibrating commitment. Two quantities set the evidence requirement. Sampling uncertainty depends on τ. Stability depends on the largest category inflation, $\Lambda = 1/\min_{k,g} Q_{k,g} \le 2G$. For phase length $n = |I_p|$, define the uncertainty width β_p, the reserve θ_p, and the gate slope s:

$$\begin{aligned} \beta_p &= 2\left(\sqrt{n\tau \log(dLT)} + \tau \log(dLT) \right), & \theta_p &= \tau + 2\beta_p, \\ s &= \frac{\rho}{e\Lambda}, & g_p(c) &= \min\{1, s(c - \theta_p)_+\}. \end{aligned} \tag{14}$$

Here $(x)_+ = \max\{x, 0\}$. The reserve covers sampling noise and one possible missing contribution. Above it, the gate rises linearly until it reaches one. For the retained proposal, the best filling increases its certificate over the factual filling by at most $e\Lambda$ in expectation, conditional on the category. The slope therefore keeps expected gate movement within ρ. Proposition C.3 gives the calibration for general fixed labels and joint probes.

Using a probability turns changes in evidence into controlled changes in decisions. Couple two phases with the same uniform draw. Their stopping decisions differ only when that draw lies between their gates, an interval whose length is the absolute gate change. When both continue, resetting the scores couples their future learning. The reserve makes simultaneous commitments agree on their slate. This links expected gate movement to stability of the entire public run.

Evaluating all completions in one step. Let z^0 contain the retained scores in the missing phase, $j = j(z^0)$ its proposal, and $b_k = A_{t,k}$ the recorded candidate in the deleted row. A filling adds

$\mathbf{1}\{a = b_k\}v_k/q_{k,b_k}$ to score $z^0_{k,a}$. The envelope is attained by adding the largest possible missing reward to a served retained winner, and zero to a served loser:

$$z^+_{k,a} = z^0_{k,a} + \frac{\mathbf{1}\{a = b_k = j_k\}}{q_{k,a}}, \qquad g_{\mathrm{env}} = g_p(C_j(z^+)). \tag{15}$$

To see why, suppose some filling changes the winner. Its local advantage is then at most one row contribution, so the reserve closes every completion gate. Otherwise all fillings keep the same proposal, and raising its scores maximizes its gate. Lemma B.5 proves both cases. On commitment, replay serves $j(z^0)$. The hypothetical reward is used to evaluate the gate; the log keeps its missing marker.

Algorithm 1. Envelope learning and retained-log replay

Inputs: T, ρ, product structure, protected groups, and private probe, tie-breaking, and phase random draws.

1. For each phase of length $1, 2, 4, \ldots$, truncated at T, acquire its probes using (11).
2. Compute present-row scores, proposal j, and reserve using (12)–(14).
3. Set $g = g_p(C_j(z))$ for an intact phase; use $g = g_p(C_j(z^+))$ for a missing phase.
4. If the saved draw $U_p < g$, stop acquisition and serve j through T.
5. Otherwise, reset the scores and start the next phase.

A request after factual acquisition ends leaves the transcript unchanged.

Theorem 2 (Envelope learner)**.** *For every bounded reward table and every fixed requested row, the Envelope learner satisfies category-wise ρ-stability* (3) *and reconstructs its specified erased execution pathwise from the ordinary retained log and saved randomness. Under the conditional-mean model* (6)*, uniformly over $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ with $0 < \Delta \le 1$, coverage $KL = \Theta(m)$, and $0 < \rho \le 1$, it satisfies*

$$R_T = O\left(C_\Delta \min\left\{K\Delta T, \frac{m\log(dLT)}{\Delta} + \frac{KG}{\rho}\right\}\right). \tag{16}$$

The learner receives neither reward gaps nor means. Let $\widehat{R}^{(-t)}_T$ be the pseudo-regret of separately replaying row t on the factual tape, and set $\widehat{R}^{(\varnothing)}_T = \widehat{R}_T$. Define

$$M_T := \max\left\{\widehat{R}^{(\varnothing)}_T, \max_{t\in[T]} \widehat{R}^{(-t)}_T\right\}. \tag{17}$$

The same upper bound controls $\mathbb{E}[M_T]$. Section B.5 proves replay, stability, and this simultaneous utility guarantee.

The rate follows the two evidence requirements in (14). Concentration makes the proposed slate correct. Beyond the uncertainty reserve, saturating the gate needs certificate height $e\Lambda/\rho$, contributing order $\Lambda/(\rho\Delta)$ exploration rounds. Each costs at most $C_\Delta K\Delta$, yielding the KG/ρ term. The logarithm multiplies identification. Theorem B.1 extends the construction to matroids; Corollary F.1 covers causally registered requests.

6 Experiments

We test the category cost law, retained-feedback replay, and the joint cost of regret and feedback collection. Envelope uses the stated constants. Large-horizon costs use exact sampling of joint sufficient statistics; replay and private-learner comparisons use chronological observations. Appendix G gives the reward laws, grids, seeds, complete results, and independent implementation checks.

Category cost appears when protection requires more evidence. Figure 3a varies the group count and stability budget on the same product instance. Its control commits as soon as the certificate clears its uncertainty reserve. At $\rho = 0.05$, all group counts coincide with this control: the evidence needed for identification already saturates the gate. At $\rho = 3 \times 10^{-6}$, protecting $G = 8$ groups costs 6.30

times $G = 1$ (paired 95% interval $[5.05, 7.83]$). The mean includes four horizon-capped runs. Four- and eight-position products show the same direction (Appendix G.1).

The equal-entropy comparison isolates category arrangement. A fixed reward law keeps $\Delta = 0.2$, while a common horizon $T = 2^{36}$ lets stability dominate identification. At $\rho = 10^{-8}$, first-choice/offset regret ratios are 4.10, 7.53, and 13.79 for $L = 4, 8, 16$ (Figure 3b). All 3456 runs commit correctly. Phase records separate the evidence needed for identification from waiting for stable commitment (Appendix G.1).

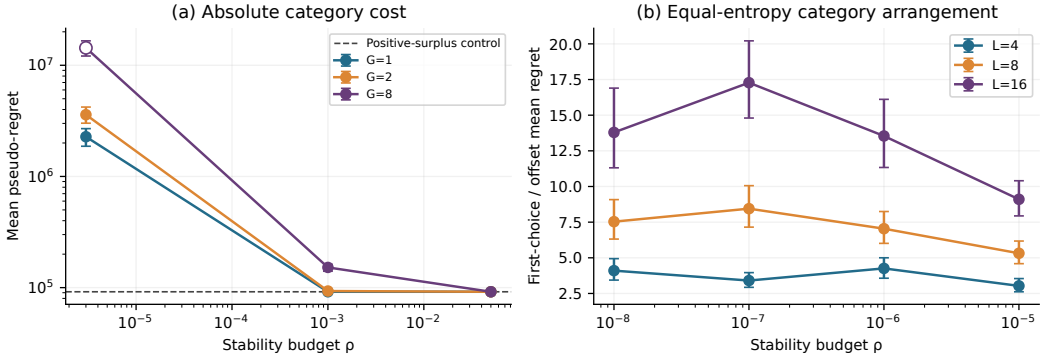

Figure 3: The two parts of the cost law. (a) Absolute regret for balanced groups at $L = 8$, $\Delta = 0.2$, $T = 2^{26}$; 64 paired seeds and a matched certificate-only control. (b) First-choice/offset regret ratios at $\Delta = 0.2$, $T = 2^{36}$; 128 paired seeds per candidate count and four budgets. Both panels have two positions. Whiskers are pointwise 95% whole-seed bootstrap intervals. Open markers in (a) include horizon-capped trajectories; the scaling grid in (b) has no caps.

Exact replay stays within the collected observations. The replay implementation receives sanitized reward slices and saved randomness. A guard rejects any read of the deleted payload or an uncollected reward. A separate reference runs on the marked table and enumerates possible completions. Every named request in the stationary Bernoulli grid erases acquired feedback. Table 1 compares Envelope with omit-only replay, which drops the reward contribution and otherwise uses the complete-data gate. Every blocked read occurs at round 4096, after factual acquisition ended at 4095. Independent reconstruction verifies each blocked case.

Stationary-grid check	Envelope	Omit-only
Acquired-feedback deletions	4,608	4,608
Exact reference matches	4,608	4,594
Blocked unavailable reads	0	14

Table 1: Replay from the same factual logs. Each rule is checked against its own erased reference, including commitment, selected slate, and public trajectory. Each of 512 independent seeds is evaluated across nine stationary reward-law and budget settings. The complete protocol is in Appendix G.4.

The gate determines when this obstacle appears. Of 4608 first positive gates, 3708 are saturated. The omitted contribution opens a continuation interval on 441 requests; 14 saved draws fall inside it. Envelope commits earlier on five requests. These records connect the access obstruction to a movable stopping boundary.

Group-aware sampling with dependent rewards and varied gaps. For groups $(1, L-1)$, mixed probing reduces category inflation from L to $4L/(L+2) < 4$. At the same partition and budget, we vary the within-instance gap spread by factors of 1, 4, and 16, using two, four, or eight positions. Winners occupy singleton groups, large groups, or alternating placements. Dependent rows share one reward shock with probability 0.8, preserving their marginal means. Across 108 comparisons, mixed probing collects 9.9–49.3% fewer rows and reduces mean regret in 100. The other eight settings illustrate the cost of extra exposure to a low-reward singleton. All 10,368 learner and control outcomes commit correctly (Appendix G.2).

Preference-derived semi-bandits. MovieLens 1M, Jester ratings, and KuaiRec viewing records supply empirical reward laws (Harper and Konstan, 2015; Goldberg et al., 2001; Gao et al., 2022). A design cohort selects four items per position. Each round resamples a complete preference vector from the evaluation cohort, revealing the two served rewards. Scaled ratings and binary viewing indicators preserve dependence and varied gaps. We use 32 paired seeds, each averaging four placements of groups $(1, 3)$.

Private AdaP-UCB and binary-reward DP-KLUCB (Azize and Basu, 2022; Azize et al., 2025) use whole-row $\epsilon = 2\,\mathrm{atanh}(\rho)$, split across positions. They protect every action category and collect feedback throughout. Envelope protects the specified groups and stops collecting at commitment. Figure 4 reports regret and collection across ten budgets at $T = 2^{28} - 1$.

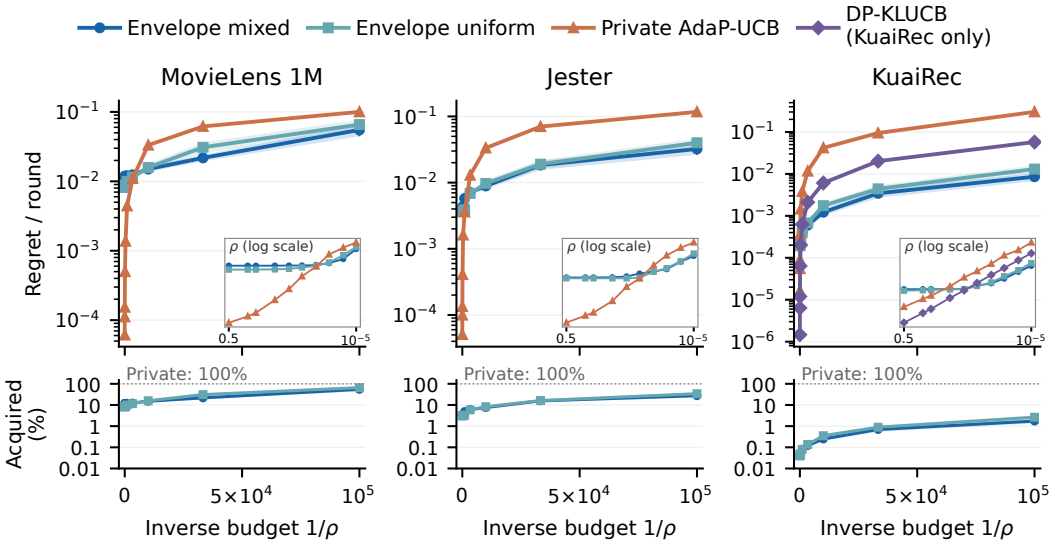

Figure 4: Regret and collection on a linear inverse-budget axis $1/\rho$. Insets show all ten budgets on a log scale. Bands are pointwise 95% whole-seed intervals. Private methods collect 100%; DP-KLUCB uses binary KuaiRec only. All 4800 Envelope runs commit correctly.

The full curves locate the crossover. Private adaptive learners lead at loose budgets. Mixed Envelope has lower mean regret than AdaP-UCB at every tested $\rho \leq 10^{-4}$ on all three sources. At 10^{-5}, the reductions are 45.1%, 72.5%, and 97.2%. Against DP-KLUCB on KuaiRec the reduction is 85.1%. All four paired 95% intervals exclude zero. Appendices G.7–G.8 give the paired intervals, four-horizon study, and private-index audit.

Categories from movie metadata. A separate MovieLens catalogue uses Action/Comedy and Action/Drama groups chosen from design availability and metadata. At $\rho = 10^{-5}$, group-aware probing reduces regret against uniform probing by 31.7% ([20.3, 42.6]%) and collects 19.1% of the horizon's feedback, versus 25.2% for uniform probing (Table 9).

After physical payload removal, full-log and affected-phase replay reconstruct all 1924 acquired-feedback erasures and 144 inactive requests, with zero forbidden reads. Omit-only also matches this dataset grid. Here 32/72 first positive gates are saturated; 40 requests have nonempty omit-only continuation intervals (Appendix G.3).

7 Conclusion

Protected categories set the cost of limiting an erased response's influence. The completion envelope makes retained feedback sufficient for the specified erased run. Together, they link stable learning to a system's feedback-collection policy.

AI USE STATEMENT

We used generative AI to help edit prose, check internal consistency, design conceptual diagrams and implement diagnostic code.We reviewed the manuscript and take responsibility for its text, claims, proofs, and figures.

REPRODUCIBILITY STATEMENT

All numbered theorems state their assumptions in their statements. A theorem-to-proof map opens the appendix. Section A formalises the erasure and access contracts and proves calendar contraction. The envelope learner, concentration, exact replay, stability, and regret are in Sections B and B.5; category specialisations are in Section C; identification lower bounds are in Section D. Heterogeneous singleton gap-class minimax is a self-contained appendix chapter (Theorem E.1 and Section E). Section G specifies the experimental grids, seeds, completion rules, and indexed-log storage contract. The observed-data experiments use MovieLens, Jester, and KuaiRec. The supplement provides official downloaders, checksums, and deterministic cohort preparation. The appendix specifies the regret grids, correlated and heterogeneous reward models, private-baseline calibration and Envelope ablations, chronological replay, and analytic stability audit. The accompanying supplement contains executable runners, fixed configurations, per-run and phase records, independent implementation checks, and figure-generation code. The project README gives the full commands for dataset preparation, new simulations, independent checks, report regeneration, and manuscript compilation. Archived phase statistics and primitive private episodes permit direct verification of recorded outcomes. All capped trajectories remain in the reported means.

REFERENCES

Achraf Azize and Debabrota Basu. When privacy meets partial information: A refined analysis of differentially private bandits. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Achraf Azize and Debabrota Basu. Concentrated differential privacy for bandits. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024. arXiv:2309.00557v3.

Achraf Azize, Yulian Wu, Junya Honda, Francesco Orabona, Shinji Ito, and Debabrota Basu. Optimal regret of bandits under differential privacy. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Mohammad Bavarian, Badih Ghazi, Elad Haramaty, Pritish Kamath, Ronald L. Rivest, and Madhu Sudan. Optimality of correlated sampling strategies. *Theory of Computing*, 16(12):1–18, 2020.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 1–10. ACM, 2014.

Xiaoyu Chen, Kai Zheng, Zixin Zhou, Yunchang Yang, Wei Chen, and Liwei Wang. (Locally) differentially private combinatorial semi-bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1757–1767, 2020.

Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6028–6073, 2023.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pp. 371–380. ACM, 2009.

Hossein Esfandiari, Alkis Kalavasis, Amin Karbasi, Andreas Krause, Vahab Mirrokni, and Grigoris Velegkas. Replicable bandits. In *The Eleventh International Conference on Learning Representations*, 2023.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 2022. doi:10.1145/3511808.3557220.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842, 2020.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16319–16330, 2021.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi:10.1145/2827872.

Thomas Holenstein. Parallel repetition: Simplifications and the no-signaling case. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, pp. 411–419. ACM, 2007.

Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jon Kleinberg and Éva Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. *Journal of the ACM*, 49(5):616–639, 2002.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 535–543, 2015.

Thanh Nguyen-Tang and Raman Arora. Exact unlearning in reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2606.04182.

Touqir Sajed and Or Sheffet. An Optimal Private Stochastic-MAB Algorithm based on Optimal Private Stopping Rule. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5579–5588, 2019.

Enayat Ullah and Raman Arora. From adaptive query release to machine unlearning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34642–34667, 2023.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A. Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4126–4142, 2021.

Siwei Wang and Jun Zhu. Optimal learning policies for differential privacy in multi-armed bandits. *Journal of Machine Learning Research*, 25(314):1–52, 2024.

Zichun Ye, Runqi Wang, Xuchuang Wang, Xutong Liu, Shuai Li, and Mohammad Hajiesmaili. Unlearning offline stochastic multi-armed bandits. *arXiv preprint arXiv:2605.00638*, 2026.

PROOF GUIDE

Result	Proof route
Theorem 1	Identification lower bound: Appendix D and Corollary D.1. Category lower bound: Lemma C.1 and Proposition C.1. Upper bound: Theorem 2.
Corollary 1	Conditional inclusion and the two category-specific lower bounds: Proposition C.3 and Appendix C.
Theorem 2	Concentration: Lemma B.4. Envelope containment and stability: Lemmas B.5 and B.6. Retained-log reconstruction: Lemmas A.4 and B.7. Utility: Appendix B.5.

The appendix also gives the matroid envelope (Theorem B.1), heterogeneous singleton rates (Theorem E.1), and causal request registration (Corollary F.1). Their statements and proofs are self-contained. Experimental methods and records are in Appendix G.

Appendix A.7 derives the private-bandit comparison and the exact pure-DP-to-category-TV calibration. Appendix C.5 proves the probe-design separation and overlapping-category extension. Proposition C.4 and Corollary C.5 characterize optimal equal-size category arrangements, with complete constructions and lower bounds. The stationary replay confirmation and cost-mechanism experiments are in Appendices G.4 and G.1.

A FORMAL UNLEARNING MODEL AND CALENDAR LAW

This appendix fixes the erasure and replay contracts used throughout the paper, proves calendar contraction, and records the causal-prefix form used by the lower bounds. The ordinary-log sufficiency criterion and a fixed-target obstruction appear in Appendix A.5.1 and Propositions A.1 and A.2.

A.1 TRANSCRIPT, CAUSALITY, AND CALENDAR REWARD ERASURE

Fix a horizon $T \in \mathbb{N}$ and a ground set $[m]$. A *table* is a sequence $X = (X_1, \dots, X_T)$ with $X_s \in [0,1]^m$ for each s. Learner coins ξ take values in a standard Borel space and supply all randomness in $\mathcal{L}(A(X))$. The learner is *causal*: the action S_t at round t is a measurable function of ξ and of earlier served actions and their ingested feedback slices only. After S_t is served, the learner may ingest the coordinates of X_t indexed by S_t.

The *public transcript* $A(X)$ consists of the served actions and phase continue/commit symbols, including the committed candidate. Reward slices, numerical scores, certificates, gate values, and saved learner coins remain private. They may be retained under the replay access contract below without becoming outputs of the stable mechanism. All stability and lower-bound comparisons use this same public interface; an additional numerical release requires its own stability argument.

The deletable record at a named time t is the post-action feedback X_{t,S_t}, not the system-generated action S_t and not the calendar key t. For a named request $t \in [T]$, the erasure table is

$$X^{(-t)} = (X_1, \dots, X_{t-1}, \perp, X_{t+1}, \dots, X_T),$$

as in (1). The symbol $\perp$ denotes absence of feedback at that event, not a numerical reward. The slot t remains on the timeline and still serves S_t, because both lie outside the deletion scope. Two tables are *calendar-adjacent* when they differ at exactly one time, one of the two values being $\perp$ and the other a real row in $[0,1]^m$. Replacement of one real row by another real row is a deletion followed by an insertion, and therefore factors as two calendar adjacencies.

Lemma A.1 (Erasure/zero-contribution equivalence)**.** *Fix a served-action trajectory and the learner coins. In every additive reward accumulator used here, omitting the feedback at time t is equivalent to assigning that summand the value zero. Calendar indices, phase lengths, and exposure counts remain unchanged. For marker-unaware additive-score constructions, this also identifies their full reward-dependent state. Envelope gates additionally use the missing marker, so a missing row and a real zero-valued row need not induce the same gate.*

Proof. For an arm sum, the identity is

$$Y_i^{(-t)} = \sum_{\substack{s: i \in S_s \\ s \neq t}} X_{s,i} = \sum_{s: i \in S_s} X_{s,i} \mathbf{1}\{s \neq t\}.$$

The same identity holds for block sums and propensity-weighted sums. Exposure counts $N_i = \sum_s \mathbf{1}\{i \in S_s\}$ depend on the served trajectory, not the reward value. Subsequent state is a function of these sums, counts, coins, and, for envelope algorithms, the missing markers. The identity concerns omission from a sum; it does not permit an unlearner to read the deleted payload and subtract it from a contaminated aggregate. □

A.2 STABILITY, EXACTNESS, AND ACCESS MODELS

Definition A.1 (Calendar ρ-TV stability)**.** A (possibly randomized) learner A is calendar ρ-TV-stable if, for every table X and every request $t \in [T]$,

$$d_{\mathrm{TV}}\big(\mathcal{L}(A(X)), \mathcal{L}(A(X^{(-t)}))\big) \le \rho. \tag{A.1}$$

This is (5) of the main text.

Lemma A.2 (Table-independent request times)**.** *Let A be calendar ρ-TV-stable, and let $t_\star$ be a random time independent of the table and of the learner coins (requester coins are allowed). Then*

$$d_{\mathrm{TV}}\big(\mathcal{L}(A(X)), \mathcal{L}(A(X^{(-t_\star)}))\big) \le \rho.$$

If $t_1, \dots, t_r$ are (possibly random) times whose joint law is independent of the table and of the learner coins, and if the same one-step stability guarantee holds on inputs already containing up to $r-1$ missing rows, the table with those r rows replaced by $\perp$ satisfies the same bound with ρ replaced by $r\rho$. Write $X^{(-D)}$ for the table with calendar rows in D replaced by $\perp$. Equivalently, the r-request claim assumes that for every real X, every $D \subseteq [T]$ with $|D| < r$, and every $t \notin D$,

$$d_{\mathrm{TV}}\big(\mathcal{L}(A(X^{(-D)})), \mathcal{L}(A(X^{(-(D\cup\{t\}))}))\big) \le \rho.$$

The one-request mixing statement does not use that extended domain. This lemma is a conditional composition principle: the envelope theorem establishes its one-erasure guarantee only, and does not instantiate repeated deletion. The bound does not apply when the next named time is a function of the unlearned transcript: the mixing measure would then depend on the table, as in the adaptive-request setting studied by Gupta et al. (2021); Chourasia and Shah (2023). Extending our guarantee to that setting remains open.

Proof. Write π for the law of $t_\star$. Independence of the requester coins from the table and from the learner coins makes $\mathcal{L}(A(X))$ equal to the π-mixture of the same law at every t, and makes $\mathcal{L}(A(X^{(-t_\star)}))$ the π-mixture of $\mathcal{L}(A(X^{(-t)}))$. Convexity of total variation and (A.1) at every fixed t give the one-time bound. For r times, dummy them in order $1, \dots, r$. Condition on $(t_1, \dots, t_r)$: each step dummies a fixed time on the table obtained from the previous step, which by the extended-domain hypothesis moves the transcript law by at most ρ. The triangle inequality yields $r\rho$, and mixing over the times preserves the bound. Real-table one-erasure stability (A.1) does not by itself justify that step. □

Definition A.2 (Exact unlearning)**.** Let Z_t^{unl} be the unlearner's output. Exact unlearning at time t requires

$$\mathcal{L}(Z_t^{\mathrm{unl}}) = \mathcal{L}\big(A(X^{(-t)})\big). \tag{A.2}$$

This is (2) of the main text. The target is equality of transcript laws, not pathwise reconstruction of every coin-wise continuation. The permitted inputs are specified by the access contract, not by this equality.

Write $\mathcal{L}_t$ for the *sanitized on-policy log* formed from the completed factual run for erasure of row t: the learner coins ξ, the missing marker at time t, and every played slice $\{(s, S_s, X_{s,S_s}) : s \neq t\}$ that the factual run ingested, including slices later than t, together with the relevant action and probe metadata. This retained log is private and is distinct from $A(X)$. This object is not the prefix available at calendar time t. An *environment oracle* is a map $\mathcal{O}_X(s, S) := X_{s,S}$ returning an arbitrary coordinate slice of the underlying table.

Definition A.3 (Ordinary-log and oracle-assisted exactness)**.** *Ordinary-log exact unlearning* is a (possibly randomized) map Φ, depending on the request time but not on the unknown table, such that $\Phi(\mathcal{L}_t) \overset{d}{=} A(X^{(-t)})$ for every table X. This is a realization of (A.2) whose only input is the sanitized log. *Oracle-assisted exact unlearning* is a map $U^{\mathcal{O}_X}$ that may query $\mathcal{O}_X$ and that satisfies (A.2).

Calendar ρ-TV stability guarantees a coupling of $\mathcal{L}(A(X))$ and $\mathcal{L}(A(X^{(-t)}))$ with disagreement probability at most ρ (Ullah et al., 2021; Ullah and Arora, 2023). Realizing an exact unlearning map from the sanitized ordinary log requires a separate sufficiency argument (Lemma A.4). Access to the underlying reward table is sufficient for an exact rerun. The coupling statement alone is neither an ordinary-log implementation nor an end-to-end runtime bound. Ordinary-log exactness is the stronger operational guarantee. Oracle-assisted exactness is still exact in the sense of Definition A.2.

A.3 PROOF OF CALENDAR CONTRACTION

Theorem A.1 (Calendar contraction)**.** *Let A be calendar ρ-TV-stable in the sense of Definition A.1. Let ν and ν' be two laws on reward rows in $[0,1]^m$, and let n be a calendar length with $n \le T$. Write ν^n and $(\nu')^n$ for the product laws of n independent rows, padded by an arbitrary common continuation on rounds $n+1, \dots, T$ if $n < T$. Then*

$$d_{\mathrm{TV}}\big(\mathcal{L}(A(\nu^n)), \mathcal{L}(A((\nu')^n))\big) \le 2\rho n\, d_{\mathrm{TV}}(\nu, \nu'). \tag{A.3}$$

Implementation	Gap information	Stability	Replay access proved
Product category label, frozen covering	class gap lower bound	sitewise group-conditioned transcript	ordinary log
Product category label, omit-only doubling	unknown gap	sitewise group-conditioned transcript	table oracle
Envelope learner	unknown gap	sitewise group-conditioned transcript	ordinary log
Heterogeneous gap-class attainer	unknown min/max gap	calendar trajectory, singleton scope	ordinary log

Table 2: Access contracts for the body product learners and the heterogeneous singleton appendix. Ordinary log means a table-independent map of the sanitized factual log. A table oracle may return coordinates that the factual run never ingested.

Proof of Theorem A.1. Let $\eta = d_{\mathrm{TV}}(\nu, \nu')$. Couple the two row laws maximally and independently at the n initial calendar slots, with a common continuation thereafter. The resulting tables D, D' have expected Hamming distance $\mathbb{E}H = n\eta$. For fixed D, D', replace their H differing rows one at a time. At each replacement, the current real table and the next real table share the intermediate table with that one row missing. Calendar stability bounds each of the two edges by ρ; all other rows are real, so the one-erasure domain suffices. The triangle inequality gives

$$d_{\mathrm{TV}}\big(\mathcal{L}(A(D)), \mathcal{L}(A(D'))\big) \le 2\rho H.$$

Convexity of total variation in its two arguments now yields

$$d_{\mathrm{TV}}\big(\mathcal{L}(A(\nu^n)), \mathcal{L}(A((\nu')^n))\big) \le \mathbb{E}[2\rho H] = 2\rho n\eta.$$

The coupling is of whole rows and requires no independence within a row. Its rows and learner coins are analysis variables, not additional public outputs. □

A.4 Causal prefixes

The lower bounds in Appendix D apply Theorem A.1 to action prefixes rather than to full transcripts. The next lemma records that reduction.

Lemma A.3 (Causal calendar prefix)**.** *Let A be causal and calendar ρ-TV-stable. For $n \le T$ the prefix of actions $(S_1, \dots, S_n)$ is ρ-TV-stable as a mechanism on the first n rows: two laws on rows at total variation η induce prefix laws at total variation at most $2\rho n\eta$. In particular, on a two-point Bernoulli pair with gap $\Delta \in (0, 1/4]$,*

$$\sup_{\theta \in \{\nu, \nu'_a\}} R_T(A, \theta) = \Omega\big(\min\{\Delta T, 1/\rho\}\big). \tag{A.4}$$

The argument never reindexes records by pull count, and it does not force $\mathbb{E}_\nu[N_a] = \Omega(1/(\rho\Delta))$ on the base.

Proof. Append any common continuation to two n-row inputs and project their full transcripts onto $(S_1, \dots, S_n)$. Causality makes this prefix independent of the continuation. An erasure at $t < n$ inherits stability by data processing; erasure at n cannot change the prefix because S_n precedes its feedback. Applying Theorem A.1 proves the prefix contraction.

For the two-point claim, let ν have a baseline arm of mean $1/2$ and a disputed arm a of mean $1/2 - \Delta$. Let ν'_a raise only arm a to $1/2 + \Delta$, with $0 < \Delta \le 1/4$. Their Bernoulli row laws have TV distance 2Δ. For $N_a(n)$, the number of plays of a in the first n rounds,

$$\begin{aligned} R_n(A, \nu) + R_n(A, \nu'_a) &= \Delta\big(\mathbb{E}_\nu N_a(n) + n - \mathbb{E}_{\nu'_a} N_a(n)\big) \\ &\ge \Delta n(1 - 4\rho n\Delta). \end{aligned}$$

The inequality uses $0 \le N_a(n) \le n$ and the prefix TV bound. Choose $n = \min\{T, \lfloor 1/(32\rho\Delta)\rfloor\}$ when this is positive. Then $4\rho n\Delta \le 1/8$, and one environment has regret at least $7\Delta n/16$. Nonnegative instantaneous regret gives the same lower bound at horizon T. If the floor is zero, $\rho\Delta > 1/32$; the first action has the same law in both environments, so their first-round regrets sum to Δ and one is at least $\Delta/2 > 1/(64\rho)$. These cases imply (A.4). For $\rho = 0$, use $n = T$: the two prefix laws coincide and the same sum is ΔT. □

Corollary A.1 ($K = 1$ multi-armed bandit)**.** *For k singleton arms, $K = 1$ and $m = k$. Calendar contraction has the factor n, the number of calendar rows. The two-point pair of Lemma A.3 therefore has unlearning cost*

$\Omega(\min\{\Delta T, 1/\rho\})$*; statistical identification of that pair requires order* $1/\Delta^2$ *observations. The restriction* $\Delta \le 1/4$ *belongs to that mean-*$1/2$ *Bernoulli family; the full product-gap range is handled by the separate constructions in Appendix D.*

A.5 Replay access

Exactness from a sanitized log requires a table-independent kernel from the retained log to the erased-transcript law. Lemma A.4 is a shared-tape sufficient criterion. Propositions A.1 and A.2 test a fixed learner and its existing missing-data target: a stable learner with positive coverage can still leave that target more distinguishing than the sanitized log.

