NON-STATIONARY BANDITS WITH KNAPSACKS WITHOUT VARIATION KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

In pay-per-click advertising, dynamic pricing, and crowdsourcing, each pull spends a hard budget that is not refilled, and the means can change over time. A better arm may appear only after the budget that would have paid for it has already been spent. If the variation V is known, existing work achieves a regret bound that depends on V. However, V is usually unknown in advance, so the change and the rewards must be learned on the same hard budget. Throwing away old pulls does not return the budget already spent. We prove that an extra gap Γ_b is necessary even if V is given. Two worlds share the first half of the rounds and the same V, and then move in opposite directions. Any algorithm pays Γ_b on one of them, since being told V does not say whether to spend early or to save the budget for later. We design Vector-MASTER, which is not told V and spends one shared budget. For constant b its regret is $\widetilde{O}\big(\sqrt{mT} + m^{1/3}(V_1 + dV_2)^{1/3}T^{2/3}\big) + \Gamma_b$. The added term is this necessary gap, and nothing larger is charged when the gap is zero.

1 INTRODUCTION

Bandits with Knapsacks (BwK) charges every pull to a budget that is not refilled (Badanidiyuru et al., 2013; Agrawal et al., 2016). The means can also change over the T rounds, as in pay-per-click advertising, dynamic pricing, and crowdsourcing (Besbes et al., 2014; Liu et al., 2022). A better arm can arrive only after the budget that would have paid for it is gone.

Write V for how much the means change. If V is known, existing methods can already track the means (Liu et al., 2022). V is rarely known. Without a budget, old pulls can be discarded. Here, discarding them does not return what was spent, and every pull used to check the old means spends the same budget.

Even if V is known, two worlds can share the first half of the rounds and the same V, and then the same arm becomes better in one and worse in the other. A policy that knows which world it is in spends early or saves the budget for later. An algorithm must spend before the worlds can be told apart, and spent budget cannot be moved. Spending in the first half is then a mistake if the arm later becomes better, and saving is a mistake if the arm later becomes worse. Being told V does not say which mistake to avoid, because both worlds have the same V. We ask how large the extra regret is when V is not known, and whether an algorithm can pay only what is necessary.

When V is not known, how much extra regret must an algorithm pay, and can an algorithm pay only that necessary cost?

We make three contributions.

Vector-MASTER. We give an algorithm that spends one shared budget and is not told V. It keeps a long run of pulls and also starts shorter runs at random. A short run can still pull an arm the long run has dropped, and it can see that the cost has changed. If the recent pulls disagree with the long run, the old counts are dropped and the budget already spent stays spent. On a hidden-arm change and on a change that swaps which arm is expensive, the algorithm moves to the new arm and a method that never drops its counts does not.

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.

Tracking the change without being told V. The regret splits into a tracking term and the extra cost below. For constant b, and without being told V, the regret is $\widetilde{O}\big(\sqrt{mT} + m^{1/3}(V_1 + dV_2)^{1/3}T^{2/3}\big) + \Gamma_b$. The $\widetilde{O}$ piece is the tracking term, the same leading terms a method given V already has. One unit of budget is held back so that every sample path remains feasible. That piece does not include the cost of having spent too early or too late.

The extra cost is necessary, and the algorithm pays only that cost. On the two worlds above, any algorithm, even one that is told V, has regret at least Γ_b on the worse world. Telling it V does not help, because both worlds have the same V. Γ_b grows linearly with T whenever the late change stays a constant, while the tracking term on these worlds stays of order $T^{2/3}$. The regret of Vector-MASTER is the tracking term plus Γ_b. The bound does not replace Γ_b by a coarser deviation: that deviation can grow linearly with T on instances where $\Gamma_b = 0$, because every round's best decision is already best for the whole budget.

2 Related work

Table 1 places the closest methods by whether V is given and whether the budget is hard.

Table 1: Closest methods. Known V means a bound on the change is given to the algorithm. Hard budget means every pull spends one shared budget.

Reference	Known V	Hard budget	Regret
Cheung et al. (2019)	×	×	$\widetilde{O}((V+1)^{1/4}T^{3/4})$
Wei & Luo (2021)	×	×	$\widetilde{O}(\min\{\sqrt{LT}, V^{1/3}T^{2/3}\})$
Liu et al. (2022)	✓	✓	$\widetilde{O}(\sqrt{mT}/b + m^{1/3}(V_1 + dV_2)^{1/3}T^{2/3} + W)$
Zhang & Cheung (2024)	×	✓	$O(\log(\eta_{\max}/\eta_{\min}))$ competitive
This paper	×	✓	$\widetilde{O}(\sqrt{mT} + m^{1/3}(V_1 + dV_2)^{1/3}T^{2/3}) + \Gamma_b$

Hard-budget BwK. Stochastic BwK introduced exploration with a non-replenishing budget (Badanidiyuru et al., 2013; 2018). Later work added concave rewards, contexts, and primal-dual prices (Agrawal & Devanur, 2014; Agrawal et al., 2016; Li et al., 2021; Kesselheim & Singla, 2020), and adversarial or best-of-both-worlds costs (Immorlica et al., 2019; Rangi et al., 2019; Castiglioni et al., 2022; Sankararaman & Slivkins, 2021). These results do not study unknown gradual change against a dynamic policy that shares one budget. The closest hard-budget result is Liu et al. (2022). Their comparator and round-dependent observation model are the ones we use. Their regret grows with the total amount of change, and their algorithm is given an upper bound on that amount. Nearby models replace this with other structure: a change-point assumption and a competitive ratio (Zhang & Cheung, 2024), demand advice (Lyu & Cheung, 2023), or a bounded modeling error (Fikioris & Tardos, 2023). The gap is arbitrary drift with a hard budget and no variation input. We keep their comparator and do not take the variation as input. Their $W_1 + \bar{q}W_2$ term can be large when every local LP is already globally optimal. The necessary object is Γ_b: any learner already pays it on a pair with constant variation, even if V_1 and V_2 are given.

Prior-free non-stationary learning. The stationary baseline is UCB and the non-stochastic baseline is EXP3 (Auer et al., 2002a;b). Discounting, sliding windows, and change detection trade old data against sample size (Garivier & Moulines, 2011; Besbes et al., 2014; Auer et al., 2019; Chen et al., 2019; Faury et al., 2021). When an upper bound on total change is known, this trade-off is explicit (Besbes et al., 2015; 2014; Russac et al., 2019). Treating each window length as an arm of a second bandit removes that bound, but adds a $T^{3/4}$ cost even when the means never change (Cheung et al., 2019). The closest prior-free method is MASTER (Wei & Luo, 2021). It matches the best window with randomized replays and no variation input. Whether that black-box guarantee extends to reinforcement learning is separate (Gerogiannis et al., 2025). The gap is that MASTER has no budget. A replay there can restart for free. In BwK the same restart spends the shared budget, and a private remainder for each length would make the comparison with one global LP false. We keep one budget and watch unused arms and consumption as well as reward.

Other resource models. Replenishable inventories and general constraints (Bernasconi et al., 2024b; 2025) and a different data model (He & Mintz, 2024; Guo & Liu, 2025) change the resource or the observation process. Corralling several BwK learners (Agarwal et al., 2017; Pacchiano et al., 2020) again gives each length a private remainder. Soft constraints measure violation rather than a finite remaining budget (Bernasconi et al., 2024a). The gap is a finite remaining budget, not a penalty. We keep one hard budget S.

3 Problem setting

There are m arms and d resources, including a known null arm with zero reward and consumption. Each resource has budget $B = bT$, where $b \in (0, 1]$ and $B \geq 2$. Before play, the environment fixes a possibly time-varying distribution $\mathcal{P}_t$ over the complete reward–consumption table. Tables are sampled independently across rounds, have entries in $[0, 1]$, and may have within-round correlations. Let $\mu_{t,i}$ and $C_{t,j,i}$ denote the mean reward and resource-j consumption of arm i, forming $\mu_t \in [0, 1]^m$ and $C_t \in [0, 1]^{d\times m}$. The learner observes only the reward and consumption of its selected arm.