A.5.1 A sufficient replay criterion

Lemma A.4 (On-policy replay closure)**.** *Couple the factual and erased executions on the learner tape* ξ *recorded in* $\mathcal{L}_t$*. Suppose every slice ingested by* $A(X^{(-t)})$ *is either the missing row at* t *or a slice stored in* $\mathcal{L}_t$*. Then a measurable map* Φ *of the sanitized log reconstructs* $A(X^{(-t)})$ *on this shared tape and hence is ordinary-log exact.*

Proof. Replay the learner with ξ. At t, supply the missing marker; at each other ingest, supply the stored slice. By induction, the replay has the erased execution's history, chooses its action, and receives its next observation. All later actions or releases that consume no feedback are functions of this reconstructed history and the tape. Thus the complete erased transcript is a function of $\mathcal{L}_t$. □

The criterion concerns *ingested slices*: equality of served actions is insufficient when the factual learner has stopped collecting feedback. It is a sufficient condition, not a characterization. Ordinary-log exactness requires only a common kernel Φ with $\mathcal{L}(\mathcal{L}_t)\Phi = \mathcal{L}(A(X^{(-t)}))$ for every table. A missing coordinate that changes a particular tape's continuation need not change this mixture law. The envelope constructions in Theorems 2 and B.1 satisfy the stronger shared-tape criterion by stopping acquisition no later after erasure.

Lemma A.5 (Two-copy pathwise conversion)**.** *Run two base copies with independent tapes on odd and even calendar slots, respectively, and couple the factual execution and every separate single-erasure execution on their saved tapes. Suppose the healthy copy's complete execution remains factual, and the affected copy's served actions agree with its factual served actions up to and including its retirement time. Retirement occurs after one of its owned actions. Subsequently the affected copy ingests no feedback and serves the healthy copy's immediately preceding action. For a fixed stationary mean vector and nonnegative arm gaps, if* G_a *and* G_h *are the two factual pathwise pseudo-regrets, then* $G_{\mathrm{erased}} \le G_a + 2G_h$*. In particular, for the same factual odd/even pair* (G_o, G_e) *and the simultaneous maximum* M_T *of* (17)*,*

$$M_T \le G_o + G_e + \max\{G_o, G_e\} \le 2\widehat{R}_T^{(\varnothing)}. \tag{A.5}$$

These are separate erasures of the original table, not cumulative deletions.

Proof. The affected prefix is a factual prefix and costs at most G_a. The healthy path costs G_h. Under odd/even ownership, every affected slot after retirement has a preceding healthy slot, and each such healthy action is copied at most once. Its total copied gap is therefore at most G_h. Adding gives the inequality. The same two factual regrets apply to every named row, proving (A.5), including the factual member of the maximum. If a sampler terminates almost surely, the statement is on the common probability-one event where all required factual calls terminate. There are finitely many such calls at a finite horizon; saved-prefix replay adds no new call and no deletion-index failure charge. A nondecreasing one-copy expected bound b therefore implies $\mathbb{E}M_T \le 2[b(n_+) + b(n_-)]$, with $n_+ = \lceil T/2 \rceil$ and $n_- = \lfloor T/2 \rfloor$. □

A.5.2 Log laws versus erased transcripts

Proposition A.1 (Log-versus-target separation)**.** *Let* μ_X *be the sanitized-log law,* ν_X *the erased-transcript law, and* Φ *a table-independent kernel from logs to transcripts. For* $\varepsilon = \sup_X d_{\mathrm{TV}}(\mu_X\Phi, \nu_X)$ *and any pair of tables,*

$$\varepsilon \ge \frac{1}{2}\big[d_{\mathrm{TV}}(\nu_{X+}, \nu_{X-}) - d_{\mathrm{TV}}(\mu_{X+}, \mu_{X-})\big]_+.$$

Proof. The triangle inequality through $\mu_{X+}\Phi$ and $\mu_{X-}\Phi$, followed by data processing, gives $d_{\mathrm{TV}}(\nu_{X+}, \nu_{X-}) \le 2\varepsilon + d_{\mathrm{TV}}(\mu_{X+}, \mu_{X-})$. □

Proposition A.2 (Positive coverage does not yield ordinary-log exactness)**.** *For every* $\rho \in (0, 1]$*, there is a four-arm, three-round causal learner that is calendar* ρ*-TV-stable on every bounded reward table, has strictly positive round-2 ingestion probability for every arm on every table, yet every table-independent ordinary-log kernel has worst-case transcript error at least* $3\rho/16$*. The bound remains valid when the log retains all*

independent learner coins. The constant kernel returning $(3, 4, 4)$ *has error at most* ρ*, so the order in* ρ *is sharp for this generic obstruction.*

Proof. Fix $\eta = 1/2$. Draw $B \sim \mathrm{Bernoulli}(\rho)$ independently of the table, and retain this coin together with all subsequent action-sampling coins. Serve arm 3 at round 1 and ingest its feedback when present. If $B = 0$, serve arm 4 at round 2. If $B = 1$ and the first reward exceeds $1/2$, draw the round-2 arm uniformly from $[4]$. Otherwise, including missingness of the first reward, use $q = (1 - \eta)\delta_2 + \eta\,\mathrm{Unif}([4])$. Ingest the served round-2 feedback when present. Serve arm 1 at round 3 exactly when $B = 1$ and the second reward exceeds $1/2$; otherwise serve arm 4. A missing second reward makes that test false. All choices precede the current row's feedback.

Couple factual and erased executions on identical coins. On $\{B = 0\}$ both action paths are $(3, 4, 4)$ for every table and every deleted row, so disagreement probability is at most $\Pr(B = 1) = \rho$. The erased-transcript law therefore assigns probability at least $1 - \rho$ to $(3, 4, 4)$, proving the constant kernel's error bound without requiring the factual transcript as a replay input. Every arm has round-2 factual ingestion probability at least $\rho\eta/4 > 0$.

Take binary tables X^0, X^1 with $X^b_{1,3} = 1$, $X^b_{2,2} = b$, and all other entries zero, and delete row 1. Factually, conditional on $B = 1$, the round-2 arm is uniform. The two sanitized logs agree unless $B = 1$ and $A_2 = 2$, an event of probability $\rho/4$; on that event the retained row-2 rewards have disjoint support. The event "round-2 arm is 2 and its retained reward is 1" distinguishes the log laws with difference $\rho/4$, so $d_{\mathrm{TV}}(\mu_0, \mu_1) = \rho/4$ even with all saved coins. In the direct erased run the first reward is missing, so conditional on $B = 1$ the round-2 arm has law q. The event $(A_2, A_3) = (2, 1)$ has probability zero on X^0 and $\rho q_2 = \rho(1 - 3\eta/4)$ on X^1, hence $d_{\mathrm{TV}}(\nu_0, \nu_1) = \rho(1 - 3\eta/4)$. Proposition A.1 yields residual at least $3(1 - \eta)\rho/8 = 3\rho/16$. □

The witness can be followed by an independent fixed-schedule tail from round 4; a single erasure affects only one block, so the composed transcript retains the corresponding ρ bound and the distinguishing pair may share an identical tail. This coexists with a continuation that inherits an existing regret guarantee up to the first three rounds; it is not a minimax-rate claim.

Proposition A.3 (The omit-only doubling contrast)**.** *For the marker-unaware doubling gate of Appendix B.1, omitting the named summand can decrease the gate and extend acquisition into a later phase absent from the factual log. A rerun using the environment oracle is exact; ordinary-log exactness is not established for this implementation. Reward-independent frozen acquisition satisfies Lemma A.4.*

Proof. Keep the tape and recompute the affected phase after erasure. Its observations are in the sanitized log. If the factual gate commits but the erased gate continues, later counterfactual probes may request unobserved slices. The oracle supplies them and thereby reproduces the erased execution. Under frozen acquisition, the same probes are ingested through the same fixed endpoint in both executions, so every required slice is stored. A missing slice obstructs this direct shared-tape construction; it alone does not prove a law-level impossibility for omit-only doubling. □

A.6 Comparison with sequential reinforcement-learning unlearning

Nguyen-Tang and Arora (2026) establish exact sequential unlearning with regret and computational guarantees in tabular reinforcement learning. Table 3 compares the contracts; its middle-column references point to their paper. Their fallback resumes the learner with an environment oracle (Algorithm 3, line 11; Algorithm 2, line 3). The completion envelope supplies retained-feedback sufficiency for our specified marker response. On this paper's public interface, (4) averages conditional TV distances, whereas the category requirement controls each supported label. Appendix G.9 implements a common-bandit comparison at calendar ρ, retaining each method's feedback access and specified erased reference.

A.7 Quantitative comparison with private semi-bandits

This comparison uses pure event-level differential privacy, with one entire feedback row as the protected record and privacy budget $0 < \epsilon \leq 1$. Its two parts are a translation between guarantees and attainable private regret bounds on the product family. The translation applies to every deterministic table. For the cited regret bounds we use the independent, identically distributed row subclass of $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$; dependence among coordinates within a row is allowed. Our TV results hold under the broader conditional-mean assumption in (6). The private comparators release their served actions plus fixed continue symbols, a deterministic augmentation of the action-only interface.

A.7.1 Privacy, category conditioning, and the missing marker

Lemma A.6 (Pure DP implies category-wise TV stability)**.** *Suppose the law of the full public transcript is* ϵ*-DP for symmetric adjacency between a present row and its missing marker. Then, for every fixed row* t *and*

Aspect	Nguyen-Tang and Arora (2026)	This paper
Setting	Tabular episodic MDPs; anytime requests independent of the algorithm (Section 2, Remark 2.2).	Product semi-bandits; known horizon; gap-regular utility class.
Deletion target	Any fixed dummy replacement; equality in distribution (Definition 2.1).	Specified missing-marker response; same-tape reconstruction.
Stability and outputs	Unconditional TV on model sequences and metadata (Definition 2.2).	TV within protected action categories; public actions and continue/commit decisions.
Replay inputs	Stored tree, including the requested leaf before replacement, and environment oracle (Algorithms 2–3).	Completed sanitized factual log, marker, action metadata, and private saved randomness.
Mechanism	Gaussian binary-tree prefix sums and maximal coupling (Section 3).	Calibrated commitment gate, completion envelope, and phase resets.
Utility and work	Gap-free RL regret and expected unlearning computation (Sections 3–4).	Matching product regret up to logarithms and sufficiency of retained feedback for replay.

Table 3: Sequential unlearning contracts. The access guarantee here establishes a source for every replayed reward; it is not an end-to-end computation bound.

every supported action category $V_t = \phi(S_t)$,

$$d_{\mathrm{TV}}\Big(\mathcal{L}(A(X) \mid V_t = v), \mathcal{L}(A(X^{(-t)}) \mid V_t = v)\Big) \le \rho_\epsilon := \frac{e^\epsilon - 1}{e^\epsilon + 1} = \tanh(\epsilon/2). \tag{A.6}$$

The same statement holds for calendar TV. The constant is sharp for arbitrary pairs of ϵ-DP output laws.

Proof. Let P, Q be the two transcript laws. Pure DP in both directions gives $P(E) \le aQ(E)$ and $Q(E) \le aP(E)$ for every event E, where $a = e^\epsilon$. If $h = P(E) - Q(E) \ge 0$, applying these inequalities to E and its complement yields $Q(E) \ge h/(a-1)$ and $P(E) \le 1 - h/(a-1)$. Hence $h \le (a-1)/(a+1)$; reverse the roles of P, Q for negative differences. The case $\epsilon = 0$ follows directly. Binary laws assigning probabilities $a/(a+1)$ and $1/(a+1)$ to one outcome attain the bound.

Causality gives $P(V_t = v) = Q(V_t = v) = p(v)$: the action and its label precede the changed feedback. For $p(v) > 0$, apply the DP inequalities to $E \cap \{V_t = v\}$ and divide by the same $p(v)$. The conditional laws therefore satisfy the same two DP inequalities, and the first part of the proof applies. The argument also applies to any fixed event determined by S_t, including an overlapping protected category. □

Thus pure DP protects every such category at once. A TV-only calendar guarantee instead controls the weighted average in (4), allowing different costs when individual categories receive their own TV guarantees. For a generic conditioning event with unequal probabilities p, q, the conditional likelihood ratio gains the factor q/p; the preceding same-parameter DP argument uses the equal probabilities supplied by causality.

The adjacency convention matters. A substitution-private bandit can be extended to marked inputs by replacing $\perp$ with a fixed zero feedback vector before its usual update. This specifies its missing-feedback response without changing its regret on unmarked inputs. A present row and the imputed row then differ by one substitution, so (A.6) uses ϵ. Conversely, if a substitution is represented by a deletion followed by an insertion, two ϵ-DP edges give 2ϵ-DP and TV at most $\tanh(\epsilon)$. A TV guarantee alone gives at most 2ρ by the same two-edge argument. We use the one-edge marker extension for the private comparators. For $0 < \rho < 1$ its exact calibration is $\epsilon = \log((1+\rho)/(1-\rho))$; in particular $\rho_\epsilon = \Theta(\epsilon)$ for $0 < \epsilon \le 1$.

A.7.2 Putting the private regret bounds on the product scale

Let $m_0 = dwL = KL$ be the active arm count, so $m_0 = \Theta(m)$ by coverage. Average the w rewards in each served block. This produces $M = dL$ bounded base-arm rewards, of which d are observed each round. Let $\mu_{k,j}$ be a block's average mean, let $j_k^\star$ be its position's winner, and put $\delta_{k,j} = \mu_{k,j_k^\star} - \mu_{k,j} \ge \Delta$ for $j \ne j_k^\star$. The original regret is w times the regret for the sum of these block averages. This reduction retains one observation per selected block, which is semi-bandit feedback.

Positionwise use of Azize–Basu. Run one L-armed AdaP-UCB instance at each position, using the observed block averages and privacy parameter ϵ/d. A row can change all d feedback values, so basic composition gives ϵ-DP for the joint action transcript. For each fixed action path, the scalar input streams under two neighboring tables differ at just that row; the per-position likelihood bounds multiply to at most e^ϵ. This also verifies

Guarantee and construction	Observation term	Additional term	Status
Pure DP: positionwise AdaP-UCB	m/Δ	md/ϵ	Upper bound
Pure DP: block-aggregated CUCB-DP	md/Δ	md/ϵ	Upper bound
Calendar TV, $G = 1$	m/Δ	K/ρ	Matching bounds
Category-wise TV, $1 \le G \le L$	m/Δ	KG/ρ	Matching bounds

Table 4: Regret scales on the product family, with $d = K/w$. Each row adds its two terms and is capped by $O(C_\Delta K \Delta T)$. The DP rows are attainable bounds obtained from Azize and Basu (2022, Theorem 7) and Chen et al. (2020, Theorems 7–8), respectively; their guarantees imply every action-category TV guarantee at $\rho = \rho_\epsilon$. Logarithms are suppressed only in this table and specified below. The TV rows have lower bounds without logarithms and upper bounds with $\log(dLT)$ only on their observation term, and admit ordinary-log exact replay.

the translation from their reward-stream definition to deterministic row tables. Theorem 7 of Azize and Basu (2022), with a fixed confidence parameter $\alpha > 3$, gives

$$R_T \le w \sum_{k=1}^{d} \sum_{j \ne j_k^\star} \left\{ \frac{16\alpha \log T}{\min\{\delta_{k,j}, \epsilon/d\}} + \frac{3\alpha}{\alpha - 3} \right\} \le O\left(\frac{m \log(eT)}{\Delta} + \frac{md \log(eT)}{\epsilon} \right). \tag{A.7}$$

This is an application of their single-arm result to d observed components. Their linear-bandit bounds concern scalar total-reward feedback and are not used here.

Block-aggregated use of Chen et al. For the $M = dL$ block-average arms, the maximum action size is d, the sum reward has $B_1 = 1$, and the exact product optimizer gives approximation parameters $\alpha = \beta = 1$. Their base-arm gap is the smallest positive *action* gap among actions containing that arm. In the block-average problem this gap, denoted $\delta_i^{\min}$, is at least Δ whenever it is defined; in the original unaveraged problem the corresponding action gap is at least $w\Delta$. Their Theorem 8 has additive observation and privacy terms. Its nonlogarithmic factors after aggregation and multiplication by w are

$$w \sum_{i:\delta_i^{\min} > 0} \frac{d}{\delta_i^{\min}} \le \frac{m_0 d}{\Delta}, \qquad \frac{wMd}{\epsilon} = \frac{m_0 d}{\epsilon}. \tag{A.8}$$

The observation term carries $\log^2 T$. The privacy term carries $\log^3 T$ and a further logarithm in their problem parameters. Thus the second table row is $\widetilde{O}(md/\Delta + md/\epsilon)$, with only logarithms in $T, d, L, 1/\Delta, 1/\epsilon$ suppressed. Their CUCB-DP privacy statement is Theorem 7. Aggregating whole blocks is useful here: using its original m_0 coordinates directly would give a privacy coefficient $m_0 K/\epsilon$ instead of $m_0 d/\epsilon$.

What the comparison establishes. The DP rows in Table 4 are upper bounds on a stronger contract. The singleton specialization $d = w = K = 1$, $m = L$, already has an additive $m \log T/\epsilon$ privacy cost in the Azize–Basu results. For interior Bernoulli means with comparable gaps, their Theorems 3 and 7 match $m \log T/\Delta + m \log T/\epsilon$ up to constants under their consistency assumption. Their lower bound is asymptotic over consistent policies; our TV lower bound is a finite-horizon minimax statement on the declared gap class. The general semi-bandit lower bound of Chen et al. (2020, Theorem 9) ranges over general action families, so it is not used to assert optimality of either DP product row.

Our contribution is the matching dependence on which action categories receive a TV guarantee, from K/ρ to KL/ρ, together with the equal-entropy separation in Corollary 1. The completion envelope also supplies a retained source for every observation used in its pathwise exact reconstruction. A DP inequality controls output laws; the completion argument establishes this additional access property. Category-wise TV stability and this access property hold for the same learner in Theorem 2.

B Envelope learners: shared proof and specialisations

This appendix proves Theorem 2 and its matroid extension from one completion-envelope argument. We first define the scores and record concentration and exchange facts. Appendices B.5 and B.6 then prove gate domination, conditional stability, retained-log exactness, and utility. The final specialisations identify the uniform factual process and the frozen learner used in the experiments.

B.1 PRODUCT CLASS AND PHASE SCORES

A covered product has d sites, $L \geq 2$ blocks per site, and block width w. After common-mode reduction, $K = dw$ and $KL = \Theta(m)$. Let $X_{t,k,a} \in [0,1]$ be the average reward of block a at site k in row t, with mean $\mu_{k,a}$. For utility, use (6): the bounded reward table is exogenous, independent of the entire learner and requester tapes, and has fixed coordinate means conditional on the full past reward rows. Block averages inherit this property by linearity. Within-row dependence and time-dependent conditional distributions are allowed. Only served block averages are observed. The stability and replay arguments below hold for every bounded deterministic table, without this stochastic assumption or any gap condition.

Definition B.1 (Gap-regular product class)**.** For $0 < \Delta \leq 1$ and $C_\Delta \geq 1$, the class $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ consists of exogenous reward laws satisfying (6) and its tape-independence assumption, whose mean vectors have a unique optimal block $j_k^\star$ at each site,

$$\Delta_{\mathrm{loc}} := \min_{k, a \neq j_k^\star} (\mu_{k,j_k^\star} - \mu_{k,a}) \geq \Delta, \qquad \Delta_{\max} := \max_{S \in \mathcal{F}} \Delta(S) \leq C_\Delta K \Delta. \tag{B.1}$$

For every $0 < \Delta_0 \leq 1$, the homogeneous packing $\Theta_{\mathrm{prod}}(\Delta_0)$ of Appendix D.1 uses the two levels in (C.1) and is contained in $\mathcal{C}_{\mathrm{prod}}(\Delta_0, 1)$.

A covered product packing is generated by a swap system. The following definitions specify the geometry used by the identification floor.

Definition B.2 (Swap system)**.** A *swap system of order* (d, w, ℓ) for a reference $S^\star \in \mathcal{F}$ consists of d pairwise disjoint sites. Site k has a base block $O_k \subseteq S^\star$ with $|O_k| = w$ and ℓ mutually disjoint alternative blocks $A_k^{(1)}, \ldots, A_k^{(\ell)}$ of size w, disjoint from $S^\star$ and from all other sites, such that (S1) for every $\varepsilon \in \{\pm 1\}^d$ and all indices $j_1, \ldots, j_d$ there exist feasible $S^\pm \in \mathcal{F}$ with

$$\mathbf{1}_{S^+} - \mathbf{1}_{S^-} = \sum_{k=1}^{d} \varepsilon_k \big(\mathbf{1}_{A_k^{(j_k)}} - \mathbf{1}_{O_k} \big),$$

and (S2) $S^\star \setminus O_k \cup A_k^{(j)} \in \mathcal{F}$ for all k, j.

Definition B.3 (Swap-generated family)**.** $\mathcal{F}$ is *generated* by a swap system of order (d, w, ℓ) if it equals

$$\big\{ S^\star \setminus \textstyle\bigcup_{k \in T} O_k \cup \bigcup_{k \in T} A_k^{(j_k)} : T \subseteq [d],\ j_k \in [\ell] \big\}.$$

After common-mode reduction, a generated family satisfies $S^\star = \bigcup_k O_k$ and $K = dw$. It is *covered* when $dw\ell = \Theta(m)$.

Phases and coins. For the fixed known horizon T, partition $[T]$ into consecutive phases I_p, $p = 0, 1, \ldots$, of length 2^p, truncating the last phase at T. Before commitment, probes are fresh independent draws from a fixed, reward-independent distribution, using private coins independent of terminal and tie coins. Every reached phase is acquired in full before its terminal test. A phase is reached using only earlier phases. Its scores are discarded on continuation and are never public. At its end draw $U_p \sim \mathrm{Unif}[0,1)$ and commit exactly when $U_p < g_p$; a zero gate never fires. Commitment stops all further ingestion. Retain probe, tie-breaking, and phase coins for replay. Counterfactual replay changes neither this tape nor the calendar partition.

Raw uniform coordinates. For independent uniform block probes $A_{t,k} \sim \mathrm{Unif}([L])$, write $c_{k,a} = \sum_{t \in I_p} \mathbf{1}\{A_{t,k} = a\} X_{t,k,a}$ and $\hat{\jmath}_k = \arg\max_a c_{k,a}$ with shared tie coins. The raw certificate is

$$C(c) = -\log \sum_{k=1}^{d} \sum_{a \neq \hat{\jmath}_k} \exp(c_{k,a} - c_{k,\hat{\jmath}_k}). \tag{B.2}$$

With $n_p = |I_p|/L$, $\iota_T = \log(dLT)$, and $\beta_p^{\mathrm{raw}} = 2(\sqrt{n_p \iota_T} + \iota_T)$, the calendar gate is

$$\theta_p^{\mathrm{raw}} = 1 + 2\beta_p^{\mathrm{raw}}, \qquad s_L = \rho L/e, \qquad g_p(C) = \min\{1, s_L(C - \theta_p^{\mathrm{raw}})_+\}. \tag{B.3}$$

When all scores are in raw units, we write $\beta_p = \beta_p^{\mathrm{raw}}$. These formulas specify the factual uniform process used by E1. The erased-input rule in Appendix B.5 uses an envelope; the omit-only rule evaluates the same formula on retained sums. Those two erased-input rules are distinct. The omit-only rule remains useful for the frozen E5 implementation and the access comparison in Appendix G.4 and Proposition A.3.

B.2 MATROID SCORES AND EXCHANGE GEOMETRY

Let $\mathcal{M} = (E, \mathcal{I})$ have m elements, rank K, and basis family $\mathcal{B}$. Remove loops and contract common-mode coloops before defining these parameters. Fix a reward-independent probe law q on bases with $p_i = \Pr_{A \sim q}(i \in$

$A) > 0$, coverage $\gamma(q) = \min_i p_i$, and $\tau = 1/\gamma(q)$. For a phase of length n, let

$$z_i = \sum_{t \in I_p} \frac{\mathbf{1}\{i \in A_t\} X_{t,i}}{p_i}, \qquad \widehat{B}(z) \in \arg\max_{B \in \mathcal{B}} \sum_{i \in B} z_i. \tag{B.4}$$

Use a fixed shared priority order to select among maximizing bases. For a fixed basis B, its feasible exchanges and frozen certificate are

$$\mathcal{E}(B) = \{(a,b) : a \in B,\ b \notin B,\ B - a + b \in \mathcal{B}\}, \quad C_B(z) = -\tau \log \sum_{(a,b) \in \mathcal{E}(B)} e^{(z_b - z_a)/\tau}. \tag{B.5}$$

The empirical certificate is $C_{\mathcal{M}}(z) = C_{\widehat{B}(z)}(z)$. If $\mathcal{E}(B)$ is empty, B is the unique basis and the learner plays it without acquiring data. Otherwise the certificate is bounded above by every margin $z_a - z_b$ in its exchange list.

Definition B.4 (Exchange-gap class)**.** For a unique optimum $B^\star$, define

$$\Delta_{\mathrm{ex}} := \min_{(a,b) \in \mathcal{E}(B^\star)} (\mu_a - \mu_b), \qquad \Delta_{\max} := \max_{B \in \mathcal{B}} \Delta(B). \tag{B.6}$$

For $0 < \Delta \le 1$, the class $\mathcal{C}_{\mathcal{M}}(\Delta, C_\Delta)$ consists of exogenous reward laws with the conditional-mean and tape-independence assumptions of (6), mean vectors in $[0,1]^m$, a unique optimum, $\Delta_{\mathrm{ex}} \ge \Delta$, and $\Delta_{\max} \le C_\Delta K \Delta$. The homogeneous vector

$$\mu_i = \tfrac{1}{2}\ (i \in B^\star), \qquad \mu_i = \tfrac{1}{2} - \Delta_0\ (i \notin B^\star) \tag{B.7}$$

lies in $\mathcal{C}_{\mathcal{M}}(\Delta_0, 1)$ for $0 < \Delta_0 \le 1/2$. For $1/2 < \Delta_0 \le 1$, use means Δ_0 on $B^\star$ and zero off it.

Lemma B.1 (Single-element symmetric exchange)**.** *If B, B' are bases and $a \in B \setminus B'$, some $b \in B' \setminus B$ makes both $B - a + b$ and $B' - b + a$ bases.*

Proof. Let $H = \mathrm{cl}(B - a)$, a hyperplane, and let $C \subseteq B' + a$ be the fundamental circuit of a relative to B'. We have $a \notin H$. If $C \setminus \{a\} \subseteq H$, circuit closure would give $a \in H$. Thus some $b \in C \setminus \{a\}$ lies outside H; it belongs to $B' \setminus B$. Being outside H gives $B - a + b$ a basis, and removing b from the fundamental circuit gives $B' - b + a$ a basis. □

Lemma B.2 (One-sided exchange bijection)**.** *For bases B, B', there is a bijection $\phi : B \setminus B' \to B' \setminus B$ such that $B - a + \phi(a)$ is a basis for every $a \in B \setminus B'$.*

Proof. Join $a \in B \setminus B'$ to $b \in B' \setminus B$ exactly when $B - a + b$ is a basis. For $Y \subseteq B' \setminus B$, let $N(Y)$ be its neighbors. Each $b \in Y$ is in the closure of $(B \cap B') \cup N(Y)$, because its fundamental circuit relative to B uses no other element of $B \setminus B'$. The set $(B \cap B') \cup Y$ is independent, being contained in B'. Consequently $|B \cap B'| + |Y| \le |B \cap B'| + |N(Y)|$. Hall's condition gives a perfect matching. No simultaneous two-sided matching is used. □

Lemma B.3 (Changed basis has a small exchange margin)**.** *If $z' = z + v$ with $0 \le v_i \le \tau$ and the shared selector gives distinct bases $B = \widehat{B}(z)$ and $B' = \widehat{B}(z')$, then*

$$C_{\mathcal{M}}(z) \le \tau, \qquad C_{\mathcal{M}}(z') \le \tau. \tag{B.8}$$

The same conclusion holds with the insertion reversed.

Proof. Take a symmetric exchange (a, b) from Lemma B.1. Optimality gives $z_a - z_b \ge 0$ and $z'_b - z'_a \ge 0$, while

$$(z_a - z_b) + (z'_b - z'_a) = v_b - v_a \le \tau. \tag{B.9}$$

Both nonnegative margins are therefore at most τ, and the two certificates are bounded by these feasible exchange margins. This comparison concerns one nonnegative row insertion. Comparing its optimistic and factual completions later requires the separate 2τ reserve of Theorem B.1. □

B.3 One concentration event

Lemma B.4 (Simultaneous IPS Bernstein event)**.** *For $D \ge 2$ scored coordinates, let $\iota_T = \log(DT)$ and $\beta_p = 2(\sqrt{|I_p|\tau\iota_T} + \tau\iota_T)$. Under the exogenous conditional-mean model* (6)*, use fresh fixed-law probes with inclusion probabilities $p_i \ge 1/\tau$. Then*

$$\mathcal{E} = \left\{|z_i^{(p)} - |I_p|\mu_i| \le \beta_p \text{ for every virtual phase and coordinate}\right\}$$

has $\Pr(\mathcal{E}^c) \le 2P/(DT^2) = O((\log T)/T)$, where $P \le 1 + \lfloor \log_2(T+1) \rfloor$ is the number of phases. Virtual scores use the saved probe tape on every calendar phase, including phases not ingested after commitment; they add no actual acquisition. The coordinate bound also holds conditionally on a phase-entry history when its probe law and length are fixed by that history.