Write i_t for the chosen arm in round t, and write $r_{t,i}$ and $c_{t,j,i}$ for the realized reward and resource-j consumption of arm i. Play stops at the first time any resource exceeds B:

$$\tau_B = \min\left(\left\{t : \exists j,\ \sum_{s=1}^{t} c_{s,j,i_s} > B\right\} \cup \{T + 1\}\right).$$

Following Liu et al. (2022), $\mathrm{OPT}(T)$ is the expected reward of a dynamic policy that knows every $\mathcal{P}_t$, may use a different arm mixture each round, and obeys the same budgets and stopping rule. Regret is

$$\mathrm{Reg}_T = \mathrm{OPT}(T) - \mathbb{E}\left[\sum_{t<\tau_B} r_{t,i_t}\right].$$

The learner sees only the selected arm, and it never receives the means or the change totals below. The comparator may save a unit now and spend it later, when the same inventory buys more reward.

The local path lengths are

$$V_1 = \sum_{t<T} \|\mu_{t+1} - \mu_t\|_\infty, \qquad V_{2,j} = \sum_{t<T} \|C_{t+1,j} - C_{t,j}\|_\infty, \quad V_2 = \max_j V_{2,j}.$$

For multi-resource charging we also use the additive path length

$$\widetilde{V}_2 = \sum_{t<T} \max_j \|C_{t+1,j} - C_{t,j}\|_\infty \leq dV_2.$$

V_2 sums one resource at a time. $\widetilde{V}_2$ lets the largest resource change from round to round. None of these quantities is given to the algorithm. They tell us how long a local estimator stays valid.

For the simplex Δ_m, the per-round LP value is

$$\phi_t(a) = \max_{x\in\Delta_m} \{\mu_t^\top x : C_t x \leq a\mathbf{1}\}.$$

This is the best expected reward in round t under a per-round budget a. The adaptive term in Section 5 competes with $\sum_t \phi_t(b)$.

Example 1 (An unused arm, and a cheap arm)**.** Two arms, consumption zero, rewards $(1/2, 0)$ then $(1/2, 1)$. A long window drops arm 2 and never sees the jump, so the regret is linear for $V_1 = 1$. Both arms reward one, consumptions $(b/2, b)$ then $(1, b)$. After the jump a reward-only check stays silent while the shared budget burns.

Later checks must look at unused arms and at consumption. A window on $\sum_t \phi_t(b)$ still misses a planner that moves budget across rounds; that gap is Theorem 3.

4 Vector-MASTER

This section specifies Vector-MASTER. Because a replay in BwK is not free, we first keep one shared budget.

4.1 ONE SHARED BUDGET

In an unconstrained bandit, a replay can restart its counts. In BwK that restart does not refund what was already spent. If each history length kept a private budget, the same inventory would be counted more than once, and the comparison with one global LP would be false. If the budget were split, a remainder that one length never uses could not be spent by another. The global LP we compare with in Section 5 has one budget. A split would give each length a private remainder, and that comparison would no longer hold.

So we keep one consumption vector $S \in \mathbb{R}^d_+$. Every pull adds to S once. Counts may live at several history lengths; the budget does not. A long run alone will stop pulling an arm that looks poor, and it will keep a stale cost. Only a shorter run still tries that arm and starts its cost estimate over. Those pulls spend the same S. A short run is how the algorithm notices a change, not a second budget.

Each learner plans to spend at the per-round rate

$$b' = \frac{B-1}{T} = b - \frac{1}{T}.$$

The missing unit is a reserve for a last pull. After any coordinate of S exceeds $B-1$, we play the null arm.

4.2 THE BASE LEARNER

An instance is one UCB-BwK counter. It stores only the rounds it played, and it treats those rounds as if the means had not moved. It first tries every real arm once. Then it takes an optimistic reward and a pessimistic consumption, and it solves the same per-round LP a stationary learner would solve. That LP does not ask whether the samples are still valid.

Write R_t and Z_t for the realized reward and consumption vector of the arm played in round t. These are the r_{t,i_t} and $c_{t,:,i_t}$ of Section 3. The acting instance stores counts $N_{a,i}$ and empirical means $\widehat{\mu}_{a,i}, \widehat{C}_{a,j,i}$. For failure probability δ, define

$$\ell = \log \frac{24mdT^4(\lfloor \log_2 T \rfloor + 1)^2}{\delta}, \qquad \beta_{a,i} = \min\left\{1, \sqrt{\frac{2\ell}{N_{a,i}+1}}\right\},$$

and the reward UCB and consumption LCB

$$U_{a,i} = \min\{1, \widehat{\mu}_{a,i} + \beta_{a,i}\}, \qquad L_{a,j,i} = \max\{0, \widehat{C}_{a,j,i} - \beta_{a,i}\}.$$

Empirical means start at zero, and the null arm's bounds stay zero. On its first $m-1$ active rounds the instance plays each non-null arm once, in random order. For a forced pull of arm i, set $x_t = e_i$, $\widetilde{g}_t = U_{a,i}$, and $\widetilde{h}_t = L_{a,:,i}$. After that, with U_t, L_t the current bounds, solve

$$x_t \in \arg\max_{x \in \Delta_m} \{U_t^\top x : L_t x \le b'\mathbf{1}\}. \tag{2}$$

and output

$$\widetilde{g}_t = U_t^\top x_t, \qquad \widetilde{h}_{t,j} = L_{t,j}^\top x_t \le b'. \tag{3}$$

4.3 WHAT THE TESTS CHECK

Example 1 says when to throw the history away. A finished short instance can pull an arm the long instance has dropped. If that short instance then looks better, the long counts are stale. That is T1. We may not want to wait for a finished replay. If the current block has been too optimistic on reward, or too cheap on consumption, we drop the counts now. Those are T2R and T2C.

Write $J_a = [s_a, e_a]$ for the calendar interval the short instance was planned to cover. T1 compares that finished instance with the long instance on J_a. The two need not have spent the same amount of S. If we compared raw reward, a short instance that only spent more would look better, and we would restart on luck. So we first fix one dual price for the interval and turn extra consumption into reward units. After that conversion, T1 asks whether the short instance earned more in priced reward.

Write $\widetilde{g}_t, \widetilde{h}_t$ for the acting instance's predictions, and write $\mathcal{U}_t = \min_{s \le t} \widetilde{g}_s$ on the current block. The threshold $\widehat{\rho}(n)$ is of the same order as the usual width of a window of length n; the formula is in Appendix D. A new epoch starts if any test fires:

$$\frac{1}{|J_a|} \sum_{\substack{t \in J_a \\ t_0+n-1}} R_t \qquad \ge \mathcal{U}_{e_a} + K_1 \widehat{\rho}(|J_a|) \qquad \text{for a completed candidate } a \qquad \text{(T1)}$$

$$\frac{1}{n} \sum_{\substack{t=t_0 \\ t_0+n-1}} (\widetilde{g}_t - R_t) \qquad \ge K_R \widehat{\rho}(n) \qquad \text{(T2R)}$$

$$\frac{1}{nb'} \sum_{\substack{t=t_0 \\ t_0+n-1}} (Z_{t,j} - \widetilde{h}_{t,j}) \qquad \ge K_C \widehat{\rho}(n) \qquad \text{for some } j \qquad \text{(T2C)}$$

T1 is checked when a short instance reaches the end of J_a. T2R and T2C are checked after every prefix of the current block. The constants K_1, K_R, K_C are fixed in Appendix D.

4.4 Epochs and replays

A test that fires starts a new epoch: all counters are forgotten, and S is left as it is. Inside an epoch a long instance tries to play for a long stretch, as a window would. That is not enough. The long instance will drop arms, and it will keep a stale consumption mean. So each block also tosses coins for shorter instances. When a shorter one is present, it plays instead of the long one. Those are the only rounds that can still pull a dropped arm, and the only rounds that see consumption with a fresh counter.

An order-n block starts at time s and covers $[s, \min\{s + 2^n - 1, T\}]$. The block always keeps the long instance. For each $k < n$, every full length-2^k cell is a candidate. Before the block starts, each candidate is kept with probability

$$p_{n,k}^{(\gamma)} = \gamma \frac{\rho(2^n)}{\rho(2^k)}, \qquad \gamma \in (0, 1].$$

The theorem uses $\gamma = 1$. The coins are tossed before the block outcomes are seen. Write $\rho(s) = C(s)/s$, with $\rho(s) = \min\{c_1/\sqrt{s} + c_2/s, c_3\}$; the constants are in Appendix E.