Proof. Let $\mathcal{F}_t$ contain all reward rows through time t and learner coins exposed by then, including completed terminal releases, but no future acquisition coins. The filtration is only an analysis device. Independence of the entire reward table from the tapes preserves $\mathbb{E}[X_{t,i} \mid \mathcal{F}_{t-1}] = \mu_i$. Expose a virtual fresh probe at each calendar round. For $Z_{t,i} = \mathbf{1}\{i \in A_t\}X_{t,i}/p_i$, exogeneity gives

$$\mathbb{E}[Z_{t,i} \mid \mathcal{F}_{t-1}] = \mu_i, \qquad \mathbb{E}[Z_{t,i}^2 \mid \mathcal{F}_{t-1}] \le p_i^{-1} \le \tau, \qquad |Z_{t,i} - \mu_i| \le \tau.$$

To obtain the needed conditional Bernstein inequality, let a centered increment Y satisfy $|Y| \le b$ and $\mathbb{E}[Y^2 \mid \mathcal{F}] \le v$. For $0 \le \lambda < 3/b$, expansion and $k! \ge 2\,3^{k-2}$ give

$$\mathbb{E}[e^{\lambda Y} \mid \mathcal{F}] \le 1 + \sum_{k\ge 2} \frac{\lambda^k b^{k-2} v}{k!} \le \exp\left\{\frac{\lambda^2 v}{2(1-\lambda b/3)}\right\}. \tag{B.10}$$

The same bound holds for $-Y$. Iterating it for the centered IPS increments of a phase of length n, and applying exponential Markov with $\lambda = u/(n\tau + \tau u/3)$, yields

$$\Pr\{|z_i - n\mu_i| \ge u\} \le 2\exp\{-u^2/(2n\tau + 2u\tau/3)\}. \tag{B.11}$$

This calculation also holds conditional on the entry history, including a random entry time: the remaining increments keep the same conditional means, and the length and variance budget are fixed there. For $a = \sqrt{n\tau\iota_T}$, $c = \tau\iota_T$, and $u = 2(a+c)$,

$$u^2 - 2\iota_T(2n\tau + 2u\tau/3) = (16/3)ac + (4/3)c^2 \ge 0.$$

Thus each coordinate-phase failure probability is at most $2(DT)^{-2}$; summing over DP pairs proves the event bound. No independence among coordinates, inclusions within one probe, or phase events is used. □

Corollary B.1 (Raw uniform Bernstein event)**.** *With the raw product scores above,* $|c_{k,a}^{(p)} - (|I_p|/L)\mu_{k,a}| \le \beta_p^{\mathrm{raw}}$ *simultaneously outside probability* $O((\log T)/T)$.

Proof. Apply Lemma B.4 with $D = dL$, $\tau = L$, and $z = Lc$, and divide by L. Conditional on entry to a fresh screening phase with an entry-measurable active set of size M, the same proof applies with $D = M$ and $L = M$ for $M \ge 2$. □

B.4 Ordinary-log exact unknown-gap constructions

The envelope changes the rule at a missing record: it may commit earlier, and never requests acquisition absent from the factual log. We prove this pathwise property separately from the expected gate-movement bound that gives conditional TV. The known horizon, full phase acquisition, reset after continuation, private scores, and independent half-open gate coins are those defined above. Structural claims below hold on every bounded deterministic table. Utility uses the exogenous conditional-mean model (6). Each replay removes one row from the original factual execution; the utility comparison will hold simultaneously for all such separate replays.

B.5 Envelope learner

Let q be a fixed law on product probe tuples, redrawn independently each round, with positive marginals $q_{k,a} = \Pr_q(A_k = a)$. For a fixed deterministic label $V = \varphi(A)$, define

$$\tau = \max_{k,a} q_{k,a}^{-1}, \qquad \Lambda = \max_{k,a,v:\Pr_q(V=v)>0} \frac{\Pr_q(A_k = a \mid V = v)}{q_{k,a}}. \tag{B.12}$$

Probe coordinates within a tuple may be dependent. The general proof uses these actual conditional marginals.

For Theorem 2, use a covered product and a partition of $[L]$ into $G \ge 1$ nonempty groups at every site, with group sizes $h_{k,g}$ allowed to differ. Let $\mathrm{grp}_k(a)$ name candidate a's group at site k. The concrete mixed design probes sites independently with

$$q_{k,a} = \frac{1}{2L} + \frac{1}{2G\,h_{k,\mathrm{grp}_k(a)}}, \qquad Q_{k,g} = \frac{h_{k,g}}{2L} + \frac{1}{2G}. \tag{B.13}$$

For the group tuple $V = (\mathrm{grp}_k(A_k))_{k=1}^d$, independence across sites gives $\tau \le 2L$ and $\Lambda = \max_{k,g} Q_{k,g}^{-1} \le 2G$. With equal group size $h = L/G$, $q_{k,a} = 1/L$, $\tau = L$, and $\Lambda = G$. The reciprocal group mass formula is specific to this independent design; a correlated joint probe uses (B.12). The $L = 1$ family has a single action and needs no acquisition.

Scores, certificate, and gate. Use the doubling phases of Appendix B.1, acquiring every reached phase in full with fresh probes from q. For a phase I_p of length n, form inverse-propensity scores

$$z_{k,a} = \sum_{s \in I_p} \frac{\mathbf{1}\{A_{s,k} = a\} X_{s,k,a}}{q_{k,a}}.$$

Use retained coin-based tie-breaking and let $j_k(z)$ maximise $z_{k,a}$. For a fixed candidate tuple j,

$$C_j(z) = -\tau \log \left(\sum_{k=1}^{d} \sum_{a \neq j_k} \exp\left(\frac{z_{k,a} - z_{k,j_k}}{\tau} \right) \right), \qquad C(z) = C_{j(z)}(z). \tag{B.14}$$

The temperature τ occurs both outside the logarithm and in every exponent. For $0 < \rho \leq 1$, set

$$\beta_p = 2\left(\sqrt{n\tau \log(dLT)} + \tau \log(dLT) \right), \qquad \theta_p = \tau + 2\beta_p,$$
$$g_p(c) = \min\left\{ 1, \frac{\rho}{e\Lambda}(c - \theta_p)_+ \right\}.$$

The logarithm uses the reduced block count dL. On a phase with no missing record, commit to $j(z)$ if and only if $U_p < g_p(C(z))$; after commitment ingest nothing further. A zero gate never fires.

On a table with no missing marker, the construction uses exactly the displayed factual gate. Equal groups recover the raw uniform certificate of (B.2) by the scaling $z = Lc$ and $C_\tau(z) = LC(c)$, under which the slope $\rho L/(eG)$ in raw units agrees with $\rho/(eG)$ in inverse-propensity units.

Definition B.5 (Optimistic completion)**.** Fix a phase p that contains a named time $t \in I_p$, with coin-driven probe $b_k = A_{t,k}$ and leave-one-out scores z^0 formed by omitting the observation at time t (Lemma A.1). Put $j = j(z^0)$. For $v \in [0,1]^d$ write

$$z^{(v)}_{k,a} = z^0_{k,a} + \frac{\mathbf{1}\{a = b_k\} v_k}{q_{k,b_k}}.$$

The optimistic completion is $v^\star_k = \mathbf{1}\{b_k = j_k\}$ and

$$z^+_{k,a} = z^0_{k,a} + \frac{\mathbf{1}\{a = b_k = j_k\}}{q_{k,j_k}}.$$

The envelope gate is

$$g_p^{\mathrm{env}} := g_p\big(C_j(z^+)\big). \tag{B.15}$$

Algorithm. While still probing, play the fresh coin-driven probes from q. At the end of phase p, draw the same coin $U_p \sim \mathrm{Unif}[0,1)$. If I_p contains no missing marker, commit to $j(z)$ iff $U_p < g_p(C(z))$. If I_p contains a missing marker, continue into I_{p+1} if and only if $U_p \geq g_p(C_j(z^+))$; otherwise envelope-commit to $j(z^0)$ and ingest nothing further. Phase scores remain ephemeral and not transcript-visible.

Lemma B.5 (Envelope dominates the factual gate)**.** *Fix a phase p, retained scores z^0, and the served block b_k at each site k. Use the same tie-breaking coins throughout, and let $j = j(z^0)$. Then, for every $v \in [0,1]^d$,*

$$C_j(z^+) \geq C_j(z^{(v)}).$$

Moreover

$$g_p\big(C_j(z^+)\big) = \max_{v \in [0,1]^d} g_p\big(C(z^{(v)})\big).$$

In particular, this envelope dominates the factual gate. If any completion changes a retained winner, all completion gates are zero.

Proof. For a fixed candidate j, increasing a winner score increases C_j, while increasing a loser score decreases C_j. At each site exactly one score can vary. Consequently $v^\star$ maximises C_j over the completion cube. Optimistic insertion only strengthens retained winners, so $j(z^+) = j$ and $C(z^+) = C_j(z^+)$.

Suppose some completion changes a retained winner at a site k. The served block at this site must be a retained loser: $b_k \neq j_k$. Since its increment lies in $[0, 1/q_{k,b_k}]$,

$$0 \leq z^0_{k,j_k} - z^0_{k,b_k} \leq \frac{1}{q_{k,b_k}} \leq \tau.$$

At this site the optimistic insertion is zero, and therefore

$$C_j(z^+) \leq z^+_{k,j_k} - z^+_{k,b_k} = z^0_{k,j_k} - z^0_{k,b_k} \leq \tau \leq \theta_p.$$

The first inequality follows because the corresponding exponential term occurs in the certificate sum. Every completion retaining j has empirical certificate equal to its frozen-candidate certificate, hence at most $C_j(z^+)$, and therefore has gate zero. Every completion changing a winner has a top-two margin at most τ at a changed site, and also has gate zero. This proves the assertions in this case. Notice that z^+ can still differ from z^0 at other sites; only the changed site's margin is used.

If no completion changes a retained winner, then $C(z^{(v)}) = C_j(z^{(v)})$ throughout the cube. Monotonicity of g_p and the frozen-candidate maximum show that the largest gate is attained at $v^\star$. The factual served-slice contribution is one point of $[0,1]^d$, which proves domination. □

Proposition B.1 (Minimal commitment under acquisition containment)**.** *Fix retained phase information and a missing served slice. For each admissible filling v, let $g_v \in [0,1]$ be the factual commit gate, with commitment event $\{U < g_v\}$ for a shared $U \sim \mathrm{Unif}[0,1)$, and put $g_\star = \sup_v g_v$. Let D be any erased commitment event measurable from the retained phase information, the shared coin, and optional additional randomness. Require*

$$D^c \subseteq \{U \geq g_v\} \quad \textit{almost surely for every } v.$$

Then D must contain $\{U < g_\star\}$ almost surely, and this condition is sufficient. The envelope is therefore the least erased commitment event, up to null sets, among local shared-coin rules satisfying the displayed nesting. This is not optimality over arbitrary distributional replay kernels. For every filling,

$$\Pr(D \triangle \{U < g_v\}) = \Pr(D) - g_v \geq g_\star - g_v.$$

The envelope attains equality for every filling. If all positive-gate completions commit to a common candidate, it also minimizes local continue/commit-candidate total variation against every completion among rules satisfying this containment requirement.

Proof. The containment requirement is equivalent to $\{U < g_v\} \subseteq D$ for every v. A countable sequence of gates approaching their supremum shows that D contains $\{U < g_\star\}$ up to a null set; the converse is immediate. Since factual commitment is contained in erased commitment, their disagreement probability is $\Pr(D) - g_v$, proving the displayed bound. For local phase outputs, the difference in commit probabilities is a total-variation lower bound. Committing to the common candidate with probability $g_\star$ attains it. Lemma B.5 computes this envelope at the optimistic vertex and shows that a possible winner change closes all completion gates. □

This proposition fixes the factual phase rule and its shared acquisition tape. Minimality is among rules obeying the displayed nesting, not among arbitrary distributional ordinary-log kernels.

The two-point comparison of the all-zero and all-one fillings does not dominate the factual gate when some sites hit winners and others hit losers: the certificate is not monotone along that diagonal. The implemented value is the optimistic vertex, not an enumeration of 2^d vertices.

Lemma B.6 (Optimistic frozen certificate)**.** *Let $\mathcal{H}_t$ contain the pre-phase state, all tie-breaking coins, and all other probes and consumed observations in the phase, excluding the target probe A_t and the gate coin U_p. Write $j = j(z^0)$ and z^{real} for the factual completion. Then, for every label v in the support,*

$$\mathbb{E}\left[C_j(z^+) - C_j(z^{\mathrm{real}}) \mid \mathcal{H}_t, V_t = v\right] \leq e\Lambda. \tag{B.16}$$

The difference on the left is pointwise nonnegative. Winner-changing completions have zero real and frozen-candidate gates by Lemma B.5.

Proof. Fix the bounded table. Phase reachability depends only on earlier phases, and complete acquisition exposes no target-dependent release inside the phase. Thus the target probe still has law q given $\mathcal{H}_t$. By (B.12),

$$\mathbb{E}\left[\frac{\mathbf{1}\{A_{t,k} = a\}}{q_{k,a}} \;\middle|\; \mathcal{H}_t, V_t = v\right] = \frac{\Pr_q(A_k = a \mid \varphi(A) = v)}{q_{k,a}} \leq \Lambda.$$

No independence across sites is needed. Let $\alpha_{k,a}$ be the retained odds weights of $C_j(z^0)$, so $\sum_{k,a\neq j_k} \alpha_{k,a} = 1$, fixed given $\mathcal{H}_t$.

A winner hit at site k contributes the exponent increment $-\mathbf{1}\{A_{t,k} = j_k\}/(\tau q_{k,j_k})$. Jensen's inequality gives

$$\begin{aligned} C_j(z^+) - C_j(z^0) &= -\tau \log \sum_{k,a\neq j_k} \alpha_{k,a} \exp\left(-\frac{\mathbf{1}\{A_{t,k} = j_k\}}{\tau q_{k,j_k}}\right) \\ &\leq \sum_{k,a\neq j_k} \alpha_{k,a} \frac{\mathbf{1}\{A_{t,k} = j_k\}}{q_{k,j_k}}. \end{aligned}$$

The conditional expectation of the right-hand side is at most Λ.

For the factual row, write $x_{k,a} \in [0,1]$ for the block-average potential rewards. The exponent increment

$$\delta_{k,a} = \frac{1}{\tau}\left(\frac{\mathbf{1}\{A_{t,k}=a\}x_{k,a}}{q_{k,a}} - \frac{\mathbf{1}\{A_{t,k}=j_k\}x_{k,j_k}}{q_{k,j_k}}\right)$$

lies in $[-1,1]$, because only one indicator can fire at a site and each contribution has magnitude at most τ. With $Y = \sum \alpha e^{\delta}$, the chord bound $(e^x-1)_+ \le (e-1)x_+$ on $[-1,1]$ yields

$$\big(C_j(z^0) - C_j(z^{\mathrm{real}})\big)_+ = \tau(\log Y)_+ \le (e-1)\sum_{k,a\neq j_k} \alpha_{k,a}\frac{\mathbf{1}\{A_{t,k}=a\}x_{k,a}}{q_{k,a}}.$$

The conditional expectation is at most $(e-1)\Lambda$. Adding the two displays yields (B.16). The identity is applied after the gate comparison of Lemma B.5: winner-changing completions contribute zero to phase-output TV, and the probe law is not restricted to the same-winner event. □

Lemma B.7 (Envelope-commit is log-measurable)**.** *The objects z^0, A_t, $j(z^0)$, z^+, $g_p(C_j(z^+))$, and the envelope-commit bit are functions of $\mathcal{L}_t$. They do not query $\mathcal{O}_X$.*

Proof. Probes are coin-driven, so A_t is in the log. Leave-one-out scores omit the named observation and use only retained ingested slices. The optimistic vertex is a function of A_t and of the z^0 argmax coins, hence of the log. The gate is a known Lipschitz map. □

For one factual table and learner tape, write $S_u^{(-t)}$ for the action in the separate replay erasing row t. Write $\widehat{R}_T^{(\varnothing)}$ for factual pseudo-regret and define

$$\widehat{R}_T^{(-t)} = \sum_{u=1}^{T}\Delta(S_u^{(-t)}), \qquad M_T = \max\left\{\widehat{R}_T^{(\varnothing)}, \max_{t\in[T]}\widehat{R}_T^{(-t)}\right\}. \tag{B.17}$$

These are sums of mean action gaps, including the named slot and all post-commitment rounds, as in (17); always $M_T \le T\Delta_{\max}$.

Proposition B.2 (Fixed probe and label upper)**.** *For any fixed full-cover product probe q and deterministic label φ, use* (B.12) *in the displayed IPS gate and optimistic completion. For every bounded deterministic table, this gap-agnostic learner is ρ-TV-stable conditionally on each supported label and reconstructs every separately erased execution pathwise from its ordinary retained log and saved coins. Under* (6)*, if the mean vector has a unique optimum and actual minimum gap $\Delta_{\mathrm{loc}} > 0$, then, for an absolute constant C,*

$$\mathbb{E}M_T \le C\Delta_{\max}\min\left\{T, \frac{\tau\log(dLT)}{\Delta_{\mathrm{loc}}^2} + \frac{\Lambda}{\rho\Delta_{\mathrm{loc}}}\right\}. \tag{B.18}$$

The structural claims require no gap, coverage, or stochastic assumption. The utility bound imposes neither gap regularity nor independence of coordinates within a probe tuple, and makes no minimax claim for an arbitrary label.

Proof of Theorem 2 and Proposition B.2. Ordinary-log. Couple on the learner coins, including probes, tie coins, and U_p. If the named time t lies after the eventual commit, the dummy is never ingested and Lemma A.4 applies pathwise. If t lies in a probing phase p, the dummy continues into I_{p+1} only on $\{U_p \ge g_p(C_j(z^+))\}$. Lemma B.5 gives $g_p(C_j(z^+)) \ge g_p(C(z^{\mathrm{real}}))$, so that event is contained in the factual continuation. Every later dummy probe is therefore a factual probe already stored in $\mathcal{L}_t$. The dummy ingest at time t is the missing marker, omitted from every accumulator. Envelope-commit uses $j(z^0)$, which is log-measurable by Lemma B.7. The fibre criterion of Lemma A.4 holds, with no unavailable reward query. A dummy that envelope-commits while the factual run continues does not ingest I_{p+1}, so it cannot request a missing slice.

Label-conditioned TV. Write $\mathcal{R}_t$ for the event that phase p containing t is reached. Off $\mathcal{R}_t$, the erasure is inert. On reaching it, the transcripts differ only through the phase-p symbol $Y_p \in \{\perp, j\}$. If any completion changes a retained winner, Lemma B.5 gives every completion gate zero, so the kernels agree. Otherwise the kernels are supported on $\{\perp, j(z^0)\}$ and disagree iff U_p falls between the two gates, an event of probability $g_p(C_j(z^+)) - g_p(C(z^{\mathrm{real}}))$. The gate is s-Lipschitz with $s = \rho/(e\Lambda)$, so this is at most $s(C_j(z^+) - C_j(z^{\mathrm{real}}))$. Lemma B.6 supplies the conditional expectation bound $e\Lambda$ without restricting the probe to that event. Continuation after a common $\perp$ resets both runs to the same state; after a common commit nothing further is ingested. Thus, writing D_t for shared-tape transcript disagreement, for every supported label v,

$$\Pr(D_t, V_t = v) \le \rho\Pr(\mathcal{R}_t, V_t = v) \le \rho\Pr(V_t = v).$$

This mixes the reached-phase probe law with the inert post-commitment case, whose served action need not have law q. The label agrees in both executions because the target action precedes its feedback. Dividing by its common positive probability and applying the coupling bound proves (3), hence also calendar ρ-TV.

One utility event for all erasures. Use the complete-data virtual scores of Lemma B.4 with $D = dL$, $\ell_T = \log(dLT)$, and the unchanged width $\beta_p = 2(\sqrt{|I_p|\tau\ell_T} + \tau\ell_T)$. The event $\mathcal{E}$ there has $\Pr(\mathcal{E}^c) \le \varepsilon_T := \min\{1, 2P/(dLT^2)\}$ and names no erased row. Set $\delta = \Delta_{\mathrm{loc}} \in (0, 1]$. On $\mathcal{E}$, a wrong factual winner has margin against the optimum at most $-n\delta + 2\beta_p < \theta_p$, so its gate is zero. For any target row in the phase, any retained wrong candidate j_k, and every completion v, comparison with the factual completion v^{f} gives

$$\begin{aligned}&\left[z^{(v)}_{k,j_k} - z^{(v)}_{k,j^\star_k}\right] - \left[z^{(v^{\mathrm{f}})}_{k,j_k} - z^{(v^{\mathrm{f}})}_{k,j^\star_k}\right]\\ &\qquad = \left(\mathbf{1}\{b_k = j_k\} - \mathbf{1}\{b_k = j^\star_k\}\right)\frac{v_k - v^{\mathrm{f}}_k}{q_{k,b_k}} \le \tau.\end{aligned} \tag{B.19}$$

Only one of the two site blocks is served, so there is one correction budget. In particular, $C_j(z^+) \le z^+_{k,j_k} - z^+_{k,j^\star_k} \le -n\delta + 2\beta_p + \tau < \theta_p$. No wrong envelope gate fires for any row on this same event.

Both runs acquire the full affected phase with the same probes. Envelope domination prevents erased acquisition from outlasting factual acquisition; a common continuation resets to the same state, and earlier erased commitment serves an optimum. Consequently, simultaneously for every $t \in [T]$,

$$\widehat{R}^{(-t)}_T \le \widehat{R}^{(\varnothing)}_T \quad \text{on } \mathcal{E}, \qquad M_T = \widehat{R}^{(\varnothing)}_T \quad \text{on } \mathcal{E}. \tag{B.20}$$

Sufficient phase and expectation. When $n\delta > 2\beta_p$, all empirical winners are correct and $C(z) \ge n\delta - 2\beta_p - \tau\log[d(L-1)]$. Thus the factual gate equals one once

$$n\delta \ge 4\beta_p + \tau\log[d(L-1)] + \tau + e\Lambda/\rho. \tag{B.21}$$

Writing $M = d(L-1)$ and applying $8\sqrt{n\tau\ell_T} \le n\delta/2 + 32\tau\ell_T/\delta$, a sufficient length is

$$n_{\mathrm{suff}} = \left\lceil \frac{64\tau\ell_T}{\delta^2} + \frac{16\tau\ell_T + 2\tau\log M + 2\tau + 2e\Lambda/\rho}{\delta} \right\rceil. \tag{B.22}$$

For every $0 < \delta \le 1$, this is $O(\tau\ell_T/\delta^2 + \Lambda/(\rho\delta))$. The complete doubling phases through the first sufficient phase use fewer than $4n_{\mathrm{suff}}$ rounds; if no such phase fits, the horizon cap applies. Therefore

$$\mathbb{E}M_T \le \Delta_{\max}\min\{T, 4n_{\mathrm{suff}}\} + \Delta_{\max}T\varepsilon_T.$$

Charge the complement once: even after taking the maximum it costs at most $\Delta_{\max}T$. There is no union over erasures. Together with the pathwise cap and $T\varepsilon_T = O(\log(T+1))$, this proves (B.18). Finally use $\delta \ge \Delta$, $\Delta_{\max} \le C_\Delta K\Delta$, $\tau \le 2L$, $\Lambda \le 2G$, and coverage $KL = \Theta(m)$ to obtain (16) throughout $0 < \Delta \le 1$. The logarithm multiplies identification; multiplying the forgetting duration $\Lambda/(\rho\Delta)$ by $\Delta_{\max}$ gives $O(C_\Delta KG/\rho)$. □

The environment oracle is not used. A dummy continuation is nested in a factual continuation: there is no deletion-delayed productive acquisition of a missing slice. Calendar $G = 1$ and action-conditioning $G = L$ are the endpoints of the same construction.

B.6 Matroid envelope

Let $\mathcal{M}$ be a finite loopless matroid of rank K on a ground set of size m, with basis family $\mathcal{B}$ and a reward-independent full-cover probe q of coverage $\gamma = \min_i p_i > 0$, writing $p_i = \Pr_{A\sim q}(i \in A)$ and $\tau = 1/\gamma$. Let φ be a deterministic category label of the served basis, and let Λ be the essential supremum of $\Pr(i \in A \mid \varphi(A) = v)/p_i$, the same inclusion inflation as above. Use the inverse-propensity scores and frozen-exchange certificate of Appendix B.2, with the larger reserve

$$\theta_p = 2\tau + 2\beta_p, \qquad g_p(c) = \min\left\{1, \frac{\rho}{e\Lambda}(c - \theta_p)_+\right\}.$$

On a phase containing the named time t, form leave-one-out scores z^0, put $B_0 = \widehat{B}(z^0)$, and insert the optimistic vertex $z^+_i = z^0_i + \mathbf{1}\{i \in A_t \cap B_0\}/p_i$. Commit to B_0 iff $U_p < g_p(C_{B_0}(z^+))$. Use the half-open phase coins of Appendix B.1 and the unique-basis convention following (B.5): $U_p \sim \mathrm{Unif}[0, 1)$, and an empty exchange list is the already-trivial unique basis.

Theorem B.1 (Matroid Envelope learner)**.** *The learner just described uses the known horizon and no gap parameter. For every bounded deterministic table it is ρ-TV-stable conditionally on $\varphi(A_t)$ and reconstructs every separately erased execution pathwise from the ordinary retained log and saved coins, in the sense of Definition A.3. Under* (6)*, on $\mathcal{C}_{\mathcal{M}}(\Delta, C_\Delta)$, $0 < \Delta \le 1$, it satisfies*

$$\mathbb{E}M_T \le CC_\Delta \min\left\{K\Delta T, \frac{K\log(mT)}{\gamma\Delta} + \frac{K\Lambda}{\rho}\right\},$$

where M_T includes the factual run and all separate single erasures as in (B.17)*. The structural claims do not require unique optimality or gaps. On $U_{K,m}$ and spanning trees of K_n, $n \geq 3$, calendar category label has $\Lambda = 1$ and $\gamma = K/m$, giving $\widetilde{O}(\min\{K\Delta T, m/\Delta + K/\rho\})$. No matching lower is claimed for these families, an arbitrary matroid, or an arbitrary label.*

Proof. Deterministic envelope. Optimistic insertion raises only coordinates of B_0, so B_0 remains a maximiser. For every frozen exchange (a, b) with $a \in B_0$ and $b \notin B_0$, the optimistic vertex maximises z_a and minimises z_b, hence maximises C_{B_0}. If some completion selects $B_1 \neq B_0$, the symmetric exchange property supplies $a \in B_0 \setminus B_1$ and $b \in B_1 \setminus B_0$ with both one-element swaps feasible. Optimality at the two endpoints forces $0 \leq z^0_a - z^0_b \leq \tau$. Optimism can raise a by at most τ and leaves b unchanged, so $C_{B_0}(z^+) \leq z^+_a - z^+_b \leq 2\tau \leq \theta_p$. Every completion retaining B_0 then has gate zero, and a completion selecting a different basis has empirical-candidate certificate at most τ by Lemma B.3, applied to the nonnegative insertion from the retained scores; see (B.8). Thus the optimistic gate equals the maximum empirical gate over the completion box, and dominates the factual gate.

Surplus. Condition on all other phase probes, rows, and tie coins, excluding the target basis and phase coin. The retained exchange weights $\alpha_{a,b} = e^{(z^0_b - z^0_a)/\tau} / \sum_{(i,j)\in\mathcal{E}(B_0)} e^{(z^0_j - z^0_i)/\tau}$ are fixed. If the disclosed label is v, then $\mathbb{E}[\mathbf{1}\{i \in A_t\}/p_i \mid v] \leq \Lambda$ for every coordinate. For the factual completion z^x, its exponent increment is

$$\delta_{a,b} = \frac{\mathbf{1}\{b \in A_t\}x_b/p_b - \mathbf{1}\{a \in A_t\}x_a/p_a}{\tau} \in [-1, 1].$$

Both endpoints may be served, but their nonnegative increments each lie in $[0, \tau]$, so their difference still lies in $[-\tau, \tau]$. Jensen's inequality and the exponential chord bound, respectively, give

$$C_{B_0}(z^+) - C_{B_0}(z^0) \leq \sum_{(a,b)} \alpha_{a,b} \frac{\mathbf{1}\{a \in A_t\}}{p_a},$$

$$\big(C_{B_0}(z^0) - C_{B_0}(z^x)\big)_+ \leq (e-1) \sum_{(a,b)} \alpha_{a,b} \frac{\mathbf{1}\{b \in A_t\}x_b}{p_b}.$$

Thus the nonnegative envelope surplus has conditional expectation at most $\Lambda + (e-1)\Lambda = e\Lambda$. No independence among the inclusion indicators, and no conditioning on basis agreement, is used.

TV and access. A changed candidate has zero gate by the deterministic envelope argument. Otherwise the common coin gives phase-output TV equal to the gate difference, at most $\rho/(e\Lambda)$ times the frozen surplus. Average over the common conditional probe law and apply the preceding bound. Scores are discarded on continuation, so the remaining transcript is a common kernel. Mix the reached-phase event with the inert post-commitment case exactly as in the product proof to obtain conditional TV at most ρ. Envelope domination gives $\{U_p \geq g^{\mathrm{env}}_p\} \subseteq \{U_p \geq g^{\mathrm{real}}_p\}$. The erased process therefore uses only retained factual acquisition slices; when it commits earlier, the obsolete factual suffix is discarded. The same replay argument as for Theorem 2 gives exactness with no unavailable reward query.

Factual utility. Use the single virtual-phase event $\mathcal{E}$ of Lemma B.4 with $D = m$ under (6). If an empirical basis $B \neq B^\star$ were selected, Lemma B.1 supplies an exchange (a, b) feasible out of B and its reverse feasible out of $B^\star$. Hence $\mu_a - \mu_b \leq -\Delta_{\mathrm{ex}}$ and $C_B(z) \leq z_a - z_b \leq -n\Delta_{\mathrm{ex}} + 2\beta_p < \theta_p$: no wrong gate fires. For identification, Lemma B.2 pairs $B^\star \setminus B$ to $B \setminus B^\star$ by feasible exchanges out of $B^\star$. If this difference has size r, then

$$\sum_{i\in B} z_i - \sum_{i\in B^\star} z_i \leq -r(n\Delta_{\mathrm{ex}} - 2\beta_p).$$

Consequently $n\Delta_{\mathrm{ex}} > 2\beta_p$ makes $B^\star$ the unique empirical maximizer. Every exchange out of $B^\star$ then has empirical margin at least $n\Delta_{\mathrm{ex}} - 2\beta_p$, so

$$C_{B^\star}(z) \geq n\Delta_{\mathrm{ex}} - 2\beta_p - \tau \log |\mathcal{E}(B^\star)|.$$

Because $|\mathcal{E}(B^\star)| \leq K(m-K)$, the gate equals one once

$$n\Delta_{\mathrm{ex}} \geq 4\beta_p + \tau \log(K(m-K)) + 2\tau + e\Lambda/\rho. \tag{B.23}$$

A sufficiently large absolute constant makes this true at $n = \widetilde{O}(\tau/\Delta^2_{\mathrm{ex}} + \Lambda/(\rho\Delta_{\mathrm{ex}}))$. This follows by absorbing the square-root term with $8\sqrt{n\tau \log(mT)} \leq n\Delta_{\mathrm{ex}}/2 + 32\tau \log(mT)/\Delta_{\mathrm{ex}}$ and using $0 < \Delta_{\mathrm{ex}} \leq 1$ for the remaining linear terms. The total length through the first sufficient doubling phase is at most a constant times this budget.