The owner of a round is the shortest scheduled instance that covers it. A shorter instance pauses a longer one. Each instance updates only on the rounds it owns. T1 scores every reward on J_a: the question is whether that calendar stretch was stale. The remaining schedule details are in Appendix A.

The same S is why a replay is not free. A short instance that owns a round spends the shared budget on its own forced pulls and its own LP. Smaller γ keeps fewer short instances, so the long instance owns more rounds, but a hidden arm can be missed.

Algorithm 1: Vector-MASTER

Input: horizon T, budget B, confidence δ, replay factor γ.
Initialize: $t \leftarrow 1$, global consumption $S \leftarrow 0$.

1. **while** $t \le T$ and $\max_j S_j \le B - 1$ **do**:
   (a) Start a fresh epoch. For block orders $n = 0, 1, \ldots$, until a restart or $t > T$:
   (b) Create the mandatory order-n instance and independently schedule each shorter candidate with probability $p_{n,k}^{(\gamma)}$.
   (c) For each round in the block, let the smallest-order scheduled instance a_t own the round. Obtain x_t by forced initialization or LP (2), sample $i_t \sim x_t$, and observe (R_t, Z_t).
   (d) Update only a_t's statistics and set $S \leftarrow S + Z_t$. If $\max_j S_j > B - 1$, exit both loops.
   (e) Evaluate T2R/T2C on the current prefix and T1 when a candidate completes. If any test fires, discard all instances and start the next epoch, leaving S unchanged.
2. **end while**. Play the null arm through round T.

5 MAIN RESULTS

Write $\Delta = V_1 + \widetilde{V}_2/b'$ and write $\widetilde{O}$ for hiding only log factors. A window of length h that knows Δ faces the usual bias–variance trade-off

$$(1 + b^{-1})\Big(\sqrt{m}\,T/\sqrt{h} + mT/h\Big) \quad \text{against} \quad h\Delta.$$

The best such window gives a square-root term when the means do not change, and a $T^{2/3}\Delta^{1/3}$ term when they do (Liu et al., 2022). This is the rate we want without seeing Δ.

Write Γ_b for the gap between a sequence of per-round LPs and one global budget. A local method spends at rate about b in every round. A global planner can save a unit now and spend it later, when the same unit buys more reward. Write $\bar{\mu} = T^{-1}\sum_t \mu_t$, $\bar{C} = T^{-1}\sum_t C_t$, and $\|A\|_1 = \max_{i\in[m]} \sum_{j=1}^{d} |A_{j,i}|$. The deviations from the time-average environment are

$$W_1 = \sum_{t=1}^{T} \|\mu_t - \bar{\mu}\|_\infty, \qquad W_2 = \sum_{t=1}^{T} \|C_t - \bar{C}\|_1.$$

These sums can be large even when consecutive-round changes V_1 and V_2 are small. Let q^* and q_t^* be optimal dual prices for the time-average LP and the round-t LP, and set $\bar{q} = \max\{\|q^*\|_\infty, \max_t \|q_t^*\|_\infty\} \le 1/b$. The global fluid LP chooses a mixture x_t in each round, subject to the total expected budget:

$$\begin{aligned} \mathrm{LP}_{\mathrm{glob}}(B) = \max_{x_t \in \Delta_m} & \sum_t \mu_t^\top x_t \\ \text{s.t.} & \sum_t C_t x_t \le bT\mathbf{1}. \end{aligned}$$

Putting the per-round LP solutions together is feasible for the global LP, so

$$\Gamma_b := \mathrm{LP}_{\mathrm{glob}}(B) - \sum_{t=1}^{T} \phi_t(b) \ge 0.$$

The same comparison as Liu et al. (2022) then gives

$$\begin{aligned} \mathrm{OPT}(T) &\le \mathrm{LP}_{\mathrm{glob}}(B+1) \le \mathrm{LP}_{\mathrm{glob}}(B) + \frac{1}{b} \\ &\le \sum_{t=1}^{T} \phi_t(b) + \Gamma_b + \frac{1}{b}. \end{aligned} \tag{1}$$

Lemma 17 shows $\Gamma_b \le W_1 + 2\bar{q}W_2$.