Simultaneous erased utility and the horizon branch. For every row in a phase and every wrong retained basis B_0, choose its symmetric exchange (a, b) with $B^\star$ as above. Relative to the factual completion, any completion

changes each endpoint score by at most τ, and can increase $z_a - z_b$ by at most 2τ because both endpoints may be served. In particular,

$$C_{B_0}(z^+) \le -n\Delta_{\mathrm{ex}} + 2\beta_p + 2\tau < \theta_p.$$

Thus all wrong envelope gates vanish on the same $\mathcal{E}$, simultaneously for every row. Envelope domination and phase reset give $\widehat{R}_T^{(-t)} \le \widehat{R}_T^{(\varnothing)}$ for all t on $\mathcal{E}$, so $M_T = \widehat{R}_T^{(\varnothing)}$ there. Charge acquisition by $\Delta_{\max}$ per round and the complement only once by $\Delta_{\max} T \min\{1, 2P/(mT^2)\}$. The sufficient phase calculation and the pathwise horizon cap yield the actual-gap upper

$$\mathbb{E}M_T \le C\Delta_{\max} \min\left\{T, \frac{\log(mT)}{\gamma(q)\Delta_{\mathrm{ex}}^2} + \frac{\Lambda}{\rho\Delta_{\mathrm{ex}}}\right\}. \tag{B.24}$$

This bound requires a unique optimum and positive actual exchange gap, but no gap regularity. For $\mathcal{C}_{\mathcal{M}}(\Delta, C_\Delta)$ substitute $\Delta_{\mathrm{ex}} \ge \Delta$ and $\Delta_{\max} \le C_\Delta K\Delta$. The stated special cases use the coverage calculations in Corollaries B.2 and B.3. □

A product site serves exactly one block, so the two endpoint edits cannot occur together and the product reserve $\tau + 2\beta_p$ of Theorem 2 remains valid. The matroid argument must not copy that one-unit product budget.

B.7 Uniform factual law and frozen replay

Proposition B.3 (Uniform product regret)**.** *The complete-feedback uniform doubling process* (B.2)–(B.3) *uses no gap parameter and on $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ has*

$$R_T = \widetilde{O}\big(\min\{K\Delta T, m/\Delta + K/\rho\}\big). \tag{B.25}$$

Proof. Set $G = 1$ in Theorem 2. The probe is uniform and the exact coordinate change $z = Lc$, $\tau = L$, $\beta_p = L\beta_p^{\mathrm{raw}}$ gives $C_{\mathrm{IPS}}(z) = LC(c)$. Thus the factual gates, coins, and action law agree. This equality of factual laws does not identify their erased-input rules. □

Lemma B.8 (Raw fixed-winner sensitivity)**.** *Fix retained raw scores c^0, a retained winner tuple j, and a target row with block contributions $x_{k,a} \in [0,1]$. Define $C_j(c) = -\log \sum_{k, a \ne j_k} e^{c_{k,a} - c_{k,j_k}}$. Condition on all other phase probes, observations, and tie coins, excluding the phase coin and target probe. Under a uniform target probe A_t,*

$$\mathbb{E}|C_j(c^x) - C_j(c^0)| \le e/L. \tag{B.26}$$

More generally, after conditioning on a target-action label, the bound is $e\pi$ if every candidate's conditional inclusion probability is at most π. The expectation is taken before any restriction to the same-winner event.

Proof. Normalize the retained exponential terms to weights $\alpha_{k,a}$ summing to one. Insertion changes a term's exponent by $\delta_{k,a} = \mathbf{1}\{A_{t,k} = a\}x_{k,a} - \mathbf{1}\{A_{t,k} = j_k\}x_{k,j_k} \in [-1,1]$. Write $Y = \sum \alpha e^{\delta}$. The chord bound and Jensen give

$$(\log Y)_+ \le (e-1)\sum \alpha(\delta)_+, \qquad (-\log Y)_+ \le \sum \alpha(-\delta)_+.$$

Each conditional expectation of a positive or negative increment is at most π, so the total is at most $e\pi$. Uniform probing gives $\pi = 1/L$. □

Lemma B.9 (One-row top-two margin)**.** *If two vectors differ at one coordinate by at most one, their top-two margins differ by at most one. If their shared coin-tied maximizers differ, both margins are at most one.*

Proof. Hold all unchanged coordinates fixed, and let $a \ge b$ be their largest two values. As a function of the changed coordinate x, the top-two margin is $a - b$ for $x \le b$, $a - x$ for $b \le x \le a$, and $x - a$ for $x \ge a$. This continuous piecewise-linear function is 1-Lipschitz. A winner change crosses $x = a$, where the margin is zero, and both endpoints lie within distance one. For two coordinates the same formula uses the sole unchanged value a and the margin is $|x - a|$. □

Lemma B.10 (Changed product winner closes the raw gates)**.** *A single served-row insertion or deletion that changes a retained product winner makes both empirical certificates at most one. Hence both raw gates are zero.*

Proof. At a changed site only one coordinate moves by at most one. Its top-two margin bounds the joint certificate from above, so Lemma B.9 applies. □

Proposition B.4 (Raw omit-only calendar stability)**.** *The raw omit-only rule with gate* (B.3) *is calendar ρ-TV-stable. It is not asserted to be ordinary-log exact under doubling.*

Proof. At the affected phase, a winner change closes both gates by Lemma B.10. Otherwise the shared phase coin gives output TV equal to the gate difference, bounded by $(\rho L/e)|C_j(c^x) - C_j(c^0)|$. Apply Lemma B.8 over the full probe law, without conditioning on winner agreement. The public probe has the same marginal in both worlds, and

$$d_{\mathrm{TV}}(P_A K, P_A K') = \mathbb{E}_{A\sim P_A} d_{\mathrm{TV}}(K_A, K'_A). \tag{B.27}$$

Thus public visibility of the probe preserves the averaged bound ρ. After a common failed gate, scores have been discarded and future evolution is a common kernel; after a common commit, neither run ingests again. Data processing proves transcript stability. Access for differing continuation symbols is the separate issue in Proposition A.3. □

Proposition B.5 (Known-gap ordinary-log calendar match)**.** *For a known class parameter Δ, choose a single deterministic uniform-probe horizon sufficient for*

$$(N/L)\Delta \geq 4\beta^{\mathrm{raw}} + \log[d(L-1)] + 2 + e/(\rho L). \tag{B.28}$$

Such a length is

$$N = \widetilde{O}(L/\Delta^2 + 1/(\rho\Delta)). \tag{B.29}$$

Run the raw omit-only gate once, then ingest nothing further; on a failed gate play a fixed, data-independent fallback action. If $N > T$, acquire throughout the horizon. This frozen learner is ordinary-log exact and calendar ρ-TV-stable on every bounded table. On $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$, the expected maximum $\mathbb{E}M_T$ of (B.17) *has rate* (B.25)*.*

Proof. The factual and erased executions use exactly the same fixed acquisition slices, with the named one replaced by the missing marker, so retained sums and the shared coin reconstruct the erased policy. Stability is the one-phase proof of Proposition B.4. On the event of Corollary B.1, a wrong empirical winner has certificate at most $2\beta^{\mathrm{raw}}$, while a correct tuple has certificate at least $(N/L)\Delta - 2\beta^{\mathrm{raw}} - \log[d(L-1)]$. For every target row on this same event, the extra unit in the saturation condition covers deletion of its served contribution: the wrong-vs-true margin increases by at most one and the correct certificate decreases by at most one. The threshold $1 + 2\beta^{\mathrm{raw}}$ prevents a wrong commit, and the strengthened sufficient length makes the correct erased gate equal one. This sufficient length specifies the fixed-acquisition rule. These constant slacks do not change (B.29). All separate erasures obey this argument on the same event. Charge acquisition by $\Delta_{\max} N$ and the maximum on failure once by $\Delta_{\max} T \Pr(\mathcal{E}^c)$; use the pathwise cap $\Delta_{\max} T$. □

Indexed ordinary-log replay.

Proposition B.6 (Local reconstruction from a phase-addressable log)**.** *For one erasure from the factual product envelope execution, suppose the retained log supports direct access to a calendar phase and the factual public transcript is available for reuse. If the named row was acquired in a phase of length n, reconstructing the compact erased transcript requires rebuilding scores only in that phase. In unit-cost arithmetic this takes $O(nK + dL)$ operations from raw served slices, where $K = dw$, or $O(nd + dL)$ from retained block averages. Given deletion-safe phase sums, the candidate, optimistic completion, certificate, and gate take $O(dL)$ operations. An unacquired named row requires no reward reads.*

Proof. The factual prefix through the affected phase's probes is unchanged. Form z^0 from zero by reading that phase's retained slices, and use Lemma B.7 to decode its erased boundary symbol. If both runs continue, their discarded scores leave the same reset state, so the factual suffix can be reused. If both commit, Lemma B.5 implies the same candidate: a completion changing a retained winner has zero gate and cannot commit. Envelope domination excludes factual commitment followed by erased continuation. The remaining case is earlier erased commitment, for which the compact output consists of the common prefix and the constant policy $j(z^0)$ after this boundary; later factual acquisition is omitted. Thus any transcript change starts with an earlier envelope commitment, and no other phase needs score reconstruction. Averaging each retained raw slice costs $O(K)$, adding its d block contributions costs $O(d)$, and initializing and decoding the dL scores costs $O(dL)$. These operations use retained values only. □

The compact representation uses transcript and log views with a phase boundary and, when needed, a replacement policy. Materializing a full action transcript adds its output size, $\Theta(Td)$ for block tuples or $\Theta(TK)$ for explicit arm lists; copying a retained log likewise adds its payload size. Phase offsets, saved coins, and the factual transcript are maintained inputs to this accounting. With per-phase, per-block sum trees, a deletion overwrites the named payload and leaf, then recomputes ancestors from their retained children. Bottom-up repair makes each queried child deletion-safe before its parent is used, and overwrites the affected cached total. The update costs $O(K + d\log n)$ before the $O(dL)$ decoder; tree construction and storage are additional costs. Such trees can be built in $O(NK + dLP)$ operations and $O(Nd + dLP)$ auxiliary words for N acquired rows in P phases, using a linear routing pass, in addition to raw-log storage. The frozen-product host measurements in Section G.3 report their own construction, storage, and update costs; the envelope bounds here are operation

counts. The deterministic checks in `run_replay.py` compare local reconstruction with full marked-table runs on earlier commitment, shared continuation, shared commitment, and inert erasure.

B.8 Coverage and matroid special cases

Define the maximin coverage

$$\gamma^\star(\mathcal{M}) = \max_{q\in\Delta(\mathcal{B})} \min_i \Pr_{A\sim q}(i \in A). \tag{B.30}$$

Proposition B.7 (Coverage characterisation)**.** *If $P_{\mathcal{B}}$ is the base polytope, then $\gamma^\star = \max\{t : \exists x \in P_{\mathcal{B}},\ x_i \geq t\ \forall i\}$. Moreover $\gamma^\star = K/m$ exactly when $(K/m)\mathbf{1} \in P_{\mathcal{B}}$, equivalently $r(A) \geq (K/m)|A|$ for every $A \subseteq E$.*

Proof. Base distributions have precisely the marginals in the base polytope. Their coordinates sum to K, so the only possible marginal vector with every coordinate at least K/m is $(K/m)\mathbf{1}$. Substituting this vector into the base-polytope constraints $x(A) \leq r(A)$ proves the equivalence. □

Corollary B.2 (Uniform matroids)**.** *Uniform probing on $U_{K,m}$ has $p_i = K/m = \gamma^\star$. Thus the calendar Envelope learner of Theorem B.1 attains*

$$\widetilde{O}(\min\{K\Delta T, m/\Delta + K/\rho\}) \tag{B.31}$$

on $\mathcal{C}_{\mathcal{M}}(\Delta, C_\Delta)$.

Proof. Symmetry gives the uniform marginal K/m; sum of marginals bounds any minimum by K/m. Substitute into Theorem B.1 with calendar category label. □

Corollary B.3 (Complete-graph spanning trees)**.** *For spanning trees of K_n, $n \geq 3$, we have $K = n-1$, $m = \binom{n}{2}$, and uniform spanning-tree probing gives $p_i = 2/n = K/m = \gamma^\star$. Calendar Envelope learner attains* (B.31)*.*

Proof. Edge transitivity makes every edge marginal equal, and their sum is $n-1$. Apply Theorem B.1. □

Unequal partition parts of sizes 3 and 2 have $\gamma^\star = 1/3 < 2/5 = K/m$. The arbitrary-matroid upper therefore retains $\gamma(q)$; no matching lower is claimed for an arbitrary matroid. In a product the reduced matroid has one choice per site. Scaling $z = Lc$ identifies its certificate with $LC(c)$, but the general matroid envelope uses reserve $2\tau + 2\beta$; its gate is not literally the product gate with the sharper one-unit reserve.

C Category lower bounds and equal-entropy comparisons

The matching group law combines a causal lower bound, proved here for every sitewise partition with a common group count, with Theorem 2. We then compare category labels whose labels have the same entropy under uniform actions. Their conditional inclusion geometry, rather than that entropy, determines the sufficient forgetting cost.

Hard two-level family. Fix $0 < \Delta \leq 1$ and define winner and loser means by

$$(p, q) = \begin{cases} (1/2, 1/2 - \Delta), & 0 < \Delta \leq 1/2, \\ (\Delta, 0), & 1/2 < \Delta \leq 1. \end{cases} \tag{C.1}$$

At each site one winner has mean p and every other block has mean q. Each width-w block clones one Bernoulli variable, independently across blocks and rows. Write $\Theta_{\mathrm{prod}}(\Delta)$ for this family; it has every local gap equal to Δ and $\Delta_{\max} = K\Delta$. Fixing the other-site winners, let P_j make block j the winner at a distinguished site k. The auxiliary reference P_0 gives every block there mean q and need not belong to the class. All these witnesses are admissible under (6).

C.1 A causal group lower bound

Fix a covered product of order (d, w, ℓ), with $L = \ell + 1 \geq 2$ and $K = dw$. At each site partition the L candidates into G nonempty groups; group sizes may differ. Write V_t for the disclosed group tuple and $G_{t,k}$ for its site-k entry. Let $N_j(n)$ count visits to candidate j at a fixed site, and $N_j^G(n)$ count visits to the group containing it. The action and its label are chosen before seeing row t, so a change to that row preserves the label marginal. For equal group sizes h, the identities

$$\sum_{j=1}^{L} N_j(n) = n, \qquad \sum_{j=1}^{L} N_j^G(n) = hn \tag{C.2}$$

hold pathwise for every causal learner.

Lemma C.1 (Group-indexed contraction)**.** *Let $\mathcal{F}$ be generated by a swap system of order (d, w, ℓ) and fix a site k. Let P_j and the auxiliary reference P_0 be the site-k environments defined above. For every causal learner that is ρ-TV-stable conditionally on the group tuple V_t of a sitewise partition, in the sense of* (3)*, every $j \in [L]$, and every event E measurable with respect to the first n actions,*

$$|P_j(E) - P_0(E)| \le 2\rho\Delta\mathbb{E}_0[N_j^G(n)]. \tag{C.3}$$

Proof. Localization. Let X and X' be any two tables that agree except in row t, where they differ only in the coordinates of block j. Couple the coins. A generated action selects exactly one block at site k. If $G_{t,k} \neq G(j)$, the consumed slice at time t misses block j entirely; the two conditional laws given $V_t = v$ then coincide for every such label. On $\{G_{t,k} = G(j)\}$, factor the replacement through the dummy $\perp$ and apply (3) twice, giving 2ρ for each such v. Writing μ_v and μ'_v for the two transcript laws conditioned on $V_t = v$,

$$d_{\mathrm{TV}}(A(X), A(X')) = \sum_v \mathbb{P}(V_t = v) d_{\mathrm{TV}}(\mu_v, \mu'_v) \le 2\rho\mathbb{P}(G_{t,k} = G(j)). \tag{C.4}$$

The bound holds at every surrounding table, which is what lets it be chained. The localization does not use equal group sizes.

Coupling. Block j is one Bernoulli replicated w times, with means q under P_0 and p under P_j from (C.1); no other block differs. The row laws therefore have $d_{\mathrm{TV}} = p - q = \Delta$. Draw X from P_0 and independent $F_1, \dots, F_n$ with $\mathbb{P}(F_s = 1) = \Delta/(1-q) \le 1$, and let $R := \{s : \text{block } j \text{ reads 0 in row } s,\ F_s = 1\}$. Raising block j to 1 on R produces a table distributed as P_j, and $\mathbb{P}(s \in R) = \Delta$ independently across rows.

Reverse-calendar telescoping. For $1 \le s \le n$, let Z_s be X with block j raised on $R \cap [s, n-1]$. Thus $Z_n = X$, and Z_1 has the P_j law as far as the first n actions are concerned. The triangle inequality bounds the prefix TV by the sum of the one-row replacement distances. At the replacement of row s, all earlier rows still have the reference law; causality therefore makes the action's group occupancy have its P_0 law. The indicator $\mathbf{1}\{s \in R\}$ depends only on row s and its independent raising coin, so it is independent of those earlier rows and the learner's coins. Averaging (C.4), each term is at most $2\rho\Delta \Pr_0(G_{s,k} = G(j))$. Summing over $s < n$ and bounding by the count through n proves (C.3). This argument uses the guarantee for each supported group tuple, not just its average over labels. □

Proposition C.1 (Floor under a G-group partition)**.** *Let $\mathcal{F}$ be generated by a covered swap system of order (d, w, ℓ), and fix a sitewise partition of $[L]$ into $G \ge 1$ nonempty groups, the same G at every site, with group sizes allowed to differ. For $0 < \Delta \le 1$, $0 < \rho \le 1$, and $T \ge 1$, every causal learner that is ρ-TV-stable conditionally on the group tuple V_t, in the sense of* (3)*, has*

$$\sup_{\theta \in \Theta_{\mathrm{prod}}(\Delta)} R_T(A, \theta) = \Omega\big(\min\{dw\Delta T, dwL/\Delta + dwG/\rho\}\big), \tag{C.5}$$

which under coverage is $\Omega(\min\{K\Delta T, m/\Delta + KG/\rho\})$. The argument does not assume that the learner probes uniformly. When $G = 1$ this is the calendar floor of Corollary D.1. For equal group sizes, the grain statement below follows by $G = L/h$.

Proof. If $G = 1$, the label V_t is constant and (3) is calendar ρ-TV. Corollary D.1 gives (C.5). Henceforth $G \ge 2$. Fix a site k, with the winners at all other sites frozen. Choose one representative block from each of the G groups at site k, and write $\mathcal{R}$ for that set of G representatives. Under P_a for $a \in \mathcal{R}$, block a is the unique site-k winner at mean p, and every other site-k block has mean q from (C.1). These G environments lie in $\Theta_{\mathrm{prod}}(\Delta)$. Write $M_g(n)$ for the number of the first n rounds whose site-k group equals g. A generated action selects exactly one block, hence exactly one group, so

$$\sum_{a \in \mathcal{R}} M_{G(a)}(n) = n \tag{C.6}$$

pathwise, for every causal policy. Lemma C.1 applies to this (possibly unbalanced) partition and yields $d_{\mathrm{TV}}(P_0^n, P_a^n) \le 2\rho\Delta\mathbb{E}_0[M_{G(a)}(n)]$. Put $n := \min\{T, \lfloor G/(8\rho\Delta)\rfloor\}$. If $n = 0$, then $\rho\Delta > G/8$, and $KG/\rho = \Theta(dwG/\rho) = O(dw\Delta)$. The first action is a function of coins alone, and averaging over the uniform choice of representative winner independently at every site gives $R_1 \ge dw\Delta(1 - 1/G) \ge dw\Delta/2 = \Omega(KG/\rho)$. Henceforth $n \ge 1$.

Since $0 \le N_a(n) \le n$, the contraction gives $\mathbb{E}_a[N_a(n)] \le \mathbb{E}_0[N_a(n)] + n d_{\mathrm{TV}}(P_0^n, P_a^n) \le \mathbb{E}_0[N_a(n)] + 2\rho\Delta n\mathbb{E}_0[M_{G(a)}(n)]$. Averaging over $a \in \mathcal{R}$ and using (C.6) together with $\sum_{a \in \mathcal{R}} N_a(n) \le n$, $n \le G/(8\rho\Delta)$, and $G \ge 2$ (so $n/G \le n/2$),

$$\frac{1}{G}\sum_{a \in \mathcal{R}} \mathbb{E}_a[N_a(n)] \le \frac{n}{G} + \frac{2\rho\Delta n^2}{G} \le \frac{n}{2} + \frac{n}{4} = \frac{3n}{4}.$$

Under P_a every round that does not select block a incurs site regret $w\Delta$, so the Bayes site-k regret over the uniform choice of representative is at least $w\Delta n/4$. If $T \le G/(8\rho\Delta)$ this is $w\Delta T/4$; otherwise $n \ge G/(16\rho\Delta)$ and it is at least $wG/(64\rho)$. Hence the Bayes site-k regret is $\Omega\big(\min\{w\Delta T, wG/\rho\}\big)$ for every frozen configuration of the other sites. Averaging the remaining winners and summing the d sites, the representative-product Bayes regret, and therefore the supremum, is $\Omega\big(\min\{dw\Delta T, dwG/\rho\}\big)$.

Lemma D.1 uses no stability and an independent uniform all-L winner prior, giving $\Omega\big(\min\{dw\Delta T, dwL/\Delta\}\big)$. For nonnegative u, v, w, $\max\{\min\{u,v\}, \min\{u,w\}\} \ge \frac{1}{2}\min\{u, v+w\}$, so the two branches combine to Proposition C.1 with $dwL = \Theta(m)$ under coverage. The identification and representative-group priors are separate lower bounds on the same worst-case regret. □

Corollary C.1 (Equal-size groups)**.** *For $0 < \Delta \le 1$ and $0 < \rho \le 1$, if every group has size $h = L/G$, every causal grain-h stable learner satisfies*

$$\sup_{\theta \in \Theta_{\mathrm{prod}}(\Delta)} R_T = \Omega(\min\{K\Delta T, m/\Delta + m/(\rho h)\}). \tag{C.7}$$

Proof. Substitute $G = L/h$ in Proposition C.1 and use $KL = \Theta(m)$. The counting identity (C.2) also gives the same contraction bound directly. Neither proof assumes uniform probing. □

C.2 Attainment without a gap parameter

Corollary C.2 (Group-partition attainment)**.** *For $0 < \Delta \le 1$ and $0 < \rho \le 1$, on $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ the learner of Theorem 2 is group-conditionally ρ-TV-stable, ordinary-log exact, gap-agnostic, and satisfies*

$$\mathbb{E}M_T = \widetilde{O}(\min\{K\Delta T, m/\Delta + KG/\rho\}) \tag{C.8}$$

for the expected maximum over the factual run and all separate single-erasure replays in (B.17)*, including unequal groups. Together with Proposition C.1, this is the product law of Theorem 1.*

Proof. The mixed probe (B.13) gives coverage at least $1/(2L)$ and group inclusion inflation at most $2G$. These are exactly the two quantities used in the proof of Theorem 2. □

Corollary C.3 (Equal-group attainment)**.** *For $0 < \Delta \le 1$, $0 < \rho \le 1$, and $h \mid L$, on a covered class $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ with fixed C_Δ, the equal-group learner has the simultaneous factual and single-erasure rate*

$$\mathbb{E}M_T = \widetilde{O}(\min\{K\Delta T, m/\Delta + m/(\rho h)\}), \tag{C.9}$$

is ordinary-log exact, and is ρ-TV-stable conditionally on the group tuple.

Proof. Set $G = L/h$ in Theorem 2. Now $q_{k,a} = 1/L$, $\tau = L$, $\Lambda = G$, and $z = Lc$. Its factual gate in raw units therefore has slope

$$s_h = \rho h/e \tag{C.10}$$

and the raw certificate and confidence reserve of (B.2)–(B.3). This is an exact change of coordinates in the factual law. On a missing record, use the envelope branch proved in Appendix B.5. □

Proposition C.2 (Known-gap frozen grain learner)**.** *On a covered class $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ with fixed C_Δ, known $0 < \Delta \le 1$, $0 < \rho \le 1$, and $h \mid L$, the frozen raw learner with slope* (C.10) *and a sufficient deterministic length*

$$N = \widetilde{O}(L/\Delta^2 + L/(\rho h\Delta)) \tag{C.11}$$

is ordinary-log exact, grain-h ρ-TV-stable, and attains the expected maximum bound (C.9)*. It ingests nothing after its single gate and uses a fixed fallback action if that gate fails. If $N > T$, it acquires throughout the horizon.*

Proof. The exact saturation condition in raw units is

$$(N/L)\Delta \ge 4\beta^{\mathrm{raw}} + \log[d(L-1)] + 2 + e/(\rho h), \tag{C.12}$$

including the one-unit slack that makes the erased gate saturate as well. The conditional version of Lemma B.8 has $\pi = 1/h$, so raw omit-only gate movement averages to at most $(\rho h/e)(e/h) = \rho$. The fixed schedule supplies all retained slices. Concentration, the one-unit row correction, failure charge, and horizon cap are exactly the frozen proof of Proposition B.5, with $e/(\rho L)$ replaced by $e/(\rho h)$. □

The mixed probe balances two costs: candidate coverage remains at least $1/(2L)$, while every disclosed group has mass at least $1/(2G)$. Substituting the smallest group size into a uniform-probe bound would miss this tradeoff. For groups of sizes 1 and $L-1$, the mixed design retains an $O(K/\rho)$ forgetting term. The half-mixture is not claimed to be constant-optimal among all probe designs.

C.3 A SUFFICIENT BOUND FOR ANY FIXED CATEGORY LABEL

Proposition C.3 (Fixed-probe category label upper)**.** *Let q be any fixed law on product actions, redrawn independently each round, with positive candidate marginals $q_{k,a} = \Pr_q(A_k = a)$. Let φ be a fixed deterministic action label, and set*

$$\tau = \max_{k,a} q_{k,a}^{-1}, \qquad \Lambda = \max_{k,a,v:\Pr_q(\varphi(A)=v)>0} \frac{\Pr_q(A_k = a \mid \varphi(A) = v)}{q_{k,a}}. \tag{C.13}$$

The product completion-envelope learner with these IPS scales is ordinary-log exact, gap-agnostic, and ρ-TV-stable conditionally on every supported label. On $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$, for $0 < \Delta \le 1$ and $0 < \rho \le 1$, its simultaneous regret obeys

$$\mathbb{E}M_T = \widetilde{O}\big(C_\Delta \min\{K\Delta T, K\tau/\Delta + K\Lambda/\rho\}\big), \tag{C.14}$$

where M_T is defined in (B.17)*. The actual-gap bound* (B.18) *does not require gap regularity. The coordinates of the probe tuple need not be independent.*

Proof. Apply Proposition B.2 with the actual conditional marginal ratio (C.13), which equals (B.12). The deterministic completion comparison uses one served block per site. Conditional inclusion expectations bound the two frozen-certificate movement terms by Λ and $(e-1)\Lambda$; no independence among positions is used. Substitute $\Delta_{\mathrm{loc}} \ge \Delta$ and $\Delta_{\max} \le C_\Delta K\Delta$ into (B.18) to obtain (C.14). □

For independent uniform block probes, $\tau = L$ and $\Lambda = L \max_{k,a,v} \Pr(A_k = a \mid \varphi(A) = v)$. Coverage therefore reduces (C.14) to $\widetilde{O}(\min\{K\Delta T, m/\Delta + K\Lambda/\rho\})$ for fixed C_Δ. For this uniform design, a known-gap frozen raw learner has the same expected maximum rate with slope

$$s_\Lambda = \rho L/(e\Lambda). \tag{C.15}$$

Indeed, take $\pi = \Lambda/L$ in Lemma B.8 and apply the fixed-acquisition proof of Proposition B.5. Under uniform probing, a constant label has $\Lambda = 1$, full action has $\Lambda = L$, and equal groups of size h have $\Lambda = L/h$. A partition with a singleton group can have $\Lambda = L$ even if $G = 2$; the mixed probe instead gives $\tau \le 2L$ and $\Lambda \le 2G$. For correlated probes, the reciprocal of the smallest site-group mass is not a substitute for the ratio in (C.13). No matching minimax claim is made for an arbitrary fixed label or probe law.

Corollary C.4 (Matroid category label upper)**.** *For a matroid probe q of full coverage and any fixed deterministic category label φ, the learner of Theorem B.1 is ordinary-log exact and conditionally ρ-TV-stable. On the exchange-gap class $\mathcal{C}_{\mathcal{M}}(\Delta, C_\Delta)$, with fixed C_Δ, its simultaneous regret satisfies*

$$\mathbb{E}M_T = \widetilde{O}(\min\{K\Delta T, K/(\gamma(q)\Delta) + K\Lambda/\rho\}). \tag{C.16}$$

Calendar category label gives $\Lambda = 1$; full-basis category label gives $\Lambda = 1/\gamma(q)$. In particular, on the uniform and complete-graph tree families with $\gamma(q) = K/m$, the sufficient full-basis forgetting term is m/ρ.

Proof. This is Theorem B.1. Use the arm-inclusion ratio $\Lambda = \max_{i,v} \Pr_q(i \in A \mid \varphi(A) = v)/p_i$ over supported labels. For full category label it is either zero or $1/p_i$; take the maximum. No independence among basis inclusion indicators is required. □

The two terms lead to the probe-design objective

$$\inf_{q:\gamma(q)>0} \left\{ \frac{K}{\gamma(q)\Delta} + \frac{K\Lambda(q,\varphi)}{\rho} \right\}. \tag{C.17}$$

This is an upper-bound design objective, not a minimax characterization for every category label. The mixed probe controls both terms for sitewise groups. Uniform probing already has $\Lambda = 1$ for the relative-offset category label below.

C.4 EQUAL ENTROPY AND CONDITIONAL SAMPLING

Lemma C.2 (Projected-support contraction)**.** *Let $\mathcal{F}$ be generated by a swap system of order (d, w, ℓ) and let φ be a function of the served action. Fix a site k and a block $j \in [L]$, and write*

$$\mathcal{S}_k(j) := \{\varphi(a) : a \in [L]^d,\ a_k = j\}$$

for the projected label support *of that block. Let P_j and the auxiliary reference P_0 be the site-k environments defined above. Write $N_j^{\varphi,k}(n)$ for the number of the first n rounds whose disclosed label lies in $\mathcal{S}_k(j)$. For every causal learner that is ρ-TV-stable conditionally on $\varphi(S_t)$, every $j \in [L]$, and every event E measurable with respect to the first n actions,*

$$|P_j(E) - P_0(E)| \le 2\rho\Delta\mathbb{E}_0[N_j^{\varphi,k}(n)]. \tag{C.18}$$

When φ is a sitewise group tuple, $\mathcal{S}_k(j)$ is the set of labels with site-k group $G(j)$, and this is Lemma C.1.

Proof. Localization. Let X and X' be any two tables that agree except in row t, where they differ only in the coordinates of block j at site k. Couple the coins. A generated action selects exactly one block at site k. If $\varphi(S_t) \notin \mathcal{S}_k(j)$, then necessarily $A_{t,k} \neq j$, so the consumed slice at time t misses block j entirely; the two conditional laws given $\varphi(S_t) = v$ coincide for every such label. On $\{\varphi(S_t) \in \mathcal{S}_k(j)\}$, factor the replacement through the dummy $\perp$ and apply the φ-conditional ρ bound twice, giving 2ρ for each such v. Writing μ_v and μ'_v for the two transcript laws conditioned on $\varphi(S_t) = v$,

$$d_{\mathrm{TV}}(A(X), A(X')) = \sum_v \mathbb{P}(\varphi(S_t) = v) d_{\mathrm{TV}}(\mu_v, \mu'_v) \leq 2\rho \mathbb{P}(\varphi(S_t) \in \mathcal{S}_k(j)). \tag{C.19}$$

The bound holds at every surrounding table.

Coupling and telescoping. Use the raising coupling and reverse-calendar hybrid of Lemma C.1, with (C.19) replacing (C.4). At row s, earlier rows retain the reference law, and the row-change indicator has probability Δ independently of those rows and the learner coins. The term is at most $2\rho\Delta \Pr_0(\varphi(S_s) \in \mathcal{S}_k(j))$. Summing over $s < n$ gives (C.18). □

Write $\Phi_{N,d}$ for the deterministic maps $[L]^d \to [N]$ with fibres of common size $F = L^d/N$, where $N \mid L^d$. Under a uniform product reference action, every such map has label entropy $\log N$ and residual action entropy $\log F$. For fixed $C_\Delta \geq 1$, define

$$\mathfrak{R}_T^{(d)}(\varphi) := \inf_{A:\,(3)\text{ for }\varphi} \sup_{\nu \in \mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)} R_T(A, \nu),$$

where the infimum allows all causal learners. The public transcript of Section 3 is the same for every φ; the map changes only which categories receive a conditional budget.

Proposition C.4 (Equal-fibre category-design frontier)**.** *On a covered two-site product, with $K = 2w$, $L \geq 2$, $N \mid L^2$, $0 < \Delta \leq 1$, $0 < \rho \leq 1$, and $T \geq 1$,*

$$\inf_{\varphi \in \Phi_{N,2}} \mathfrak{R}_T^{(2)}(\varphi) = \widetilde{\Theta}\left(\min\left\{K\Delta T, \frac{m}{\Delta} + \frac{K}{\rho}\max\{1, N/L\}\right\}\right), \tag{C.20}$$

$$\sup_{\varphi \in \Phi_{N,2}} \mathfrak{R}_T^{(2)}(\varphi) = \widetilde{\Theta}\left(\min\left\{K\Delta T, \frac{m}{\Delta} + \frac{K}{\rho}\min\{N, L\}\right\}\right). \tag{C.21}$$

The constants absorb the fixed coverage and gap-regularity constants. Uniform probes attain the upper bounds with gap-agnostic ordinary-log exact learners, also for $\mathbb{E}M_T$. These are extrema over category designs, not a characterization of every individual map.

Proof. Occupancy lower bound. Let $\mathrm{occ}_k(v)$ count the distinct site-k candidates in $\varphi^{-1}(v)$. For any number of sites, pathwise

$$\sum_{j=1}^{L} N_j^{\varphi,k}(n) = \sum_{t \leq n} \mathrm{occ}_k(\varphi(S_t)). \tag{C.22}$$

Suppose $\mathrm{occ}_k(v) \leq r \leq L$ for every label. Freeze the other-site winners and use the uniform prior over the distinguished site's L winners. Since $N_j(n) \in [0, n]$, Lemma C.2 and (C.22) give

$$\frac{1}{L}\sum_j \mathbb{E}_j N_j(n) \leq \frac{n}{L} + \frac{2\rho\Delta r n^2}{L}.$$

Set $a = L/(8\rho\Delta r)$ and $n = \min\{T, \lfloor a \rfloor\}$. If $a \geq 1$, this is at most $3n/4$, while $n \geq \frac{1}{2}\min\{T, a\}$. If $a < 1$, the first-round mean regret is at least $w\Delta/2$ and $wL/(\rho r) < 8w\Delta$. Thus, in both cases,

$$\frac{1}{L}\sum_j R_T^{(k)}(A, P_j) \geq c\min\{w\Delta T, wL/(\rho r)\}. \tag{C.23}$$

This uses public action counts and permits arbitrary adaptive exploration.

For two-site equal fibres, $\mathrm{occ}_k(v) \leq \min\{L, F\}$, where $F = L^2/N$. Consequently $L/\min\{L, F\} = \max\{1, N/L\}$. Apply (C.23) at both sites, average the other winners under the same independent uniform prior, and sum. This proves the universal stability lower bound in (C.20).

Best construction. Identify candidates with $\mathbb{Z}/L\mathbb{Z}$. Concatenate the diagonal matchings $M_s = \{(i, i+s) : i \in \mathbb{Z}/L\mathbb{Z}\}$ over all L offsets, and split the resulting sequence into N consecutive fibres of size F. A window of F pairs meets at most $\lceil F/L \rceil + 1$ matchings, each containing every coordinate once. Under uniform probes,

$$\Lambda \leq \frac{L}{F}(\lceil F/L \rceil + 1) \leq 3\max\{1, N/L\}. \tag{C.24}$$

For $F \geq L$, the middle expression is at most $1 + 2L/F \leq 3$; for $F < L$, it equals $2L/F$. Proposition C.3, with $\tau = L$, supplies the matching upper bound.

Worst construction. Every fibre has at most $\min\{F, L\}$ pairs with a specified coordinate, so every equal-fibre map has $\Lambda \leq L \min\{F, L\}/F = \min\{N, L\}$ under uniform probes. This gives a uniform upper bound for (C.21). Now list pairs in row-major order and split into consecutive fibres of size F. Such a fibre occupies at most $r = \min\{L, \lceil F/L \rceil + 1\}$ first coordinates, with

$$\frac{L}{r} \geq \frac{\min\{N, L\}}{3}.$$

Indeed, $\lceil F/L \rceil + 1 \leq F/L + 2$, which is at most $3F/L$ when $F \geq L$ and at most 3 when $F < L$. Equation (C.23) at site 1 gives the matching stability floor, since $K = 2w$. It holds for every frozen site-2 winner. The first fibre also has $\min\{F, L\}$ copies of its first coordinate, so this construction has exactly $\Lambda = \min\{N, L\}$ under uniform probes.

Finally, Lemma D.1 supplies the identification floor for every map. Combine each stability floor with it using $\max\{\min(u, v), \min(u, z)\} \geq \frac{1}{2}\min\{u, v + z\}$, and use coverage $KL = \Theta(m)$. The full range $0 < \Delta \leq 1$ uses the raising coupling of Lemma C.2 for stability and the separate large-gap search argument of Lemma D.1 for identification. □

Corollary C.5 (Optimal equal-fibre category design at d sites)**.** *For a covered d-site product with $K = dw$, $d \geq 2$, and $N \mid L^d$, under the same gap, stability, and horizon assumptions,*

$$\inf_{\varphi \in \Phi_{N,d}} \mathfrak{R}_T^{(d)}(\varphi) = \widetilde{\Theta}\left(\min\left\{K\Delta T, \frac{m}{\Delta} + \frac{K}{\rho}\max\{1, N/L^{d-1}\}\right\}\right). \tag{C.25}$$

Concatenating diagonal matchings and cutting equal fibres attains this rate for every such divisor N.

Proof. Each fibre has $F = L^d/N$ tuples, so $\mathrm{occ}_k(v) \leq \min\{L, F\}$ at every site. Apply (C.23) for each frozen configuration of the other winners, average under the independent uniform winner prior, and sum all d sites. This gives the stability floor with $L/\min\{L, F\} = \max\{1, N/L^{d-1}\}$; combine with Lemma D.1 as above.

For attainment, concatenate the L^{d-1} matchings

$$M_\sigma = \{(i, i + s_2, \ldots, i + s_d) : i \in \mathbb{Z}/L\mathbb{Z}\}, \qquad \sigma = (s_2, \ldots, s_d) \in (\mathbb{Z}/L\mathbb{Z})^{d-1},$$

and split into fibres of size F. These matchings partition all tuples. Each contains every candidate once at each site, so the window argument of (C.24) gives $\Lambda \leq 3\max\{1, N/L^{d-1}\}$. Uniform probes and Proposition C.3 prove the upper bound, including ordinary-log exactness and the simultaneous guarantee. When $N = L^r$ for $0 \leq r \leq d - 1$, the simpler checksum label $\varphi(a) = (a_2 - a_1, \ldots, a_{r+1} - a_1)$ has equal fibres and $\Lambda = 1$: conditionally on the differences, a_1 and every remaining free coordinate are uniform. For $r = 0$ this is the constant label. □

Proof of Corollary 1. Set $N = L$ in Proposition C.4. The diagonal fibres are precisely the relative-offset labels, and the row-major fibres are precisely the first-coordinate labels. Their uniform-probe inclusion factors are 1 and L, respectively. Both labels have L equally likely values and L actions per fibre, so their label and residual action entropies both equal $\log L$, as in Figure 1. The resulting stability terms are K/ρ and KL/ρ; the identification term is common. Their factor-L separation in total regret applies when stability dominates identification and the horizon cap is inactive. □

C.5 Probe design and overlapping protected categories

Probe design reduces conditional inclusion. The parameter in Proposition C.3 is $\Lambda(q, \varphi)$: it describes a probe and a label together. For a concrete separation, at every position let the two groups be $\{1\}$ and $\{2, \ldots, L\}$, and let φ record their tuple. Uniform product probes have $\tau = L$ and $\Lambda = L$. Instead, independently at each position, put

$$\bar{q}_{k,1} = \tfrac{1}{2}, \qquad \bar{q}_{k,a} = \frac{1}{2(L-1)} \quad (a \neq 1).$$

Both groups now have mass $1/2$, so $\tau(\bar{q}) = 2(L - 1)$ and $\Lambda(\bar{q}, \varphi) = 2$. Thus Proposition C.3 gives

$$\mathbb{E}M_T \leq CC_\Delta \min\left\{K\Delta T, \frac{KL\log(dLT)}{\Delta} + \frac{K}{\rho}\right\}. \tag{C.26}$$

The $G = 2$ lower bound of Proposition C.1 matches this scale and allows arbitrary causal exploration. In particular, fix d, w, C_Δ, take $\Delta = 1/4$, $\rho = L^{-2}$ and $T = L^4$. The minimax factual regret is $\Theta(KL^2)$, whereas $K\Lambda(q_{\mathrm{unif}}, \varphi)/\rho = KL^3$. This is a strict $o(K\Lambda/\rho)$ separation for the fixed uniform probe, even for total regret outside the horizon cap. The mechanism is coverage-preserving redistribution of probes toward

the singleton group. The example compares fixed probes within one category contract; the optimized design objective in (C.17) remains an upper bound.

Overlapping categories can be protected directly. Let $\mathscr{C}$ be a finite family of fixed subsets of the product action family. For each $C \in \mathscr{C}$, require (3) with the conditioning event $\{S_t \in C\}$ whenever it has positive probability. This specifies a guarantee for each category separately. For a fixed probe q with positive candidate marginals define

$$\Lambda_{\mathscr{C}}(q) = \max\left\{1, \max_{\substack{C\in\mathscr{C}:q(C)>0\\ k,a}} \frac{\Pr_q(A_k = a \mid A \in C)}{q_{k,a}}\right\}. \tag{C.27}$$

An empty inner maximum contributes zero.

Proposition C.5 (Overlapping-category upper bound)**.** *The fixed-probe completion-envelope learner using $\Lambda_{\mathscr{C}}(q)$ in its gate is calendar ρ-TV-stable and ρ-TV-stable within every category in $\mathscr{C}$. It is gap-agnostic and pathwise ordinary-log exact for all separate single erasures. Under the assumptions of Proposition C.3, its simultaneous regret satisfies*

$$\mathbb{E}M_T \le CC_\Delta \min\left\{K\Delta T, \frac{K\tau\log(dLT)}{\Delta} + \frac{K\Lambda_{\mathscr{C}}(q)}{\rho}\right\}.$$

Proof. Fix a deterministic table, a named row t, and one category C. On reaching the phase containing t, condition on its entry state and all other phase probes, rewards and tie coins, as in Lemma B.6. The target probe remains independent with law q. Conditional on $A_t \in C$, every IPS inclusion has expectation at most $\Lambda_{\mathscr{C}}(q)$. The two frozen-certificate movement terms in that lemma are therefore at most $\Lambda_{\mathscr{C}}(q)$ and $(e-1)\Lambda_{\mathscr{C}}(q)$. The shared-coin disagreement is at most ρ after multiplying by the gate slope $\rho/(e\Lambda_{\mathscr{C}}(q))$. Conditioning on $S_t \in C$ mixes this reached-phase case with the case of commitment before t, where erasure is inert. If $q(C) = 0$, only the inert case can support the category. Applying the same argument to the whole action family gives calendar stability, since unconditional IPS inclusion expectations equal one. This argument holds separately for every C with the same gate, so no union bound over categories is needed. Envelope domination and the utility proof of Proposition B.2 are unchanged. □

Protecting every atom of the joint membership label $\psi(S) = (\mathbf{1}\{S \in C\})_{C\in\mathscr{C}}$ also suffices, by mixing its conditional laws with their common pre-feedback probabilities. It can be strictly more costly. For even $L \ge 4$ and two positions, take the L cyclic categories $C_j = \{(a,b) : a \in \{j, j+1, \dots, j+L/2-1\} \pmod L\}$. Uniform product probes have $\Lambda_{\mathscr{C}} = 2$. The family contains complementary pairs, so its guarantee implies calendar stability; the preceding upper and the calendar lower give the $\widetilde{\Theta}(K/\rho)$ stability scale. In contrast, the full membership vector determines a and has $\Lambda = L$; protecting each membership atom is exactly the first-choice contract, whose scale is $\widetilde{\Theta}(KL/\rho)$ by Corollary 1. Both comparisons retain the common identification term and horizon cap.

What transfers from heterogeneous singleton gaps. There is also a direct calendar-stable product upper bound from Theorem E.1. Under the conditional-mean model (6), suppose each position has a unique winning block and every suboptimal block-average gap lies in $[\delta, \Gamma]$, where $0 < \delta \le \Gamma \le 1$. Run its singleton learner independently at each of the d positions, with budget ρ/d, using only the served block averages that the respective components request. Then

$$\mathbb{E}M_T \le CK\log^3\left(\frac{100Ld(T+1)}{\rho}\right)\Psi_{\mathrm{het}}(L, T, \rho/d; \delta, \Gamma). \tag{C.28}$$

Indeed, for each deterministic table the component transcript laws factor over independent tapes, so their joint TV is at most $d(\rho/d)$. This holds even for the tuple of component public transcripts; the served slate transcript is a projection. Component replay uses only its retained factual slices, so their union reconstructs the product execution pathwise. For utility, block averages inherit the conditional-mean assumption even with dependent positions, and the product's simultaneous regret is pathwise at most w times the sum of the component maxima. Theorem E.1 proves (C.28). Thus its interpolants yield a constructive upper bound for calendar stability. Dividing the budget across positions can lose a factor d in the stability term relative to the joint learner in Theorem 1.

D MATCHING CALENDAR LOWER BOUNDS

This appendix proves the identification and calendar-stability floors for the hard family (C.1), throughout $0 < \Delta \le 1$ and $0 < \rho \le 1$. The witnesses have independent rows and thus satisfy the fixed-conditional-mean model. The proofs use only public action prefixes; neither reward observations nor the learner's private coins need be public.

D.1 PRODUCT PACKING AND IDENTIFICATION

Let $\mathcal{F}$ be generated by a swap system of order (d, w, ℓ) (Definition B.3), with $L = \ell + 1 \geq 2$ and $K = dw$. An environment in $\Theta_{\mathrm{prod}}(\Delta)$ has one winner per site, with winner and loser means (p, q) from (C.1). The w coordinates in a block clone one Bernoulli variable, independently across blocks and rows. Every local gap is exactly Δ, and $\Delta_{\max} = K\Delta$, so this family belongs to $\mathcal{C}_{\mathrm{prod}}(\Delta, 1)$, including at $\Delta = 1$. If $N_{j,k}(n)$ counts plays of block j at site k, its regret when j wins is $w\Delta\{n - \mathbb{E}_j N_{j,k}(n)\}$. Site regrets add.

Lemma D.1 (Identification floor)**.** *For $0 < \Delta \leq 1$ and $T \geq 1$, every causal learner satisfies*

$$\sup_{\theta \in \Theta_{\mathrm{prod}}(\Delta)} R_T(A, \theta) \geq c_1 \min\{K\Delta T, KL/\Delta\} \tag{D.1}$$

for a universal $c_1 > 0$, without a stability assumption.

Proof. Use independent uniform winner indices in all L candidates at every site. Fix the winners at the other sites and suppress the distinguished site's index. Write P_j for the law with winner j, and P_0 for the reference law with every block at that site of mean q. The reference need not belong to the class.

Gaps $0 < \Delta \leq 1/2$. Here $p = 1/2$ and $q = 1/2 - \Delta$. Let P_j^H denote the law of the first H public actions. The likelihood chain rule for the acquired history, followed by projection, gives

$$\mathrm{KL}(P_0^H \| P_j^H) \leq \mathrm{kl}(q, p)\mathbb{E}_0 N_j(H) \leq 4\Delta^2 \mathbb{E}_0 N_j(H). \tag{D.2}$$

Only acquired observations of block j contribute to the likelihood sum; their number is at most the play count $N_j(H)$ even after ingestion stops. The direction of KL is as displayed: for $p = 1/2$, $\mathrm{kl}(q, p) \leq \chi^2(\mathrm{Bern}(q) \| \mathrm{Bern}(p)) = 4\Delta^2$, also when $q = 0$.

Put $a = L/(32\Delta^2)$ and $H = \min\{T, \lfloor a \rfloor\}$. If $a \geq 1$, then $H \geq \frac{1}{2}\min\{T, a\}$. Since $\sum_j \mathbb{E}_0 N_j(H) = H$, Pinsker's inequality and concavity give

$$\frac{1}{L}\sum_j d_{\mathrm{TV}}(P_0^H, P_j^H) \leq \sqrt{\frac{2\Delta^2 H}{L}} \leq \frac{1}{4}, \qquad \frac{1}{L}\sum_j \mathbb{E}_j N_j(H) \leq \frac{H}{L} + \frac{H}{4} \leq \frac{3H}{4}.$$

The mean site regret is at least $w\Delta H/4$, hence $\Omega(\min\{w\Delta T, wL/\Delta\})$. If $a < 1$, the first action loses at least $w\Delta(1 - 1/L) \geq w\Delta/2$ on average, while $wL/\Delta < 32w\Delta$. This proves the same bound when the displayed prefix length would be zero.

Gaps $1/2 < \Delta \leq 1$. First consider a deterministic experiment in which the winner always returns Δ and every loser returns zero. Fix the other winners, all other-site reward realizations, and the learner's coins, coupling these identically across choices of the distinguished winner. On the hypothetical all-zero observation history at the distinguished site, let $\mathcal{B}_t$ be the set of candidates served by round t, including plays after acquisition stops. Then $|\mathcal{B}_t| \leq t$. Before the first winner play, the actual execution follows that history, including its other-site actions and responses. Therefore $A_{t,k} = J$ implies $J \in \mathcal{B}_t$, and for a uniform winner J,

$$\Pr(A_{t,k} = J) \leq \Pr(J \in \mathcal{B}_t) \leq t/L.$$

For $n = \min\{T, \lfloor L/2 \rfloor\} \geq \frac{1}{3}\min\{T, L\}$, the mean site regret is at least

$$\frac{w\Delta n}{2} \geq \frac{w\Delta}{6}\min\{T, L\} \geq \frac{w}{24}\min\{\Delta T, L/\Delta\},$$

using $\Delta^2 \geq 1/4$ in the last inequality. This also lower-bounds the Bernoulli family: simulate any Bernoulli learner in the deterministic experiment by replacing each observed Δ with a fresh independent $\mathrm{Bernoulli}(\Delta)$ bit and leaving zeros unchanged, cloning within each block. The induced action and observation law is exactly that of the Bernoulli learner, with identical means and pseudo-regret. The deterministic bound applies to every causal simulator, so no transfer of stability is required.

Both arguments hold for every fixed configuration of the other-site winners and allow adaptive cross-site actions. Average those winners under the independent all-L prior and sum over sites to obtain (D.1). □

D.2 UNIVERSAL CALENDAR FLOOR

Corollary D.1 (Universal calendar floor)**.** *For $0 < \Delta \leq 1$, $0 < \rho \leq 1$, and $T \geq 1$, every causal calendar ρ-TV-stable learner satisfies*

$$\sup_{\theta \in \Theta_{\mathrm{prod}}(\Delta)} R_T(A, \theta) \geq c \min\{K\Delta T, KL/\Delta + K/\rho\} \tag{D.3}$$

for a universal $c > 0$. Coverage $KL = \Theta(m)$ gives the corresponding $\Omega(\min\{K\Delta T, m/\Delta + K/\rho\})$ bound.

Proof. Restrict each site's winner to two designated blocks, indexed by a uniform binary cube $\varepsilon \in \{0,1\}^d$. Neighboring row laws have TV at most 2Δ: couple the two Bernoulli blocks whose means p and q are swapped, each with mismatch probability $p - q = \Delta$. Set $a = 1/(32\rho\Delta)$ and $n = \min\{T, \lfloor a \rfloor\}$. If $a \geq 1$, then $n \geq \frac{1}{2}\min\{T, a\}$, and the general contraction part of Lemma A.3 gives neighboring action-prefix TV at most $4\rho\Delta n \leq 1/8$. Let N count prefix plays of the first designated block at the differing site. The two neighbors' summed site regret is at least

$$w\Delta\{n - \mathbb{E}_0 N + \mathbb{E}_1 N\} \geq \frac{7}{8} w\Delta n.$$

Average over the cube and sum sites. If $a < 1$, the first action loses at least $K\Delta/2$ on average and $K/\rho < 32K\Delta$. In either case,

$$\sup_{\theta \in \Theta_{\mathrm{prod}}(\Delta)} R_T(A, \theta) \geq c_2 \min\{K\Delta T, K/\rho\}. \tag{D.4}$$

Combine this with Lemma D.1, using

$$\max\{\min(u, v), \min(u, z)\} \geq \tfrac{1}{2}\min\{u, v + z\}.$$

The two priors need not coincide: both inequalities hold for the same learner's worst-case regret on $\Theta_{\mathrm{prod}}(\Delta)$. All records remain indexed by calendar time. □

D.3 Assembly of the product law

Proof of Theorem 1. For every $0 < \Delta \leq 1$, the hard family $\Theta_{\mathrm{prod}}(\Delta) \subseteq \mathcal{C}_{\mathrm{prod}}(\Delta, 1)$ gives the group lower bound in Proposition C.1. The gap-agnostic mixed-probe envelope gives the upper uniformly over $\mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)$ by Corollary C.2 and Theorem 2. More explicitly, for universal $c, C > 0$,

$$\inf_A \sup_{\nu \in \mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)} R_T(A, \nu) \geq c \min\{K\Delta T, KL/\Delta + KG/\rho\},$$
$$\inf_A \sup_{\nu \in \mathcal{C}_{\mathrm{prod}}(\Delta, C_\Delta)} R_T(A, \nu) \leq C C_\Delta \min\{K\Delta T, KL\log(dLT)/\Delta + KG/\rho\}.$$

The infimum is over causal learners satisfying the specified tablewise group-conditional guarantee. The G groups are nonempty at each site and may have unequal sizes. The endpoints $G = 1$ and $G = L$ are calendar and action-conditioned stability. Coverage replaces KL by $\Theta(m)$; no unused arm enters the lower bound. The case $T = 0$ is zero. □

E Heterogeneous singleton gap classes

For $0 < \delta \leq \Gamma \leq 1$ write $\mathcal{G}(m, \delta, \Gamma)$ for the singleton laws on $m \geq 2$ arms with a unique optimum and every suboptimal gap in $[\delta, \Gamma]$. The exogenous reward process satisfies (6) and is independent of the entire learner and requester tapes. No baseline mean is supplied. On an instance write $\Delta^{\mathrm{inst}}_{\max} = \max_i(\mu_\star - \mu_i) = \max_i \mu_i - \min_i \mu_i$ for the maximum gap and $\Delta^{\mathrm{inst}}_{\min} = \min_{i \neq \star}(\mu_\star - \mu_i)$ for the positive minimum suboptimal gap, so $\delta \leq \Delta^{\mathrm{inst}}_{\min} \leq \Delta^{\mathrm{inst}}_{\max} \leq \Gamma$. Throughout this appendix only, $D := \Delta^{\mathrm{inst}}_{\max}$ and $d := \Delta^{\mathrm{inst}}_{\min}$; these are analysis quantities, not learner inputs. The detector's output $\widehat{D}$ is an estimated upper bound, distinct from D. Let $\mathcal{R}^{\mathrm{cal}}_{m,\delta,\Gamma}(T, \rho)$ be the minimax factual expected pseudo-regret on that class among causal calendar ρ-TV-stable learners; the infimum may use δ and Γ. Define

$$\Psi_{\mathrm{het}}(m, T, \rho; \delta, \Gamma) = \min\Big\{\Gamma T, \max\Big[\min\{\sqrt{mT}, m/\delta\},\ 1/\rho, \min\big\{(m\Gamma T^2)^{1/3}, (m^2 T/\rho)^{1/3}, m\Gamma/\delta^2, m/\sqrt{\rho\delta}\big\}\Big]\Big\}. \tag{E.1}$$

The display is a range cap, an identification floor, a calendar floor, and a heterogeneous block of two interpolants and two plateaus.

Theorem E.1 (Adaptive minimax over heterogeneous gap classes)**.** *Fix $m \geq 2$, $T \geq 1$, and $0 < \rho \leq 1$. There are universal constants $c, C > 0$ and a learner $\mathcal{A}_{m,T,\rho}$, whose only inputs are m, T and ρ, such that simultaneously for every $0 < \delta \leq \Gamma \leq 1$,*

$$c\,\Psi_{\mathrm{het}}(m, T, \rho; \delta, \Gamma) \leq \mathcal{R}^{\mathrm{cal}}_{m,\delta,\Gamma}(T, \rho)$$

and

$$\sup_{\nu \in \mathcal{G}(m,\delta,\Gamma)} \mathbb{E}_\nu M_T \leq C \log^3\big(100m(T+1)/\rho\big)\, \Psi_{\mathrm{het}}(m, T, \rho; \delta, \Gamma). \tag{E.2}$$

Here M_T is the maximum of factual pathwise pseudo-regret and all separate single-erasure pseudo-regrets on the same table and tape, as in (17)*, for this singleton learner. The learner is causal and tablewise calendar ρ-TV-stable on the public interface of Appendix A.1 for each fixed named deletion, and is pathwise ordinary-log*

exact simultaneously for all such separate erasures, almost surely. Consequently the utility bound and same-tape replay identity hold after selecting any one factual row, including from the completed transcript. This does not enlarge the fixed-request stability claim or assert cumulative deletion.

Throughout, rewards lie in $[0,1]$ and have the fixed conditional means of (6); rows need not be independent. Write $\iota = 100\log(100m(T{+}1)/\rho)$. The two copies have independent learner tapes; their reward subsequences may be dependent. The public transcript contains served arms and continue/commit decisions, including the committed arm. Query flags, detector flags and estimated range, active sets, certificate draws, anchor draws, numerical scores, and saved coins are private replay metadata. Below, a terminal "release" is a controlled internal transition; its completion consensus also controls the resulting public transcript.

E.1 Range detector

On a unique-optimum instance the detector's inputs are only m, T, ρ. For a vector $c \in \mathbb{R}^m$ of fresh raw sums set

$$r(c) = \log \sum_{i=1}^{m} e^{c_i} + \log \sum_{i=1}^{m} e^{-c_i}. \tag{E.3}$$

Then $\max_i c_i - \min_i c_i \le r(c) \le \max_i c_i - \min_i c_i + 2\log m$. If coordinate k is missing, completed sums are $c + xe_k$ for $x \in [0,1]$. Write $A_k = \sum_{i\neq k} e^{c_i}$, $B_k = \sum_{i\neq k} e^{-c_i}$, and $x_k^\star = [\frac{1}{2}\log(A_k/B_k) - c_k]_{[0,1]}$, where $[x]_{[0,1]} = \max\{0, \min\{1, x\}\}$. Along the segment r is convex, so the completion envelope is $r_k^- = r(c + x_k^\star e_k)$ and $r_k^+ = \max\{r(c), r(c+e_k)\}$. The interior minimiser is required: an endpoint-only test is not a valid replay rule. Write $p_i^+(c) = e^{c_i}/\sum_j e^{c_j}$ and $p_i^-(c) = e^{-c_i}/\sum_j e^{-c_j}$. The derivative in the missing coordinate is $p_k^+ - p_k^-$, and for $x \in [0,1]$ one has $p_k^+(c{+}xe_k) \le e\, p_k^+(c)$ and $p_k^-(c{+}xe_k) \le p_k^-(c)$. Summing oscillations therefore yields

$$\sum_{k=1}^{m} (r_k^+ - r_k^-) \le e + 1 \le 2e. \tag{E.4}$$

Put $W_{\mathrm{rg}} = 16e/(\rho m)$. At scale $g = 2^{-j}$, starting at $j = 0$, take a fresh uniform-probe phase of length

$$L(g) = \left\lceil \frac{65536 m\iota}{g^2} + \frac{65536 e\iota}{\rho g} \right\rceil, \qquad n = L(g)/m,$$

draw $V \sim \mathrm{Unif}[0, W_{\mathrm{rg}}]$, and release only $F = \mathbf{1}\{r(c) + V \ge ng\}$. If $F = 0$, discard the sums and decrease g. If $F = 1$, return $\widehat{D} = \min\{1, 4g\}$. An unfinished phase makes no release.

On the simultaneous Bernstein event $|c_i - n\mu_i| \le \beta := 2\sqrt{n\iota} + 2\iota$ one has $nD - 2\beta \le r(c) \le nD + 2\beta + 2\log m$, and the length forces $2\beta + 2\log m + W_{\mathrm{rg}} \le ng/4$. Hence $D \le g/2$ implies $F = 0$ and $D \ge 2g$ implies $F = 1$. The first positive flag therefore satisfies $D \le \widehat{D} \le 8D$.

For a missing queried reward, retire if V lies in $[ng - r_k^+, ng - r_k^-]$. Source averaging and (E.4) bound the retirement probability by $\rho/8$. Scores are discarded after a completed flag. Lemmas A.4 and A.5 give ordinary-log exactness of the two-copy wrapper.

Detection occurs by the first scale $g \le D/2$. The planned forced-detection duration through that stage obeys $L_{\mathrm{det}}(D) \le C\iota(m/D^2 + 1/(\rho D))$, and the corresponding regret is $C\iota(m/D + 1/\rho)$. Each copy reserves at most half of its assigned slots. If a copy horizon H satisfies $H < 4$, serve a fixed arm throughout. If $H \ge 4$ and discovery has not completed by $\lfloor H/2 \rfloor$, the copy serves a fixed arm thereafter: failed discovery then forces $H \le 4L_{\mathrm{det}}(D)$, so the interrupted cost is absorbed by $\min\{DT, m/D + 1/\rho\}$ and hence by $\Psi_{\mathrm{het}}(m, T, \rho; d, D)$ at the instance minimum gap d. A successful detector's output is measurable in its stopping history; future rows retain their fixed conditional means given that history. It starts the one-copy certified learner of Appendix E.6 with input $\widehat{D}$ and global horizon T, truncated at the remaining owned slots. The phase-budget argument of that subsection bounds every entry-determined truncated prefix and does not condition a fixed-horizon theorem on a random remaining duration. Only one outer alternating pair is used.

E.2 Padded acquisition

At phase entry fix an active set A of size $N \ge 2$, a padding anchor $b \in A$, a length L, and a query rate $q_{\mathrm{acq}} \in (0,1]$. Independently at each of the L slots draw $Q \sim \mathrm{Ber}(q_{\mathrm{acq}})$ and a uniform probe $I \in A$. If $Q = 1$, serve I and ingest its reward; if $Q = 0$, serve b and ingest nothing. The phase's served actions do not depend on its current rewards. Fresh scores are

$$c_i = \sum_{s \text{ in phase}} Q_s \mathbf{1}\{I_s = i\} Y_s, \qquad \mathbb{E}[c_i \mid \text{entry}] = n\mu_i, \qquad n = q_{\mathrm{acq}} L/N.$$

Thus n is an expected exposure count, not a fixed observation count per arm. Terminal coins are independent of the probes. After a terminal release, discard the numerical scores. An unfinished phase has no terminal release.

For completeness, let σ be the phase-entry stopping time for the chronological filtration containing full past reward rows and exposed learner coins, but no future coins. The active set, N, L, q_{acq}, and anchor are fixed at entry. At each owned calendar slot s, fresh query/probe coins and (6) give

$$\mathbb{E}[Q_s \mathbf{1}\{I_s = i\} X_{s,i} \mid \mathcal{F}_{s-1}] = \frac{q_{\mathrm{acq}}}{N}\mu_i, \qquad \mathbb{E}[(Q_s \mathbf{1}\{I_s = i\} X_{s,i})^2 \mid \mathcal{F}_{s-1}] \le \frac{q_{\mathrm{acq}}}{N}.$$

The centered increment is bounded by one. The conditional exponential-moment argument of Lemma B.4, applied with variance budget $n = q_{\mathrm{acq}} L/N$, yields

$$\Pr\big(|c_i - n\mu_i| > 2\sqrt{n\iota} + 2\iota \mid \mathcal{F}_\sigma\big) \le 2e^{-2\iota}. \tag{E.5}$$

Conditioning includes intervening rows between owned slots. The parameters may be random because they are fixed at entry; no future-row independence is used. The detector is the case $q_{\mathrm{acq}} = 1, N = m$. An unfinished phase has no terminal confidence requirement.

E.3 Confident-winner certificate

Let $0 < a \le 1$ and $c \in \mathbb{R}^N$, and write $p_i(c) = e^{ac_i} / \sum_j e^{ac_j}$. Fix a threshold $\tau \in [4/5, 1)$ and set

$$\pi_i(c) = \frac{(p_i(c) - \tau)_+}{1 - \tau}, \qquad \pi_\perp(c) = 1 - \sum_i \pi_i(c). \tag{E.6}$$

At most one π_i is positive. No score-dependent numerical statistic is released.

Lemma E.1 (Source-averaged certificate sensitivity)**.** *For every c, $\sum_{k=1}^N d_{\mathrm{TV}}(\pi(c + e_k), \pi(c)) \le (1 + e)a$. The same bound covers increments $x_k e_k$ with $x_k \in [0, 1]$.*

Proof. If arm i exceeds τ anywhere on a unit segment from c, then either $k = i$ and $p_i(c) > \tau/(e^a - (e^a - 1)\tau) > 1/2$, or $k \ne i$ and $p_i(c) > \tau > 1/2$. Hence at most one arm j can carry positive certificate mass on the union of segments. Along the j-segment, π_j increases and its derivative is at most a wherever it is positive, so that segment contributes at most a. If $p_j(c) \ge \tau$, each other coordinate decreases π_j by at most $(e^a - 1)p_k(c)/(1 - \tau)$; summing those coordinates uses $\sum_{k \ne j} p_k \le 1 - \tau$ and contributes at most ea. □

Completion consensus draws a saved uniform against π_j and retires only between the endpoint masses. The two endpoints share the same non-bottom arm.

On a fresh padded phase with expected exposure n per arm set $\beta = 2\sqrt{n\iota} + 2\iota$ and $\tau = \exp(2a\beta + \log 4)/(1 + \exp(2a\beta + \log 4)) \ge 4/5$. On the Bernstein event, a non-bottom certificate is the unique optimum: $\pi_i(c) > 0$ forces $a(c_i - c_k) > 2a\beta$ for every $k \ne i$. For a target $\varepsilon \in (0, 1]$ take

$$n_0 = \lceil 512\iota/\varepsilon^2 \rceil, \quad L = \lceil N n_0 + 512\iota/(\rho\varepsilon) \rceil, \quad q_{\mathrm{acq}} = N n_0/L, \quad a = \min\{1, \rho N/(64 q_{\mathrm{acq}})\}.$$

Thus $q_{\mathrm{acq}} L/N = n_0$. If every suboptimal active gap is at least ε, the Bernstein event implies that the certificate returns $\perp$ with probability at most $5Ne^{-4\iota}$. Independently draw a Gibbs anchor J from $p(c)$ by the sampler of Appendix E.4 and save both tapes, including when the certificate commits.

E.4 Gibbs proposal sampler

We use the shared-proposal strategy of Kleinberg and Tardos (2002); Holenstein (2007), presented and analysed by Bavarian et al. (2020). That source attributes the construction and output-disagreement bound $2v/(1 + v)$ at TV distance v to the earlier work. We add the completion-consensus retirement rule and ordinary-log execution below.

Repeatedly draw independent proposals $K \sim \mathrm{Unif}(A)$ and $U \sim \mathrm{Unif}[0, 1)$. Factual sampling accepts the first proposal with $U < p_K(c)$ and retains the proposal prefix through acceptance. Each proposal is accepted with probability $1/N$, so the sampler terminates almost surely and has the Gibbs marginal $p(c)$.

With one missing queried coordinate k, form retained sums c^0 and endpoint laws $p^0 = p(c^0), p^1 = p(c^0 + e_k)$. At each saved proposal: accept if $U < \min\{p^0_K, p^1_K\}$; reject and continue if $U \ge \max\{p^0_K, p^1_K\}$; otherwise retire the affected copy. Along $c^0 + xe_k$, $x \in [0, 1]$, softmax coordinate k increases and every other coordinate decreases, so the endpoint envelopes cover every missing reward. If $v = d_{\mathrm{TV}}(p^0, p^1)$, a proposal commonly accepts with probability $(1 - v)/N$ and is ambiguous with probability $2v/N$. Summing common rejections gives retirement probability $2v/(1 + v) \le 2v$. At the factual accepted proposal the erased sampler cannot unanimously reject, because the factual probability lies inside the completion envelope. The stored factual prefix therefore always suffices: replay either accepts or retires by its end.

Certificate retirement sums to at most $(1+e)a$; Gibbs endpoint TV is at most $eap_k(c^0)$, and proposal-consensus retirement is at most twice that value. Form the union of the two ambiguity events before averaging the missing source; do not condition the second release's source on the first release's success. Averaging the uniform source

and the query flag yields retirement probability at most $(1+3e)q_{\mathrm{acq}}a/N < \rho$, including the saturated case $a = 1$. Require consensus for every required terminal output; clear the missing marker and numerical phase state after a common transition. The healthy copy remains factual; the affected copy either continues factually or stops acquiring.

E.5 Refinement and thinned filter

For refinement from a valid range γ to a target ζ, use $\varepsilon_j = \max\{\gamma 2^{-j}, \zeta\}$ for $0 \le j \le \lceil \log_2(\gamma/\zeta) \rceil$. Begin with any active anchor and use the preceding released Gibbs anchor as padding for the next phase. On a correct entry its gap is at most $2\varepsilon_j$. Use the certified parameters of Appendix E.3. On the simultaneous confidence event $|c_i - n_0\mu_i| \le \beta$, the displayed constants imply $4\beta \le n_0\varepsilon/2$ and $an_0\varepsilon \ge 8\iota$, so the Gibbs probability of selecting an arm with gap above ε is at most $N\exp[-a(n_0\varepsilon - 2\beta)] \le Ne^{-4\iota}$. This is the anchor accuracy that padding regret requires.

On a correct phase entry, expected phase regret, or that of any entry-determined truncated prefix, is at most

$$\gamma q_{\mathrm{acq}} L + 2\varepsilon L \le C\iota\left(\frac{N\gamma}{\varepsilon^2} + \frac{N}{\varepsilon} + \frac{1}{\rho}\right) + C \le C'\iota\left(\frac{N\gamma}{\varepsilon^2} + \frac{1}{\rho}\right) + C.$$

Summing over the geometric targets proves that a refinement from a valid range γ to a target ζ costs $C\iota(N\gamma/\zeta^2 + (1+\log(\gamma/\zeta))/\rho)$, including truncated prefixes. This converts a duration of order $1/(\rho\varepsilon)$ into regret of order $1/\rho$ per level. If targets first cross the unknown minimum gap d, the crossing target lies in $[d/2, d]$ and the call stops at cost $C\iota(N\gamma/d^2 + (1+\log(\gamma/d))/\rho)$ unless a score, anchor, or eligible-certificate failure occurs; these first failures receive the separate charge below.

A thinned filter at scale γ, padded by an anchor of gap at most $\eta\gamma$, uses the same acquisition rule at its chosen q_{acq}, not an all-query policy. Set

$$n = q_{\mathrm{acq}}L/N, \quad \beta = 2\sqrt{n\iota} + 2\iota, \quad \theta = 2\beta + \log N, \quad W_{\mathrm{fil}} = 4eq_{\mathrm{acq}}/\rho,$$

and retain

$$B(c, V) = \left\{ i : \log \sum_j e^{c_j} - c_i \le \theta + V \right\}, \qquad V \sim \mathrm{Unif}[0, W_{\mathrm{fil}}].$$

For a queried missing coordinate k, retained raw sums c^0, and the saved threshold V, write $B_0 = B(c^0, V)$ and $B_1 = B(c^0 + e_k, V)$. If $B_0 = B_1$, publish that common set. Otherwise retire the affected copy and switch to the healthy copy as in Lemma A.5. Write $h_i(c) = \log \sum_j e^{c_j} - c_i$. Along $c^0 + xe_k$ the derivative of h_i is $p_k - 1 \le 0$ at $i = k$ and $p_k \ge 0$ at $i \ne k$, so each membership predicate is monotone in $x \in [0, 1]$. Endpoint agreement is therefore agreement on the whole completion interval. This endpoint test is specific to the filter; the range detector retains its interior minimum. Padding or no-query rows do not increment c. Retirement occurs before an active-set release is published, and no fresh unobserved query is made. For the local disagreement calculation, condition on phase entry and all other phase probes and rewards, so the retained vector $c = c^0$ is fixed and the queried source is uniform. Put $p_k = e^{c_k}/\sum_j e^{c_j}$ and $a_k = \log(1 + (e-1)p_k)$. Then $\sum_k a_k \le e - 1$ and $h_i(c + e_k) - h_i(c) = a_k - \mathbf{1}\{i = k\}$. A common uniform threshold of width W can change membership only in intervals of these displacement lengths, so

$$\frac{1}{N}\sum_{k=1}^{N} \Pr_V\{B(c, V) \ne B(c + e_k, V)\} \le \frac{1}{NW}\sum_{k=1}^{N}[1 + (N-2)a_k] \le \frac{e}{W}. \tag{E.7}$$

Endpoint monotonicity makes this a bound for the full completion ambiguity event. The length $L = \lceil 4096N\iota/(q_{\mathrm{acq}}\gamma^2) + 256eN/(\rho\gamma) \rceil$ gives $n \ge 4096\iota/\gamma^2 + 64W_{\mathrm{fil}}/\gamma$, hence $4\beta + \log N + W_{\mathrm{fil}} \le n\gamma/2$. Every optimum survives, and every retained arm has gap at most $\gamma/2$. If the padding anchor has gap at most $\eta\gamma$, expected phase regret on a correct entry obeys

$$\gamma q_{\mathrm{acq}} L + \eta\gamma L \le C\iota N\left(\frac{1}{\gamma} + \frac{q_{\mathrm{acq}}}{\rho} + \frac{\eta}{q_{\mathrm{acq}}\gamma} + \frac{\eta}{\rho}\right) + C.$$

Source-averaged threshold disagreement, after multiplying by the actual query probability, is at most $q_{\mathrm{acq}}e/W_{\mathrm{fil}} = \rho/4$. Changing only the threshold width while retaining an all-query policy would not produce that factor.

E.6 Adaptive schedule

The certified one-copy learner receives the detector's upper bound $\widehat{D}$ and does not receive d. Every non-bottom certificate causes permanent commitment and ignores later feedback. Every noncertified refinement produces a Gibbs anchor used as the next padding arm. Filters use the padded threshold of Appendix E.5. A singleton active set is committed immediately. An unfinished phase has no terminal release.

ADAPTIVELEARNER(m, T, ρ)

Create two copies with independent tapes on odd and even slots, with owned horizons $\lceil T/2 \rceil$ and $\lfloor T/2 \rfloor$. Confidence and later branch choices use the global T, not the owned remainder.

Per copy. If the owned horizon is less than 4, serve a fixed arm. Otherwise reserve at most half of the owned slots for the detector of Appendix E.1. On the first positive flag, start CERTIFIEDLEARNER$(m, T, \rho, \widehat{D})$ and truncate it at the remaining owned slots. If detection expires, serve a fixed arm thereafter.

Named erasure. Only the owning copy can be affected. After that copy retires, copy the healthy action as in Lemma A.5. Replay uses saved query/probe coins, certificate uniforms, and factual accepted Gibbs proposal prefixes.

CERTIFIEDLEARNER$(m, T, \rho, \widehat{D})$

Cap. If $T \le m/\widehat{D}^2$, serve a fixed arm and stop.

Short. If $\widehat{D} > \rho$ and $T < m/(\rho\widehat{D})$, refine the full set from $\widehat{D}$ to $\zeta = (m\widehat{D}/T)^{1/3}$. Commit to a non-bottom certificate, or to the final Gibbs anchor if none appears.

Middle. If $\widehat{D} > \rho$ and $m/(\rho\widehat{D}) \le T < m/\rho^2$, set $g = m/(\rho T)$ and $\zeta = (\rho g^2)^{1/3}$. At dyadic scales $\gamma > g$ use $\eta = (\rho/\gamma)^{1/3}$ and $q_{\mathrm{acq}} = (\rho/\gamma)^{2/3}$: refine to $\eta\gamma$, then filter at scale γ, padded by the preceding Gibbs anchor. Once the bound is at most g, make one certified refinement from g to ζ. If no certificate appears, commit to the final anchor and ignore subsequent feedback.

Long. Otherwise set $g = \sqrt{m/T}$. At scales $\gamma > \rho$ use the same (η, q_{acq}) pair; at scales $\gamma \le \rho$ use $q_{\mathrm{acq}} = 1$ filters and no refinement, until the active-set bound is at most g. Then commit to any active arm and ignore later feedback.

Each refinement or filter has a deterministic planned length from Appendices E.2 and E.5, computed from the entry active set. Horizon truncation uses the weighted planned-prefix charge of Lemma E.2 and emits no terminal release.

There are $O(\log(T+1))$ planned coarse scales, uniformly in the unknown gaps: terminal targets are never smaller than $\sqrt{m/T}$ outside the cap, and unknown d below a terminal target creates no further phases. In the middle branch the ratio of the terminal target to $\sqrt{m/T}$ is $(m/(\rho^2 T))^{1/6} \ge 1$. Refinement subphases across all calls may number $O(\log^2(T+1))$; the log cube in Theorem E.1 accommodates them.

Lemma E.2 (Planned-prefix charging)**.** *At phase entry σ, let L, q_{acq}, the active set A, anchor b, and actual prefix length $s \le L$ be entry-measurable. In particular, s may be the minimum of L and the remaining owned horizon or detector allocation. On a correct entry with active gaps at most γ and anchor gap at most α,*

$$\mathbb{E}[\widehat{R}_{\mathrm{prefix}} \mid \mathcal{F}_\sigma] = s\left(q_{\mathrm{acq}}\frac{1}{|A|}\sum_{i\in A}\Delta_i + (1-q_{\mathrm{acq}})\Delta_b\right) \le L(q_{\mathrm{acq}}\gamma + \alpha). \tag{E.8}$$

An unfinished phase emits no terminal release. The global-T phase budgets therefore remain valid after random detection, with $\alpha = 2\varepsilon$ in refinement and $\alpha = \eta\gamma$ in thinned filtering.

Proof. Fresh query and uniform-probe coins give the displayed expectation at each of the s owned slots. Summation and $s \le L$ prove the bound. Multiply only by entry-measurable correct-history indicators before summing over phases; conditioning query counts on a future confidence event is unnecessary. Horizon truncation has no later phase, while a completed filter may change the next entry state. An arbitrary reward-selected mid-phase stopping rule is not covered. □

Write $S = \sqrt{mT}$, $P = 1/\rho$, $X = (m\widehat{D}T^2)^{1/3}$, $Y = (m^2T/\rho)^{1/3}$, $A = m/d$, $B = m\widehat{D}/d^2$, $D_{\mathrm{pl}} = m/\sqrt{\rho d}$, and $Q = A + \min\{B, D_{\mathrm{pl}}\} + P$.

Lemma E.3 (Generic finite-horizon bound)**.** *On a unique-optimum instance with maximum gap at most $\widehat{D}$, the one-copy learner above, including every truncated prefix charged at its planned budget, has expected pseudo-regret $\widetilde{O}\Big(\min\{\widehat{D}T, \max(S, P, \min(X, Y))\}\Big)$.*

Proof. The cap branch costs at most $\widehat{D}T$. In the short branch, Appendix E.5 and the terminal-anchor accuracy give $\widetilde{O}(m\widehat{D}/\zeta^2 + P + \zeta T) = \widetilde{O}(X + P)$. In the middle branch, a scale γ has refinement-plus-filter cost $\widetilde{O}\Big(m/(\rho^{2/3}\gamma^{1/3}) + P\Big)$. The geometric sum down to g is a logarithmic multiple of $m/(\rho^{2/3}g^{1/3})$, and the final refinement has the same order; these balance $\zeta T = Y$. Thus the middle cost is $\widetilde{O}(Y + P)$. In the

long branch, scales above ρ cost $\widetilde{O}(m/\rho + P)$. The $q_{\mathrm{acq}} = 1$ phases below ρ cost $\widetilde{O}(m/g)$, and terminal commitment costs gT. Since $m/g = gT = S$ and $m/\rho \le S$ whenever $\widehat{D} > \rho$ in this branch, the cost is $\widetilde{O}(S)$. If $\widehat{D} \le \rho$, the same conclusion follows from $S \ge m/\widehat{D} \ge m/\rho$. Incomplete phases are charged by Lemma E.2; the schedule need not finish inside T. □

Lemma E.4 (Certificate stopping)**.** *Suppose $d \ge 4\rho$ and write $E_d = d^{3/2}/\sqrt{\rho}$. Whenever the coarse procedure runs far enough to include a target $\varepsilon \le d$, its expected regret charged on correct entries before stopping is $\widetilde{O}\Big(\min\{m\widehat{D}/d^2, m/\sqrt{\rho d}\} + P\Big)$ including entry-determined truncated prefixes. First failures receive the separate exceptional charge below.*

Proof. If $E_d \ge \widehat{D}$, the first call has target $(\rho\widehat{D}^2)^{1/3} \le d$. Its geometrically descending targets stop on first crossing d, and Appendix E.5 bounds that cost by a logarithmic multiple of $m\widehat{D}/d^2 + P$, which is the smaller branch here. If $E_d < \widehat{D}$, completed coarse scales strictly above E_d cost a geometric sum at most a logarithmic multiple of $m/(\rho^{2/3}E_d^{1/3}) = m/\sqrt{\rho d}$. The first scale at or below E_d still uses certified refinement, has target at most d, and has first crossing target at least $d/2$, so its acquisition cost is a logarithmic multiple of $mE_d/d^2 + P = m/\sqrt{\rho d} + P$. Eligible certificates at any $\varepsilon \le d$ fail with probability at most $(T+1)^{-4}$ on a correct entry; a union over at most T completed phases and a T charge make the exceptional contribution a universal constant. □

Lemma E.5 (Plateau bound)**.** *The same one-copy learner, still without using d, also has expected pseudo-regret $\widetilde{O}(Q)$, including truncated prefixes.*

Proof. In the cap, $\widehat{D}T \le m/\widehat{D} \le A$. In the short branch, if $d \ge \zeta$, the refinement crossing bound of Appendix E.5 gives cost $\widetilde{O}(B + P)$ directly, without requiring $d \ge 4\rho$. Indeed,

$$E_d \ge E_\zeta = \sqrt{m\widehat{D}/(\rho T)} > \widehat{D}, \qquad B/D_{\mathrm{pl}} = \widehat{D}/E_d < 1.$$

If $d < \zeta$, then $B > m\widehat{D}/\zeta^2 = X$ and $D_{\mathrm{pl}} > m/\sqrt{\rho\zeta} \ge X$, so Q already exceeds the generic short-branch bound of Lemma E.3. In the middle branch, $d < \zeta$ gives $B > mg/\zeta^2 = Y$ and $D_{\mathrm{pl}} > m/\sqrt{\rho\zeta} = Y$. For $d \ge \zeta$ and $d \ge 4\rho$, Lemma E.4 applies, including the terminal refinement from g. The remaining case $\zeta \le d < 4\rho$ satisfies $Y/A = d/(\rho^{2/3}g^{1/3}) < 4$ because $g > \rho$. In the long branch, $d \ge 4\rho$ stops before fine filtering by Lemma E.4. For the remaining $d < 4\rho$, if $d \le 2g$ then $A \ge S/2$, so the generic S bound is a multiple of Q. If $d > 2g$, correct elimination leaves only the optimum: coarse scales above ρ cost a logarithmic multiple of $m/\rho < 4A$, and fine $q_{\mathrm{acq}} = 1$ filters until the bound is below d cost a logarithmic multiple of A. Thus one schedule attains the minimum of the generic bound and Q without selecting a branch using d. □

Charge each phase under the entry-measurable indicator that no earlier failure has occurred. By (E.5), a completed phase on a correct entry has score failure probability at most $2me^{-2\iota}$. Conditional on valid scores, fresh terminal coins give Gibbs-anchor failure at most $me^{-4\iota}$ and eligible-certificate failure at most $5me^{-4\iota}$. If H_j means that all earlier outputs were correct, conditional averaging gives

$$\Pr(H_j \cap \{\text{phase } j \text{ is the first failure}\}) \le (2me^{-2\iota} + 6me^{-4\iota}) \Pr(H_j \cap \{j \text{ is reached}\}).$$

At most T nonempty phases complete across both copies. Summing bounds the probability of any first failure by $T(2me^{-2\iota} + 6me^{-4\iota})$; charge all subsequent regret by DT. This exceptional contribution is absorbed by the displayed rates and their linear cap. No independence of the copies' reward subsequences is used, and query counts are never conditioned on future concentration success.

Lemma E.6 (Generic/plateau identity)**.** *For $m \ge 2$, $0 < \delta \le \Gamma \le 1$, $T \ge 1$, and $\rho > 0$, writing $S = \sqrt{mT}$, $P = 1/\rho$, $X = (m\Gamma T^2)^{1/3}$, and $Y = (m^2T/\rho)^{1/3}$,*

$$\min\Big\{\max(S, P, \min(X, Y)), \max\big(m/\delta, P, \min(m\Gamma/\delta^2, m/\sqrt{\rho\delta})\big)\Big\}$$

equals $\max\{P, \min(S, m/\delta), \min(X, Y, m\Gamma/\delta^2, m/\sqrt{\rho\delta})\}$.

Proof. Distributivity produces two cross terms. Normalise by $B_0 = m/\delta$ and set $u = T\delta^2/m$, $r = \Gamma/\delta$, $h = \rho/\delta$. If $u \le 1$, then $S \le m/\delta$ and $X \le m\Gamma/\delta^2$. The Y branch is at most $m/\sqrt{\rho\delta}$ when $h \le 1$; at most S when $u \ge h^{-2}$; and, when $u \le 1$, $h > 1$, and $u < h^{-2}$,

$$Y/B_0 = (u/h)^{1/3} < h^{-1} \le h^{-1/2}.$$

If $u \ge 1$, then $m/\delta \le S$, and if the plateau exceeds m/δ then $h < 1$ and both X and (when $u \le h^{-2}$) Y are at least S, while $u \ge h^{-2}$ forces Y to dominate the plateau. A further minimum with ΓT preserves the identity. Replacing a sum of three nonnegative terms by a maximum costs at most a factor of three. □

Lemma E.7 (Simultaneous adaptation)**.** *On a unique-optimum instance with minimum gap d and maximum gap at most $\widehat{D}$,* CERTIFIEDLEARNER *attains, without using d,*

$$\widetilde{O}\Big(\min\Big\{\widehat{D}T, \max(S, P, \min(X, Y)), Q\Big\}\Big). \tag{E.9}$$

Proof. Lemmas E.3 and E.5 bound the same execution, whose branches do not use d, including every entry-determined truncated prefix. □

Lemma E.6 identifies (E.9) with $\widetilde{O}(\Psi_{\mathrm{het}}(m, T, \rho; d, \widehat{D}))$. The detector's bound $D \le \widehat{D} \le 8D$ gives $\Psi_{\mathrm{het}}(m, T, \rho; d, \widehat{D}) \le 8\Psi_{\mathrm{het}}(m, T, \rho; d, D)$, which also absorbs detection cost. Monotonicity then gives the class bound at (δ, Γ).

To transfer the factual bound, couple every separate erasure to the same factual odd/even pair. Acquisition depends only on the common entry state and saved query/probe coins. The detector's full completion interval, the certificate's monotone endpoint masses, the Gibbs sampler's common acceptance/rejection tests, and the filter's monotone membership tests all require consensus before a transition is published. A common transition clears numerical phase state and the marker; otherwise the affected copy retires after its factual action and ingests no further rewards. Its later actions copy the immediately preceding healthy action, while the healthy copy remains factual. Every required factual Gibbs sampler terminates almost surely; finitely many calls give one common probability-one event, and every erased call ends within its saved factual proposal prefix. The hypotheses of Lemma A.5 therefore hold simultaneously for every row, and (A.5) gives $\mathbb{E}M_T \le 2\mathbb{E}\widehat{R}_T^{(\varnothing)}$, without an additional failure charge or a ρT penalty. Empty even copies and owned horizons below four use their specified fixed-arm branches, so no nonexistent healthy predecessor is requested. This is the upper bound of Theorem E.1.

E.7 MATCHING LOWER BOUND

Write $S_\star = \min\{\Gamma T, \sqrt{mT}, m/\delta\}$, $C_\star = \min\{\Gamma T, 1/\rho\}$, and

$$F_\star = \min\Big\{\Gamma T, (m\Gamma T^2)^{1/3}, (m^2T/\rho)^{1/3}, m\Gamma/\delta^2, m/\sqrt{\rho\delta}\Big\}.$$

Any prefix bound at length $s \le T$ applies at horizon T by appending a fixed real tail and using nonnegativity of per-round regret.

Statistical floor. If $\delta \le 1/64$, set $b_{\mathrm{st}} = \min\{\Gamma, 1/64\}$. For each $d \in [\delta, b_{\mathrm{st}}]$ the m Bernoulli environments with one arm at $1/2 + d$ and the others at $1/2$ lie in the class. An auxiliary all-$1/2$ law and Pinsker on a prefix of length $s = \min\{T, \lfloor m/(160d^2)\rfloor\}$ yield average regret at least $(3/2560)\min\{dT, m/d\}$. Optimising over d produces $S_\star$ up to the factor $\Gamma/b_{\mathrm{st}} \le 64$. If $\delta > 1/64$, the m deterministic one-hot environments of gap δ force regret $\Omega(\delta \min\{T, m\})$ by testing untested indices until discovery, which is again $\Omega(S_\star)$.

Calendar floor. The two one-hot $\mathrm{Bernoulli}(\Gamma)$ environments on distinct arms have row TV at most 2Γ. If $\rho\Gamma \le 1/16$, calendar contraction on the capped prefix $s = \min\{T, \lfloor 1/(16\rho\Gamma)\rfloor\}$ yields $\Omega(\min\{\Gamma T, 1/\rho\})$. If $\rho\Gamma > 1/16$, the first action has the same law under both environments, so their first-round regrets sum to at least Γ and $R \ge \Gamma/2 > 1/(32\rho)$. For $\rho = 0$, the same comparison at $s = T$ gives $R \ge \Gamma T/2$.

Heterogeneous contribution. Lemma E.8 below supplies, for m_{ch} challengers, a known mean-$1/2$ baseline, base gaps in $[d, G]$, and alternatives raising one challenger to $1/2 + d$, the bound

$$\Omega\Big(\min\{dT, m_{\mathrm{ch}}G/d^2, m_{\mathrm{ch}}/\sqrt{\rho d}\}\Big)$$

whenever $0 < d \le 1/64$ and $d \le G \le 1/4$. Take $m_{\mathrm{ch}} = m - 1$, $G = \Gamma/4$, and $b = \min\{\Gamma/16, 1/64\}$. In a raised alternative the baseline has gap d and every other gap is at most $G + d \le 5\Gamma/16$, so every such anchored instance lies in $\mathcal{G}(m, \delta, \Gamma)$ for $d \in [\delta, b]$ and has a unique optimum. The elementary identity

$$\sup_{d\in[a,b]} \min\{dT, A/d^2, B/\sqrt{d}\} = \min\Big\{bT, (AT^2)^{1/3}, (B^2T)^{1/3}, A/a^2, B/\sqrt{a}\Big\} \tag{E.10}$$

holds because $\min\{dT, f(d)\}$ for decreasing f is maximised at the unique intersection of dT with f, clipped to $[a, b]$. Applying it with $A = (m-1)G$ and $B = (m-1)/\sqrt{\rho}$ produces $F_\star$ up to a universal factor when $\delta \le b$. If $\delta > b$, then $F_\star \le 4096S_\star$, so the statistical floor covers $F_\star$. Combining the three floors and using $\max\{\min\{A, u\}, \min\{A, v\}, \min\{A, w\}\} = \min\{A, \max\{u, v, w\}\}$ proves the lower bound of Theorem E.1 with a universal $c > 0$. The infimum in $\mathcal{R}^{\mathrm{cal}}$ may know δ, Γ and need not be ordinary-log exact, so the same lower bound applies to the exact-replay subclass.

Lemma E.8 (Anchored heterogeneous family)**.** *Fix a challenger count $m_{\mathrm{ch}} \geq 1$, $0 < \delta \leq 1/64$, $\delta \leq \Gamma \leq 1/4$, and $0 < \rho \leq 1$. Let $\mathcal{C}(m_{\mathrm{ch}}, \delta, \Gamma)$ contain every Bernoulli instance with a known mean-$1/2$ baseline and challenger gaps in $[\delta, \Gamma]^{m_{\mathrm{ch}}}$, together with each alternative that raises one challenger to mean $1/2 + \delta$. Every causal tablewise calendar ρ-TV-stable learner has class regret at least $c \min\{\delta T, m_{\mathrm{ch}}/\delta + \min\{m_{\mathrm{ch}}\Gamma/\delta^2, m_{\mathrm{ch}}/\sqrt{\rho\delta}\} + 1/\rho\}$ for an absolute constant $c > 0$.*

Proof. The statistical comparison of the uniform-δ base with one raised alternative uses the adaptive likelihood identity and Pinsker, giving $\Omega(\min\{\delta T, m_{\mathrm{ch}}/\delta\})$. Calendar contraction on the same pair gives $\Omega(\min\{\delta T, 1/\rho\})$. For the interpolation term, if $\Gamma < 4\delta$ or $\rho \geq \delta/16$ then

$$\min\{m_{\mathrm{ch}}\Gamma/\delta^2, m_{\mathrm{ch}}/\sqrt{\rho\delta}\} \leq 4m_{\mathrm{ch}}/\delta,$$

so the statistical bound already covers the interpolation minimum. Otherwise set $\Gamma_{\mathrm{eff}} = \min\{\Gamma, \delta^{3/2}/\sqrt{\rho}\}$, so that $m_{\mathrm{ch}}\Gamma_{\mathrm{eff}}/\delta^2$ equals that minimum, and put $n = \lfloor 1/(1280\delta^2) \rfloor$. If $T < 2n$, then $\delta T \leq m_{\mathrm{ch}}/\delta$, and the statistical floor already covers the capped target. Otherwise put $a_0 = m_{\mathrm{ch}}\Gamma_{\mathrm{eff}}/(128\delta^3)$ and $\tau = \min\{T, \lceil a_0 \rceil\}$. Now $\Gamma_{\mathrm{eff}} \geq 4\delta$ gives $a_0 \geq 2n$ and $a_0 \geq m_{\mathrm{ch}}/(2\Gamma_{\mathrm{eff}}^2) \geq 8m_{\mathrm{ch}} \geq 1$. Using $\rho\Gamma_{\mathrm{eff}}^2 \leq \delta^3$,

$$2n \leq \tau \leq 2a_0 \leq m_{\mathrm{ch}}/(64\rho\Gamma_{\mathrm{eff}}).$$

Lemma E.9 applies on the prefix of length τ. Moreover $\delta\tau \geq \min\{\delta T, m_{\mathrm{ch}}\Gamma_{\mathrm{eff}}/(128\delta^2)\}$, so its lower bound is a universal constant times $\min\{\delta T, m_{\mathrm{ch}}\Gamma_{\mathrm{eff}}/\delta^2\}$. Combine the statistical, calendar, and heterogeneous floors by the maximum-versus-sum inequality. □

Lemma E.9 (Weighted calendar fingerprint)**.** *In the anchored class of Lemma E.8, write $n = \lfloor 1/(1280\delta^2) \rfloor$ and assume $\Gamma \geq 4\delta$. If $2n \leq T \leq m_{\mathrm{ch}}/(64\Gamma\rho)$, every admissible learner has class regret at least $c \min\{\delta T, m_{\mathrm{ch}}\Gamma/\delta^2\}$.*

Proof. We adapt the score method of Bun et al. (2014); Kamath et al. (2022) to fixed-row deletion and derive the required calendar inequality here. Tests f_i of the public action transcript take values in $[0, 1]$. Challenger means satisfy $p_i \in [1/4, 1/2]$, the baseline has mean $p_0 = 1/2$, and a baseline action contributes zero to the score. For nonnegative weights of maximum $w_{\max}$, the fixed-table score at a named time is

$$g_{t,X}(a) = w_{a_t} f_{a_t}(a) \frac{X_{t,a_t} - p_{a_t}}{p_{a_t}(1 - p_{a_t})},$$

with value zero on the baseline action. Its range lies in $[-2w_{\max}, 4w_{\max}]$. Calendar TV therefore moves its expectation by at most $6\rho w_{\max}$. In the erased execution, conditional on retained rows and learner coins, the entire action transcript is independent of the erased row. Averaging that row centers the test. The original row is averaged only in the proof; it is not available to the erased learner. A test depending on the fixed table is permitted because the TV bound holds separately for every deterministic table. Combining this with the adaptive Bernoulli likelihood identity yields

$$\sum_{i=1}^{m_{\mathrm{ch}}} w_i \frac{\partial}{\partial p_i} \mathbb{E}_p[f_i] \leq 6\rho T w_{\max}.$$

Assume for contradiction that regret is smaller than both $\delta T/32$ and $m_{\mathrm{ch}}\Gamma n/64$. Let $f_i = \mathbf{1}\{N_i \geq n\}$. On every near base vector $x_i \in [\delta, 2\delta]$, a stopped comparison with the raised alternative of arm i has KL at most $40n\delta^2 \leq 1/32$, so Pinsker and the regret hypothesis give $\mathbb{E}[f_i] \geq 3/4$. That comparison uses only the law of arm i and the number of its plays: it holds uniformly for every choice of the other challengers' gaps in $[\delta, \Gamma]$, and therefore under the product density $h^{m_{\mathrm{ch}}-1}$ on those coordinates. Let a, b, and u be the uniform densities on $[\delta, 2\delta]$, $[\Gamma/2, \Gamma]$, and $[\delta, \Gamma]$, and put

$$h = \tfrac{1}{4}a + \tfrac{1}{4}b + \tfrac{1}{2}u, \qquad W(x) = \int_\delta^x [a(s) - b(s)]\, ds, \qquad w(x) = W(x)/h(x).$$

The ordered near and far intervals give $W(\delta) = W(\Gamma) = 0$ and $0 \leq W \leq 1$. Since $h \geq 1/[2(\Gamma - \delta)]$, one has $0 \leq w \leq 2\Gamma$. The middle component is required: it bounds this nonnegative Radon–Nikodym weight on the bridge. Write $F_i(x) = \mathbb{E}_x f_i$, with means $p_i = 1/2 - x_i$. Integration by parts produces

$$\begin{aligned} D_i &:= \mathbb{E}_{h^{m_{\mathrm{ch}}-1},\, a}[f_i] - \mathbb{E}_{h^{m_{\mathrm{ch}}-1},\, b}[f_i] \\ &= -\mathbb{E}_{h^{m_{\mathrm{ch}}-1}} \int W(x_i)\, \partial_{x_i} F_i(x)\, dx_i \\ &= \mathbb{E}_{h^{m_{\mathrm{ch}}}} \left[w(x_i)\, \partial_{p_i} \mathbb{E}_p f_i \right]. \end{aligned}$$

The signed ratio $(a - b)/h$ is not a nonnegative weight and does not give this identity. Applying the score inequality with weights $w(x_i)$ therefore yields

$$\sum_i D_i \leq 12\Gamma\rho T \leq 3m_{\mathrm{ch}}/16.$$

Far arms with $x_i \geq \Gamma/2$ cannot be sampled n times often: since $h \geq b/4$ and each such arm has gap at least $\Gamma/2$,

$$R \geq \frac{\Gamma n}{8} \sum_i \mathbb{E}_{h^{m_{\mathrm{ch}}-1},\, b}[f_i],$$

so the far-probability sum is less than $m_{\mathrm{ch}}/8$. The near mass $3m_{\mathrm{ch}}/4$ then forces $\sum_i D_i > 5m_{\mathrm{ch}}/8$, a contradiction. Finally $n \geq 1/(2560\delta^2)$ converts the n expression to the stated rate. The continuous bridge cannot be replaced by a two-point prior. □

E.8 TABLE DP

Let $\mathcal{R}^{\mathrm{tab}}_{m,\delta,\Gamma}(T,\beta)$ be the factual class-minimax regret under $(0,\beta)$-Table DP in the sense of Azize and Basu (2024, Definition 3): adjacent tables differ in one whole calendar row, and the output is the action sequence.

Corollary E.1 (Approximate Table DP at epsilon zero)**.** *For $0 < \beta \leq 1$ and the classes of Theorem E.1,*

$$c\,\Psi_{\mathrm{het}}(m,T,\beta;\delta,\Gamma) \leq \mathcal{R}^{\mathrm{tab}}_{m,\delta,\Gamma}(T,\beta) \leq C' \log^3\!\big(100m(T+1)/\beta\big)\,\Psi_{\mathrm{het}}(m,T,\beta;\delta,\Gamma).$$

One gap-oblivious learner attains the upper bound and has the ordinary-log exactness and simultaneous single-erasure utility of Theorem E.1.

Proof. A $(0,\beta)$-Table-DP policy becomes calendar β-TV-stable by imputing zero at missing feedback: its erased execution equals its factual execution on the table whose named row is zero, and factual regret is unchanged. The lower bound of Theorem E.1 applies to the action projection. Conversely, if P_X is the public law of a calendar-γ policy and $Q_{X,-t}$ is its missing-row law, two real tables differing only at t satisfy $Q_{X,-t} = Q_{X',-t}$ by causality, hence $d_{\mathrm{TV}}(P_X, P_{X'}) \leq 2\gamma$. Run $\mathcal{A}_{m,T,\beta/2}$ and project to actions. Each term of $\Psi_{\mathrm{het}}(\cdot,\beta/2,\cdot)$ is at most twice the corresponding term at budget β, so $\Psi_{\mathrm{het}}(\beta/2) \leq 2\Psi_{\mathrm{het}}(\beta)$. Replay and erased utility come from this constructed learner, not from an arbitrary Table-DP policy. □

When $\beta\Gamma T \leq 1$ the rate equals the linear cap ΓT. The corollary does not characterise general (ε,β) approximate DP.

E.9 MODEL DISTINCTIONS

Wang and Zhu (2024) include epsilon zero in a stochastic, action-dependent expected-revision constraint. That constraint is not calendar TV.

Calendar TV does not imply the revision guarantee. Uniformly permute m arms and query each once in the first m slots; at the next slot serve arm 1 exactly when the observed reward of designated arm 1 was one. Calendar TV is at most $1/m$. On the deterministic instance where arm 1 returns one, revising those first-m rewards of arm 1 to zero is one revision on every feasible action vector and flips the subsequent action with probability one.

The revision guarantee does not imply calendar TV. Query arm 1 for N deterministic probes, N divisible by four and at least 16, then serve arm 1 only if the probes equal the alternating pattern $0,1,\ldots,0,1$. For any independent-row stationary law the factual flag probability is at most 2^{-N}. An admissible revision rule with expected probe revisions at most one, independent of reward values on that prefix, leaves at least $N/4$ required zeros and $N/4$ required ones untouched except on a Markov event of probability at most $4/N$, so the revised flag probability is $O(N^{-1} + 2^{-N/2})$. On a table with the alternating arm-1 rewards, erasing any named slot that requires a one flips the flag with certainty, so calendar TV equals one.

For IID bounded rewards, the pure-DP UCB bound of Azize and Basu (2022) gives a factual TV baseline of order $\widetilde{O}(\sum_{i\neq\star} 1/\Delta_i + m/\rho)$ after zero imputation. Azize et al. (2025) give matching asymptotic instance-dependent pure-DP regret bounds for Bernoulli rewards with $\mu_\star < 1$. These guarantees are distinct from the class rate (E.1).

F REQUEST SELECTION

The Envelope learner remains a single-erasure construction (Theorem 2). This appendix records the causal-registration corollary cited in the body: one request named from earlier public actions and the named row's protected category, executed after training, at the existing calibration.

Causally registered requests. Fix a deterministic reward table X. Let $Z = A(X;\xi)$ and $Z^{-t} = A(X^{(-t)};\xi)$ denote factual and erased transcripts on the same learner tape ξ. Write $\mathcal{F}_{t^-}$ for the factual public history strictly before the action S_t is drawn, and let V_t be its protected group tuple. A requester has an independent tape ζ and names one index $J \in [T]$ satisfying

$$\{J = t\} \in \mathcal{F}_{t^-} \vee \sigma(V_t, \zeta) \qquad \text{for every } t. \tag{F.1}$$

Thus the rule may inspect earlier public actions and phase decisions, and the named row's protected category, but not its within-group choice, reward, or a later phase decision. The index rule must be implementable with this chronological information, even if evaluated retrospectively. Deletion is executed after training from the completed retained log. Registration does not alter factual training, the environment is not resumed, and the factual index J is held fixed in the erased reference.

Corollary F.1 (Causal registration at unchanged calibration)**.** *The group Envelope learner of Theorem 2, at its existing budget ρ, satisfies*

$$d_{\mathrm{TV}}\Big(\mathcal{L}(Z, J, Z), \mathcal{L}(Z, J, Z^{-J})\Big) \le \rho \tag{F.2}$$

for every bounded deterministic table and every requester satisfying (F.1)*. Replay of the selected-erased execution remains pathwise ordinary-log exact. On the covered gap-regular product class, factual and selected-erased expected pseudo-regrets both obey*

$$O\left(C_\Delta \min\left\{K\Delta T, \frac{m\log(dLT)}{\Delta} + \frac{KG}{\rho}\right\}\right).$$

The shared-tape disagreement probability is also at most ρ conditional on each supported pair (J, V_J).

For $G = 1$, the registration rule is predictable before the current action: $\{J = t\} \in \mathcal{F}_{t^-} \vee \sigma(\zeta)$. On homogeneous unit-width covered products with $L \ge 2$, $0 < \rho \le 1/16$, and $0 < \Delta \le 1/4$, the minimax factual regret under (F.2) *for every such predictable requester is*

$$\widetilde{\Theta}\left(\min\left\{K\Delta T, \frac{m}{\Delta} + \frac{K}{\rho}\right\}\right).$$

Proof. Let $D_t = \mathbf{1}\{Z \neq Z^{-t}\}$, and let $\mathcal{R}_t$ be the event that the acquisition phase containing t is reached. Off $\mathcal{R}_t$, the requested feedback was not ingested and $D_t = 0$.

On $\mathcal{R}_t$, use the leave-one-out conditioning field $\mathcal{H}_t$ of Lemma B.6, augmented by the independent requester tape. It contains the pre-phase state and earlier randomness, the tie-breaking coins, and all other probes and induced rewards in the target phase; it excludes the target probe and the target phase coin U_p. A reached phase is acquired completely before its gate is tested. Potential probes define this conditioning on the whole probability space, followed by restriction to $\mathcal{R}_t$.

The indicator of $\{J = t\}$, restricted to $\mathcal{R}_t$, is measurable with respect to $\mathcal{H}_t, V_t$. Conditioning preserves the target probe's within-group law used in Lemma B.6, while U_p remains independent and uniform. By Lemma B.5, disagreements can occur only when that coin lies between the factual and envelope gates. If a completion changes a retained winner, all gates are zero. Otherwise the positive gates share a candidate. The conditional expected certificate movement and the slope $\rho/(e\Lambda)$ therefore give, on the reached event,

$$\mathbb{E}[D_t \mid \mathcal{H}_t, V_t] \le \rho.$$

Multiplying by the measurable selection and label indicators yields, for every group tuple v,

$$\Pr(D_t = 1, J = t, V_t = v) \le \rho \Pr(J = t, V_t = v, \mathcal{R}_t) \le \rho \Pr(J = t, V_t = v). \tag{F.3}$$

Sum over t, v to obtain $\Pr(D_J = 1) \le \rho$. In the shared coupling, the original transcript and selected index agree in both histories; only the final response can differ. The no-update law is supported on triples (z, j, z), while the update law restricted to this diagonal is a submeasure of the no-update law with missing mass $\Pr(D_J = 1)$, so full-history total variation equals this disagreement probability. Division of (F.3) by a positive selection and label probability proves the conditional statement.

The ordinary-log identity already holds for every tape and named row, so substitution of J preserves it. The expected-maximum bound of Theorem 2 controls $\widehat{R}_T^{(-J)} \le M_T$ on the common table and tape. Selection therefore incurs no additional failure charge. This utility argument permits any selected factual row; the TV argument additionally requires (F.1).

For calendar matching, predictable requesters include every deterministic fixed index. The full-history requirement for those indices implies calendar stability by projection. The identification and calendar lower bounds of Theorem 1 apply, while the upper follows from the displayed bound at $G = 1$. □

Corollary F.2 (Causal registration for the matroid envelope)**.** *For the learner of Theorem B.1, use $V_t = \varphi(S_t)$ and the same requester information restriction* (F.1)*. At the existing calibration,* (F.2) *holds for every bounded*

deterministic table. The shared-tape disagreement probability is at most ρ conditional on each supported (J, V_J). Selected replay is pathwise ordinary-log exact, and selected utility is bounded by the expected maximum in (B.24)*.*

Proof. Use the leave-one-out field in the proof of Theorem B.1. It fixes retained exchange weights and the reached-phase event but exposes neither the target basis nor the terminal coin. Adjoin V_t and the independent requester tape. The registration indicator is then measurable, conditional inclusion inflation bounds the expected frozen envelope surplus by $e\Lambda$, and the independent terminal coin with slope $\rho/(e\Lambda)$ gives conditional disagreement at most ρ. Multiply by the registration and label indicators and sum exactly as in (F.3). This uses no independence among elements of a served basis. The simultaneous replay identity and $\widehat{R}_T^{(-J)} \le M_T$ give the remaining assertions. □

The result can improve on a generic completed-transcript selector. For $T \ge 2$, fix an arm a at site one and take

$$J = \min\left(\{t \in \{2, \dots, T\} : A_{t-1,1} = a\} \cup \{T\}\right).$$

This names the row immediately after the first appearance of a, with fallback T. It is predictable, but on the all-zero table every gate vanishes, every probe sequence has positive probability, and every index $2, \dots, T$ can be selected. Calendar Envelope learner nevertheless retains its K/ρ forgetting term. Selecting the appearance row itself instead would inspect that row's current action. The unchanged-calibration TV assertion uses the causal information in (F.1). The next result assigns a budget to selection from the completed transcript. Same-tape replay and the expected-maximum utility bound cover one selected factual row. Neither corollary treats cumulative deletion: the composition principle in Lemma A.2 would require one-step stability on already-missing inputs, which the retained single-marker envelope does not supply.

F.1 A BUDGET FOR SELECTION FROM THE COMPLETED TRANSCRIPT

A completed-transcript requester can be described by a kernel $k(t \mid z)$: after seeing the factual transcript z, it selects row t with this probability. Its additional random draws are independent of the learner's tape. Define

$$C_{\mathrm{sel}} := \sum_{t=1}^{T} \sup_z k(t \mid z), \qquad 1 \le C_{\mathrm{sel}} \le T.$$

The supremum ranges over the factual transcript space. This quantity is one for a request distribution independent of the transcript. A deterministic selector whose range contains at most b indices has $C_{\mathrm{sel}} \le b$.

Corollary F.3 (Completed-transcript selection)**.** *Let a known constant $C \ge C_{\mathrm{sel}}$ be supplied before training. Run the calendar Envelope learner at budget ρ/C. For every bounded deterministic table, with Z, Z^{-t} defined above and $J \sim k(\cdot \mid Z)$,*

$$d_{\mathrm{TV}}\big(\mathcal{L}(Z, J, Z), \mathcal{L}(Z, J, Z^{-J})\big) \le \rho.$$

Its selected erased execution is reconstructed pathwise from retained feedback. Under the product assumptions of Theorem 2, both factual and selected-erased expected pseudo-regrets are at most

$$O\left(C_\Delta \min\left\{K\Delta T, \frac{m \log(dLT)}{\Delta} + \frac{KC}{\rho}\right\}\right).$$

Proof. The proof of Proposition B.2 bounds shared-tape disagreement, so $\Pr(Z \ne Z^{-t}) \le \rho/C$ for each fixed t. Conditional on the learner's whole tape, selection has probabilities $k(t \mid Z)$. Thus

$$\Pr(Z \ne Z^{-J}) = \sum_t \mathbb{E}[k(t \mid Z)\mathbf{1}\{Z \ne Z^{-t}\}] \le \sum_t \sup_z k(t \mid z)\ \Pr(Z \ne Z^{-t}) \le \rho.$$

The coupling inequality gives the stated joint-history bound. The pathwise ordinary-log identity holds for every row and tape, so it also holds for J. Finally, $\widehat{R}_T^{(-J)} \le M_T$ pointwise; substitute ρ/C and $G = 1$ into the simultaneous utility bound of Theorem 2. □

This reference keeps the original transcript and its selected index fixed and uses the same learner tape for the erased response. It describes one request after training. Causal registration above can give a smaller cost when the requester's available information satisfies (F.1).

G EXPERIMENTAL METHODS AND VERIFICATION

The executable supplement contains the registered configurations, every completed seed outcome, phase sufficient statistics, private episode records, independent checks, and figure generators. The present results are fresh runs of these configurations. Historical sources are preserved separately. Synthetic methods use master seed 2026092602; empirical utility uses separately addressed source, probe, priority, gate, and privacy-noise streams. The replay campaign uses master entropy 202609260731. The configurations specify every address and were saved before their new outcomes. Earlier experiments informed the budget and horizon choices. A capped run remains in its cell mean; a commitment at the horizon is a commitment with acquisition fraction one. Population means enter the environment and regret evaluator, while learners receive selected feedback only.

Campaign	Independent seed units	Method outcomes
Balanced categories, $d = 2, 4, 8$	64 per dimension	1920
Equal-entropy primary grid	128 per L	2688
Equal-entropy scaling grid	128 per L	3456
Dependent and heterogeneous rewards	32 per reward law	10368
Synthetic private comparisons	32 per (d, L)	2688
Three preference-derived laws, ten budgets, four horizons	32 per source	24320
MovieLens genre categories, three budgets	32	288
Nguyen-Tang–Arora specialization	16 clusters per cell	512 paired runs

Table 5: Campaign sizes. Several methods, budgets, or horizons share a seed. A method outcome is not an independent statistical unit. Replay requests are counted separately below.

G.1 CATEGORY COSTS AND FINITE-HORIZON GEOMETRY

All cost curves use the unknown-gap Envelope learner of Theorem 2, with its stated inverse-propensity scores, reserve, and gate constants. Complete phases have lengths $1, 2, 4, \ldots$; a final phase is truncated at T. Commitment uses the strict comparison $U_p < g_p$. Uniform probes give $\tau = L$. Balanced G-group labels have $\Lambda = G$; the offset and first-choice labels have $\Lambda = 1$ and L. The certificate-only control commits when the same certificate exceeds its uncertainty reserve. It isolates the additional waiting introduced by the stability gate. Every policy sees the same potential probes, rewards, tie ranks, and phase coins within a seed.

The balanced grid uses $L = 8$, $\Delta = 0.2$, $T = 2^{26}$, $d \in \{2, 4, 8\}$, $G \in \{1, 2, 8\}$, and $\rho \in \{0.05, 10^{-3}, 3 \times 10^{-6}\}$, with 64 seeds. The primary equal-entropy grid uses $d = 2$, $L \in \{4, 8, 16\}$, $T = 2^{26}$, $\Delta = 0.2$, budgets $\{0.05, 10^{-4}, 10^{-5}\}$, and 128 seeds. Its scaling counterpart keeps the reward law and seed count, uses $T = 2^{36}$, and budgets $\{10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$. Winner means are 0.5, and alternative means are 0.3. The balanced grid has 15 capped outcomes and no wrong commitments; the primary and scaling grids have none of either. All 402 synthetic cells, including the robustness cells below, are in `results/synthetic/cells.csv`.

Conditional on candidate occupancies, independent Bernoulli reward sums are binomial. Coupled probe designs use the common inverse-CDF intersection cells, so shared coordinates have shared rewards. This samples the exact joint phase sufficient-statistic law. Large horizons therefore do not require a chronological walk. The saved integer occupancies, reward sums, ranks, and gate coins determine every reported factual outcome. The independent audit re-evaluates 118080 phase certificates and 430119 stopping branches. It reconstructs all 18432 synthetic outcomes. Separate literal-row checks compare both same-data reduction and cross-position/cross-design moments. These are implementation checks; the conditional sampling construction supplies the distributional identity.

Intervals here use 2000 paired whole-seed bootstrap resamples and are pointwise 95% percentile intervals. Ratios are ratios of mean regrets, recomputed within each resample. For each recorded tape, the supplement also integrates the phase coins analytically: if g_p is a potential phase gate, expected acquisition is $\sum_p |I_p| \prod_{q<p}(1 - g_q)$. At $\rho = 10^{-8}$ the corresponding first-choice/offset acquisition ratios are 3.97, 7.90, and 15.66 for $L = 4, 8, 16$. The main figure reports realized learning runs and their uncertainty. Confidence requirements, dyadic phase lengths, and finite horizons determine the finite-sample ratios; the corollary compares rates up to constants and logarithms.

G.2 DEPENDENCE, HETEROGENEOUS GAPS, AND PROBE DESIGN

The grid has $L = 16$, $d \in \{2, 4, 8\}$, $T = 2^{26} - 1$, and groups $(1, 15)$. It crosses gap-spread factors $1, 4, 16$, shared-shock probabilities $0, 0.8$, three winner/singleton placements, and budgets $10^{-4}, 3 \times 10^{-4}$, with 32 paired seeds per reward law. Each winner has mean 0.8; the fifteen alternative gaps are $0.04R^{j/14}$ for $j = 0, \ldots, 14$ and spread $R \in \{1, 4, 16\}$. The configuration fixes this construction, and tape records save the

exact mean arrays. A shared-shock row uses one uniform variable U and rewards $\mathbf{1}\{U < \mu_{k,a}\}$ at every coordinate. Other rows use independent Bernoulli rewards. This preserves each mean while introducing within-row dependence. The mixture probability is not a correlation coefficient.

Uniform and mixed probing use the same joint reward/probe construction. The placements put winners in singleton groups, alternate their positions, or put the worst candidate in the singleton. Mixed probing reduces mean acquisition in all 108 paired comparisons, by 9.9–49.3%. Mean regret decreases in 100 and increases in eight; the largest increase is 25.5%. All 10368 learner and certificate-only control outcomes commit correctly. The per-placement results identify when increased singleton exposure buys protection cheaply and when it serves a poor candidate more often.

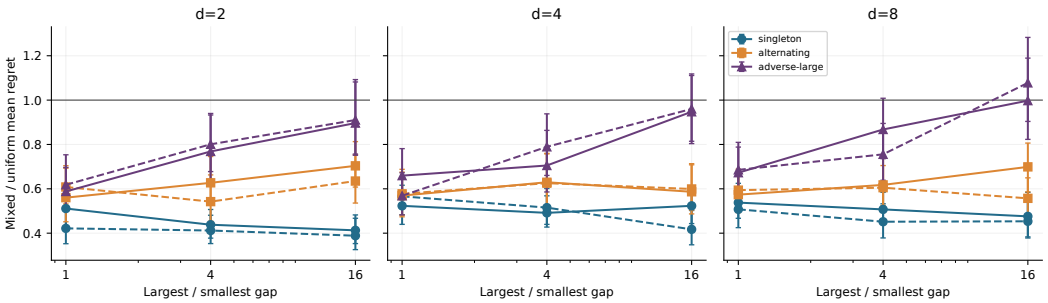

Figure 5: Paired mixed/uniform regret ratios at $\rho = 10^{-4}$ across the dependence and gap-spread grid. Dashed lines use independent rewards; solid lines use shared shocks with probability 0.8. Every placement is shown, with pointwise 95% paired intervals. The raw tables also retain the 3×10^{-4} budget.

G.3 Retained logs and chronological replay

A phase log stores the served candidate indices and integer reward units for each acquired row. Phase boundaries and saved probe, priority, and gate randomness remain available. A deletion creates a new affected-phase payload array with the requested row physically absent. The retained decoder receives these arrays and reward-free policy metadata. It receives no source-user indices, environment seed, full reward table, or cached factual score totals. Full-log replay accumulates all reached retained phases. A second decoder recomputes only the affected phase, using the saved factual public boundaries for the remaining prefix and suffix. It reads $n - 1$ retained rows for a request in a phase of length n, and zero for an inactive request. These logical read counts exclude copying the sanitized array. The guard separately rejects a deleted reward, an unserved coordinate, and a row beyond factual acquisition. Deliberate guard tests are recorded separately from denials made during actual replay.

The independent reference regenerates the exogenous source and probes, replaces the named row by the marker, and enumerates all four endpoint fillings for the two missing rewards. It recomputes the candidate separately at each vertex and uses a scalar log-sum-exp implementation. Agreement compares the commitment round, committed slate, and every public phase boundary. The common saved probes then determine all exploration actions; the committed slate determines the rest of the public action sequence. Omit-only replay is compared with its own marked-table rule. All requests are separate single deletions from the same factual run.

The empirical replay grid uses the three sources below, budgets $\{0.05, 10^{-3}, 10^{-5}\}$, uniform and mixed probes, and four independent seed/placement clusters, giving 72 factual trajectories at $T = 2^{28} - 1$. Cluster r uses singleton positions r and $(r + 1) \bmod 4$. Each reached phase contributes one uniformly sampled row request. Six additional calendar requests are $1, 32768, 131072, 1048576, 16777216, 268435455$. Duplicates remain separately named requests. Requests share their factual seed, so they are not independent trials. The results table separates acquired-feedback erasures from inactive requests after acquisition.

Source	Runs	Actual	Inert	Both replay matches	Denied reads
MovieLens	24	728	26	754/754	0
Jester	24	675	42	717/717	0
KuaiRec	24	521	76	597/597	0

Table 6: Literal empirical replay. Full-log and affected-phase decoders separately match the independent marked-table reference. Actual requests remove acquired feedback; inert requests name a row after acquisition.

A floating-point regression tests acquisition containment at a representable gate boundary. Missing endpoint rewards are inserted into integer reward sums before normalization and inverse-propensity weighting. Adding them after normalization can reverse factual/envelope gate ordering by one unit in the last place. The recorded regression uses $n = 8192$, $T = 65535$, $\rho = 10^{-4}$, denominator 2000, and retained sums $(6756432, 0, 0, 0)$ and $(11105383, 0, 0, 0)$. The corrected factual and envelope gates both equal 0.08097695985276489. The full audit also checks bounded completion vertices, independent integer accumulation, and candidate switches. It records numerical gate differences separately from decisions.

G.4 A STATIONARY ACCESS OBSTRUCTION

This grid uses $d = 2$, $L = 4$, uniform probes, calendar protection, $T = 8191$, and one fixed acquired-feedback deletion at row 2048. Winner means are $0.1 + \Delta$ and other means are 0.1; $\Delta \in \{0.65, 0.70, 0.75\}$ and $\rho \in \{0.02, 0.05, 0.20\}$. The 512 independent seeds share their potential tapes across the nine settings. Every run uses ordinary stationary Bernoulli draws and its original gate coins. All 4608 factual runs commit correctly. Envelope matches every reference from retained feedback. Omit-only requests unavailable feedback on 14 runs across seven settings; all 14 are distinct seed identities. Its denied read is row 4096, after factual commitment at 4095; its full-table reference continues to 8191. Independent literal reconstruction verifies all 8191 actions in all 4608 runs. Envelope commits earlier on five runs.

Of the first positive gates, $3708/4608$ are saturated. A nonempty omit-only continuation interval occurs on $441/4608$ requests. At $\Delta = 0.70$ and $\rho = 0.05$, $4/512$ requests block, or 0.781% (Wilson 95% interval $[0.304, 1.991]\%$). The average conditional interval mass in that cell is 0.465%. The frequency and interval mass describe different random quantities. The complete nine-cell table and each gate/coin realization are saved.

The supplement also supplies deterministic matched-access and earlier-commitment tapes, with independent four-vertex checks. Seventeen predicates verify the specified gates, candidate choices, and availability obstruction. These fixtures test boundary cases independently of the stationary frequencies above.

G.5 ANALYTIC CALIBRATION AUDIT

A finite-binomial calculation checks the raw uniform-probe gate on a specified single-position, four-candidate table. It uses $n = 16384$, $T = 65535$, and $\rho = 10^{-5}$. Conditioning on the target category and integrating the shared coin gives the exact event-probability difference for this construction. If the calendar slope is reused for $G = 1, 2, 4$, the maximum category-wise TV ratios to ρ are $0.868, 1.069, 1.472$. Calibrating the slope to the corresponding group sizes gives $0.868, 0.534, 0.368$. All binomial masses and conditional gate terms are retained, and 216 explicit-score calculations independently check the formula. This analytic fixture is separate from the unknown-gap utility grids.

G.6 PRIVATE ADAPTIVE COMPARATORS

Whole-row privacy uses $\epsilon = 2\,\mathrm{atanh}(\rho)$ and $\epsilon_{\mathrm{pos}} = \epsilon/d$. The public release is the action path. Private statistics, noise draws, and diagnostic records are evaluation artifacts. The methods acquire one reward at every position throughout the horizon. Appendix A.7 proves the present-row/missing-marker calibration.

AdaP-UCB. We implement Algorithm 1 and Equation (7) of Azize and Basu (2022) with $\alpha = 3.1$. Initialization observes every candidate once. Thereafter the selected candidate is played for its cumulative previous observation count, producing batch lengths $1, 1, 2, 4, \ldots$. A completed batch of length n_a supplies its mean plus an independent Laplace draw of scale $1/(\epsilon_{\mathrm{pos}} n_a)$. At one-based decision round u, its unbounded index is

$$\widetilde{\mu}_a + \sqrt{\frac{3.1 \log u}{2n_a}} + \frac{3.1 \log u}{\epsilon_{\mathrm{pos}} n_a}.$$

Only the last completed batch supplies its mean and confidence denominator. Independent fixed candidate ranks resolve ties. Each position has its own noise stream and episode clock; its observations use the same calendar reward rows as the other methods. Truncation at T releases no additional incomplete-batch statistic.

DP-KLUCB. For binary KuaiRec and Bernoulli synthetic rewards, Algorithm 1 and Equation (10) of Azize et al. (2025) use initial batch length one and growth factor two. Completed disjoint batches of lengths $1, 2, 4, \ldots$ contribute their reward sums plus independent $\mathrm{Lap}(1/\epsilon_{\mathrm{pos}})$ draws to a cumulative noisy sum. Divide by the cumulative completed count N_a and clip to $[0, 1]$. For clipped mean p, the index is the largest $q \in [p, 1]$ satisfying

$$\inf_{z \in [0,1]} \{\epsilon_{\mathrm{pos}} |z - p| + \mathrm{kl}(z, q)\} \le \frac{\log u}{N_a}.$$

The solver returns the endpoint $q = 1$ exactly when it is feasible. Otherwise it uses the analytic privacy branch and a monotone Bernoulli-KL bisection. Independent constrained minimization checks the divergence and

inverse. Literal global-clock implementations check episode lengths, selected rewards, privacy noise, horizon prefixes, and tie decisions for both baselines.

The synthetic comparison uses $d \in \{2, 4\}$, $L \in \{8, 16\}$, $T = 2^{20}$, $\rho \in \{0.0003, 0.003, 0.03, 0.3\}$, 32 seeds, and winner/alternative means $0.5/0.25$. Envelope protects $G \in \{1, 2, L\}$; private methods protect every pre-feedback category. All methods share chronological Bernoulli tables. There are 2688 outcomes, including 22 Envelope caps and no wrong commitments. Both directions of the regret crossover remain in the complete tables.

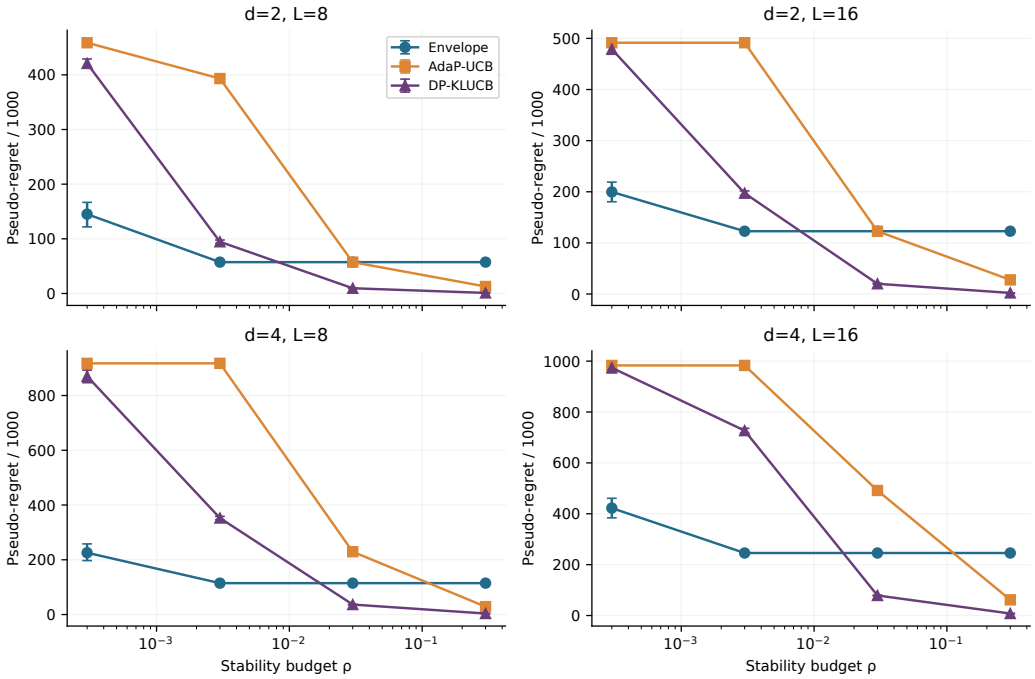

Figure 6: Private-comparator checks at $T = 2^{20}$ with $G = 2$ for Envelope. Whiskers are pointwise 95% whole-seed intervals. The full tables additionally retain $G = 1$ and $G = L$, including protection of each action.

G.7 Preference-derived reward laws and genre categories

MovieLens 1M, the original three-workbook Jester release, and KuaiRec's small matrix supply the reward vectors (Harper and Konstan, 2015; Goldberg et al., 2001; Gao et al., 2022). A user enters the design cohort when the SHA-256 hash of the decimal ID, prefixed by `movielens-design-v1:`, `jester-design-v1:`, or `kuairec-design-v1:`, is divisible by five as a big-endian integer. Jester IDs are one-based after workbook concatenation. Rank items by design-user observation count, breaking ties by increasing item ID, and select the first eight. The first four and remaining four form the two positions. Evaluation uses complete vectors from the other users. Jester and KuaiRec retain at most 2048 users ranked by the outcome-independent prefixes `jester-environment-cap-v1:` and `kuairec-evaluation-cap-v1:`. All complete MovieLens evaluation users remain.

The catalogues are $(2858, 1196, 260, 1210; 480, 2028, 2571, 589)$, $(17, 20, 15, 5; 7, 8, 13, 16)$, and $(103, 109, 120, 128; 136, 139, 142, 145)$. Ratings become $(r - 1)/4$ for MovieLens and $(r + 10)/20$ for Jester. Jester's 99 sentinel marks absence. KuaiRec uses $\mathbf{1}\{\texttt{watch_ratio} \geq 1\}$, which agrees with the duration comparison on every selected record. Exact integer reward denominators are $4, 2000, 1$. The preparation code verifies Jester's hundredth-point rating lattice. Checksums cover the official source archives, source members, user-ID arrays, and prepared reward arrays; selected catalogues are recorded. An independent source reader verifies all 154328 selected reward values, including the metadata check below.

If Y is the evaluation array, round t samples $X_t = Y_{J_t}$ with independent uniform source index J_t. Sampling a whole vector preserves empirical dependence and heterogeneous gaps. Only the two served coordinates reach the learner. This is a stationary semi-bandit law conditional on the selected cohort. The experiments assess regret, collection, and interaction erasure under that law; the confidence intervals measure simulation variability rather than uncertainty about a population of users or deployed recommendation quality. A requested record is one sampled interaction, including when the same source user is sampled again elsewhere in the run.

Source	Complete design	Evaluation	Local gap range	Max. |corr.|
MovieLens	104	486	.0134–.1214	.620
Jester	14723	2048	.0209–.1244	.306
KuaiRec	277	1131	.2140–.5057	.293

Table 7: The fixed complete-vector cohorts. Correlations are off-diagonal properties of their evaluation vectors.

Budgets, horizons, and shared randomness. The ten budgets are $\{0.5, 0.1, 0.05, 0.01, 0.003, 0.001, 0.0003, 0.0001, 0.00003, 0.00001\}$. All three sources use the horizon ladder $T \in \{2^{16}-1, 2^{20}-1, 2^{24}-1, 2^{28}-1\}$ and 32 paired seeds. At placement r, the singleton is candidate r at position one and $(r+1) \bmod 4$ at position two. Mixed probes assign $3/8$ to a singleton and $5/24$ to other candidates, with $\Lambda = 8/3$; uniform probes use $1/4$ and $\Lambda = 4$. Each seed retains all four mixed placements and one uniform/private outcome per method, budget, and horizon. Its source stream, priorities, phase coins, and privacy noise are shared across the budget/horizon ladder. Private outcomes at shorter horizons are exact prefixes of the same chronological private path. Envelope is re-evaluated for each supplied horizon using that horizon's $\log(dLT)$ reserve. Short horizons contain 8806 capped Envelope outcomes; all 4800 Envelope outcomes at $2^{28}-1$ commit correctly.

Private episode observations are exact chronological interval sums from an integer block-prefix index of the source tape. For Envelope, condition on that same source tape. Within each phase, each source vector's count is split over the Cartesian cells of the common probe-CDF partition. One joint multinomial supplies all five designs' occupancies and reward sums. The union of CDF edges is $\{0, 5, 6, 9, 10, 12, 14, 15, 18, 19, 24\}/24$. This integrates only probe order within a fully acquired phase. The separate chronological replay campaign constructs actual retained logs. Literal reductions, exact interval checks, and horizon-prefix checks verify both implementations. An independent Jester pilot at the largest horizon checks all eight coordinates' integer totals and the complete phase occupancy histogram.

Uncertainty and full curves. The four mixed placements are averaged within a seed. Paired percentile bootstrap intervals use 20000 resamples of the 32 whole seeds, preserving every method, budget, horizon, and placement together. Intervals are pointwise 95% intervals. The supplement retains unrounded values, all placements, every cap and wrong-commitment indicator, and every signed paired difference. The following table gives the full main-horizon grid; the horizon figure displays all shorter settings.

Table 8: All ten budgets at $T = 2^{28}-1$. Mean regret per round, with pointwise 95% whole-seed intervals underneath.

ρ	Mixed Envelope	Uniform Envelope	Private AdaP-UCB	DP-KLUCB
MovieLens				
0.5	0.0118 [0.0113, 0.0122]	0.00822 [0.00724, 0.0092]	6.1e-05 [6.02e-05, 6.17e-05]	–
0.1	0.0118 [0.0113, 0.0122]	0.00822 [0.00724, 0.0092]	0.000111 [0.00011, 0.000112]	–
0.05	0.0118 [0.0113, 0.0122]	0.00822 [0.00724, 0.0092]	0.000153 [0.000151, 0.000156]	–
0.01	0.0118 [0.0114, 0.0122]	0.00841 [0.00744, 0.00939]	0.0005 [0.0005, 0.0005]	–
0.003	0.0119 [0.0115, 0.0123]	0.0088 [0.00783, 0.00998]	0.00137 [0.00137, 0.00137]	–
0.001	0.012 [0.0116, 0.0124]	0.00998 [0.00881, 0.011]	0.00442 [0.00438, 0.00445]	–
0.0003	0.0124 [0.0123, 0.0125]	0.0115 [0.0108, 0.0123]	0.011 [0.011, 0.011]	–
0.0001	0.015 [0.0137, 0.0166]	0.0158 [0.0139, 0.018]	0.0332 [0.0328, 0.0338]	–
3e-05	0.0218 [0.0195, 0.024]	0.0309 [0.0254, 0.0364]	0.0618 [0.0618, 0.0618]	–
1e-05	0.055 [0.0446, 0.0658]	0.0657 [0.0544, 0.0771]	0.1 [0.1, 0.1]	–
Jester				
0.5	0.00397 [0.00372, 0.00427]	0.00366 [0.00366, 0.00366]	5.06e-05 [4.9e-05, 5.19e-05]	–
0.1	0.00397 [0.00372, 0.00427]	0.00366 [0.00366, 0.00366]	9.96e-05 [9.86e-05, 0.000101]	–
0.05	0.00397 [0.00372, 0.00427]	0.00366 [0.00366, 0.00366]	0.000134 [0.000131, 0.000137]	–
0.01	0.00405 [0.00378, 0.00438]	0.00366 [0.00366, 0.00366]	0.000409 [0.000404, 0.000416]	–
0.003	0.00437 [0.004, 0.00478]	0.00366 [0.00366, 0.00366]	0.00161 [0.00161, 0.00161]	–
0.001	0.00576 [0.00521, 0.00628]	0.00389 [0.00366, 0.00423]	0.0036 [0.00354, 0.00365]	–

ρ	Mixed Envelope	Uniform Envelope	Private AdaP-UCB	DP-KLUCB
	0.00709	0.00686	0.0129	
0.0003	[0.00675, 0.00732]	[0.00641, 0.00721]	[0.0129, 0.0129]	–
	0.0089	0.00972	0.0333	
0.0001	[0.00796, 0.00995]	[0.00847, 0.011]	[0.0332, 0.0336]	–
	0.0184	0.019	0.0702	
3e-05	[0.0156, 0.0214]	[0.016, 0.022]	[0.0702, 0.0702]	–
	0.0323	0.04	0.117	
1e-05	[0.0263, 0.0384]	[0.0336, 0.0462]	[0.117, 0.117]	–
KuaiRec				
	0.000242	0.000193	1.64e-05	1.46e-06
0.5	[0.000242, 0.000242]	[0.000174, 0.000212]	[1.6e-05, 1.67e-05]	[1.4e-06, 1.52e-06]
	0.000242	0.000204	5.5e-05	6.37e-06
0.1	[0.000242, 0.000242]	[0.000185, 0.000223]	[5.34e-05, 5.65e-05]	[6.16e-06, 6.59e-06]
	0.000242	0.000215	9.1e-05	1.2e-05
0.05	[0.000242, 0.000242]	[0.000196, 0.000231]	[9.1e-05, 9.1e-05]	[1.15e-05, 1.24e-05]
	0.000242	0.000234	0.000367	6.37e-05
0.01	[0.000242, 0.000242]	[0.000223, 0.000242]	[0.000364, 0.000372]	[6.14e-05, 6.6e-05]
	0.000249	0.000242	0.00146	0.000202
0.003	[0.000242, 0.000264]	[0.000227, 0.000261]	[0.00146, 0.00146]	[0.000192, 0.000212]
	0.000399	0.000397	0.00381	0.000627
0.001	[0.000359, 0.000436]	[0.000351, 0.000438]	[0.00376, 0.00386]	[0.000597, 0.000657]
	0.0006	0.00068	0.0116	0.00211
0.0003	[0.00052, 0.000685]	[0.000589, 0.000771]	[0.0116, 0.0116]	[0.00202, 0.00221]
	0.00122	0.00175	0.0417	0.00608
0.0001	[0.00101, 0.00144]	[0.00147, 0.00207]	[0.0415, 0.0419]	[0.00583, 0.00635]
	0.00344	0.00441	0.0945	0.0202
3e-05	[0.00267, 0.00427]	[0.00348, 0.00535]	[0.0932, 0.0959]	[0.0194, 0.021]
	0.00859	0.0131	0.302	0.0575
1e-05	[0.00673, 0.0105]	[0.0102, 0.0161]	[0.302, 0.302]	[0.0552, 0.0599]

Metadata-defined MovieLens groups. The genre check assigns each movie its first genre under a fixed alphabetical MovieLens taxonomy. At each position, it takes the highest-availability unused movie, the highest-availability unused movie with a different primary genre, and the two highest-availability remaining movies. All rankings use design availability and item-ID ties. Selection is completed before evaluation rewards are inspected. This gives $(2858, 1196, 260, 1210; 480, 2028, 2571, 593)$, with Comedy/Action groups at position one and Drama/Action groups at position two. Each has a singleton and three Action candidates. There are 93 complete design users and 429 evaluation users. The fixed genre policy replaces the four artificial rotations in this check. The 32 seeds and budgets $\{0.05, 10^{-3}, 10^{-5}\}$ yield 288 method outcomes. At 10^{-5}, mixed probing reduces regret against uniform probing by 31.7% ($[20.3, 42.6]\%$) and against private AdaP-UCB by 82.8% ($[79.1, 86.1]\%$). At 0.05, private AdaP-UCB has lower regret; the mixed/uniform difference has an interval containing zero. All 192 Envelope runs commit correctly.

ρ	Method	Regret per round [95% CI]	Acquired fraction	Caps
1e-05	Private AdaP-UCB	0.0979 [0.0979, 0.0979]	1.0000	0
1e-05	Envelope mixed	0.01684 [0.01358, 0.02045]	0.1914	0
1e-05	Envelope uniform	0.02467 [0.01979, 0.02974]	0.2520	0
0.001	Private AdaP-UCB	0.004016 [0.004016, 0.004016]	1.0000	0
0.001	Envelope mixed	0.002707 [0.002621, 0.00275]	0.0308	0
0.001	Envelope uniform	0.002916 [0.002772, 0.003059]	0.0298	0
0.05	Private AdaP-UCB	0.0001375 [0.0001337, 0.0001412]	1.0000	0
0.05	Envelope mixed	0.002578 [0.002406, 0.002707]	0.0293	0
0.05	Envelope uniform	0.002534 [0.002294, 0.002773]	0.0259	0

Table 9: The complete fixed-genre check. The same category partition and selected cohort are used for all three methods.

G.8 Private-index diagnostics and collection costs

For every diagnostic episode, the independent audit reconstructs the index from its empirical mean, cached Laplace perturbation, sampling bonus, and privacy bonus. Final regret is independently recomputed from arm counts and true mean gaps. At $\rho = 10^{-5}$, MovieLens and Jester each have one distinct final allocation and pseudo-regret across 32 seeds, but 32 distinct episode orders. On the factual decision histories, removing only the empirical-mean component changes $567/6656$ and $507/6656$ decisions, respectively. These are local index counterfactuals.

Full reruns hold priorities and privacy-noise streams fixed while replacing rewards by their coordinate means, replacing all rewards by $1/2$, or cyclically permuting candidate coordinates. Three budgets, three sources,

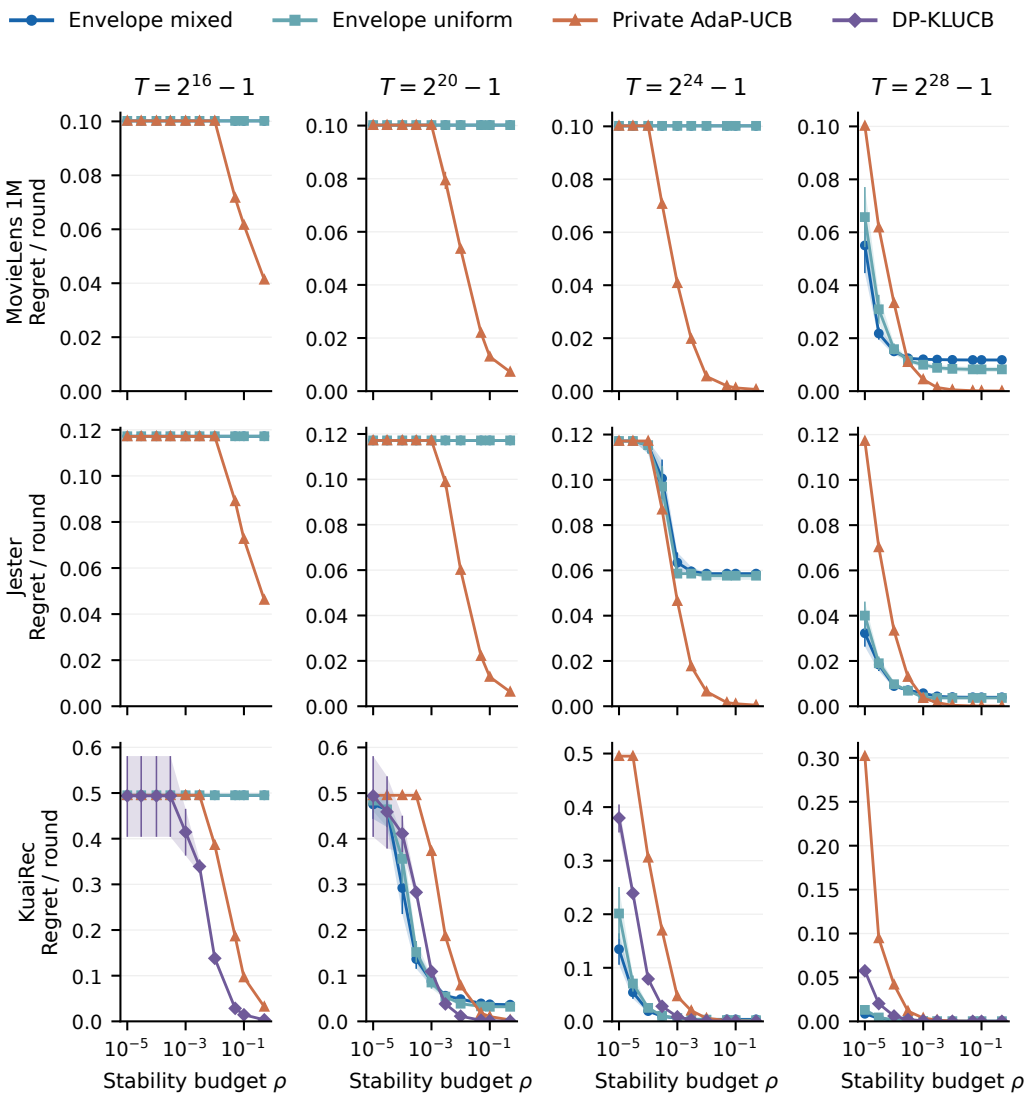

Figure 7: Fixed-horizon sensitivity, using shared source-tape prefixes and each horizon's own Envelope confidence reserve. All capped runs remain in the plotted means.

three interventions, and eight fixed seeds give 216 runs. At the strictest budget, cyclic permutation changes allocations in all eight tested seeds for each source. Constant coordinate means leave the strict-budget actions unchanged in those tested seeds. At KuaiRec's $\rho = 0.05$, constant means change all eight episode orders while preserving all final allocations. The full tables distinguish allocation, ordering, reward, and noise effects. Identical pseudo-regret can therefore coexist with reward-sensitive decisions. Bonus sizes, phase schedules, and the chosen horizon all affect the comparison, alongside the different stability and collection contracts.

For applications with a cost $c \geq 0$ per acquired feedback row, the same records support the sensitivity calculation $J_c = R_T + c\,\mathbb{E}[N_{\text{acq}}]$. Against a private learner that acquires all T rows, the signed break-even cost is

$$c_\star = \frac{R_{\text{mixed}} - R_{\text{private}}}{T - \mathbb{E}[N_{\text{mixed}}]}.$$

A negative value means mixed Envelope is better for every nonnegative c. A positive value is the collection price above which its acquisition saving compensates for higher regret. At $\rho = 0.05$, the thresholds against AdaP-UCB are 0.01317 for MovieLens, 0.00397 for Jester, and 0.000151 for KuaiRec, in reward units per acquired interaction. Their paired 95% intervals are $[0.012535, 0.013744]$, $[0.003706, 0.004297]$, and $[0.000150720, 0.000150890]$, respectively. All other cells are saved. No service price is assumed or estimated. If a resample has zero acquisition saving, its undefined threshold is counted explicitly rather than substituted.

G.9 NGUYEN-TANG–ARORA COMMON-BANDIT COMPARISON

We implement the one-state, one-step specialization of Nguyen-Tang and Arora (2026), using its Gaussian binary tree, maximal coupling, confidence constants, and environment oracle. The common reward law has four Bernoulli arms, with winner mean $0.1 + \Delta$ and other means 0.1. The grid is $T \in \{8192, 32768\}$, $\Delta \in \{0.3, 0.6\}$, and $\rho \in \{0.1, 0.5\}$, with calendar protection ($G = 1$). Each cell has 16 independent seed clusters, each containing all four winner positions. All 512 paired instances are retained.

The source specialization keeps all twelve visit, transition, and reward coordinates, uses sensitivity $B = \sqrt{3}$, and sets $\sigma = \sqrt{3 \log_2 T}/(\sqrt{2}\rho)$ with confidence parameter 0.05. The finite-horizon terminal value is $V_{H+1} = 0$, as required by its value recursion; this corrects the displayed initialization $Q_{H+1} = H$ in Algorithm 4, line 1. Exact maximizer ties use uniform random draws. Upon rejection in Gaussian coupling, every subsequent postorder node receives fresh noise, including ancestors completed at the same calendar boundary. The implementation records these conventions.

Each deletion names one-based row $T/4$. The source dummy has reward zero and retains visit and transition indicators. Its reconstruction may read the old leaf and query the full environment. Envelope uses its marker rule and a separately sanitized scalar log. The comparison measures factual regret under the same calendar budget; reconstruction is judged against each learner's own missing-record rule.

T	Δ	ρ	Envelope	NT	Difference [95% interval]
8192	0.3	0.1	1843.2	1843.2	0.0 [0.0, 0.0]
8192	0.3	0.5	1843.2	1843.2	0.0 [0.0, 0.0]
8192	0.6	0.1	3571.1	3686.4	−115.3 [−229.2, −28.9]
8192	0.6	0.5	3571.1	3686.4	−115.3 [−229.3, −28.9]
32768	0.3	0.1	7315.0	7372.8	−57.8 [−173.0, −0.2]
32768	0.3	0.5	7315.0	5846.5	1468.5 [1350.2, 1531.6]
32768	0.6	0.1	3685.9	14745.6	−11059.6 [−11059.6, −11059.6]
32768	0.6	0.5	3685.9	8527.9	−4842.0 [−4853.9, −4830.7]

Table 10: All common-bandit cells. Differences are Envelope minus the stated Nguyen-Tang–Arora specialization. Pointwise intervals use 20000 paired bootstrap resamples of the 16 seed clusters, preserving the four winner positions.

At $T = 32768$, $\rho = 0.5$, the adaptive source learner has lower regret at gap 0.3; Envelope has lower regret at gap 0.6. All source scores remain clipped at one in the other six cells. The full winner-position strata are retained, including constant cell averages. Envelope is capped in the 128 runs with $T = 8192$, $\Delta = 0.3$ and commits correctly in the other 384.

All 512 Envelope reconstructions match their references: 384 requests erase acquired feedback and 128 are inactive. The source learner acquires throughout. The next table separates retained-log reads from oracle reads and fresh counterfactual coordinates. Work units differ: a source coupling test compares twelve-coordinate Gaussian vectors, whereas an Envelope read is a retained scalar. The source method's old-leaf and dummy reads are additional fixed operations.

(T, Δ, ρ)	Env. active	Env. reads	NT tests	NT oracle	NT fresh
(8192, 0.3, 0.1)	64/64	2047	14.00	0	0.00
(8192, 0.3, 0.5)	64/64	2047	13.33	672	0.00
(8192, 0.6, 0.1)	64/64	2047	14.00	0	0.00
(8192, 0.6, 0.5)	64/64	2047	12.97	960	0.00
(32768, 0.3, 0.1)	64/64	8191	15.95	384	0.00
(32768, 0.3, 0.5)	64/64	8191	15.14	2560	981.44
(32768, 0.6, 0.1)	0/64	0	15.95	384	0.00
(32768, 0.6, 0.5)	0/64	0	14.97	3328	339.64

Table 11: Logical operations averaged over every request in a cell. Envelope makes no deleted-payload or fresh-environment query. Oracle access is part of the source method's reconstruction contract.

The source kernel's Gaussian density-ratio test retains the old sample on acceptance; reflection $y \mapsto m + m' - y$ gives the residual new law on rejection. Fresh later nodes generate the updated suffix. Independent checks reconstruct 20971520 leaves and 20970496 parents; 192 small-tree comparisons exercise acceptance, rejection, informative actions, and same-boundary ancestors. Envelope's decoder receives the sanitized scalar log and summaries of unchanged phases. The requested scalar is overwritten with the marker and its affected-phase aggregate is discarded before decoding.