Theorem 1 (Prior-free dynamic regret and pathwise feasibility)**.** *Under the model of Section 3, Vector-MASTER receives none of V_1, V_2, W_1, W_2, uses $\gamma = 1$, and satisfies every resource budget on every sample path. For every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{OPT}(T) - \sum_{t<\tau_B} R_t \le \widetilde{O}(\mathfrak{R}_\Delta) + \Gamma_b + O(b^{-1}). \tag{5}$$

Here $b' = b - 1/T$, $\Delta = V_1 + \widetilde{V}_2/b'$, $\bar{q} \le 1/b$, and $\mathfrak{R}_\Delta$ is the adaptive term: it is of order $\sqrt{T}$ when the means do not change, and of order $T^{2/3}\Delta^{1/3}$ when they do. The explicit polynomial factors in m, d, b and the restart count are in Appendix E. Setting $\delta = T^{-2}$ in the algorithm gives

$$\mathrm{Reg}_T \le \widetilde{O}(\mathfrak{R}_\Delta) + \Gamma_b + O(b^{-1}) + T^{-1}. \tag{6b}$$

The schedule draws an instance before the current outcome is seen, so ordinary confidence bounds still hold on the rounds it owns. A shorter instance can pull an arm the long instance has dropped, and it can see that consumption has moved. We fix one dual price and turn extra spending into reward

units, then run T1. A short instance that only spends more does not force a restart. T2R and T2C watch the current block: if the long instance is already too optimistic on reward or on consumption, we do not wait for a finished replay. If a test fires, the variation on that epoch cannot have been too small: otherwise the confidence widths would have kept the test below its threshold. Adding those payments over epochs recovers the same bias–variance sum as the best window of Δ. We never see Δ. After that sum, Γ_b is not a window error. It is the gap between local LPs and one global budget, plus the one-unit reserve. It is zero if consumption is always zero, or if the per-round LP solutions are also optimal for the global LP. It can be large when V_1 and V_2 are small but the means wander far from their time averages, because W_1 and W_2 measure that deviation, not consecutive-round change.

The two terms answer different questions. The tracking term is the price of not being told V, and it matches a method that is told V. The extra term Γ_b is the price of a budget that cannot be moved after it is spent. The lower bound shows that this price remains even when V is told. On the two worlds used there, the tracking term is only of order $T^{2/3}$, so it does not absorb Γ_b.

Corollary 2 (Rate under a constant budget fraction)**.** *If $b \geq b_0 > 0$, m, d are constant or grow only polylogarithmically with T, and $V_1, V_2 = o(T)$ and $\Gamma_b = o(T)$, the leading large-T terms are*

$$\mathrm{Reg}_T \leq \widetilde{O}\Big(\sqrt{mT} + m^{1/3}(V_1 + dV_2)^{1/3}T^{2/3}\Big) + \Gamma_b. \tag{6a}$$

If reward and consumption means never change, so all four change measures are zero, the bound is $\widetilde{O}(\sqrt{mT}/b + m^2/b)$. This display is sublinear only if $\widetilde{O}(\mathfrak{R}_\Delta) = o(T)$ as well. If m or d grows, the extra m, d factors in the appendix apply as well.

When b is a constant, the adaptive term is of square-root order if the means do not change, and of order $T^{2/3}\Delta^{1/3}$ if they do. The square-root term is ordinary estimation of rewards and consumptions. The $T^{2/3}$ term balances old data against confidence width, and it is the same local-drift rate as a window that knows Δ. If all four change measures are zero, the remaining cost is $\widetilde{O}(\sqrt{mT}/b)$, plus an m^2/b term from probing every non-null arm in each fresh instance. The lemma bound $\Gamma_b \leq W_1 + 2\bar{q}W_2$ can be loose (Proposition 4). Sublinear regret still needs $\Gamma_b = o(T)$ as well as $\widetilde{O}(\mathfrak{R}_\Delta) = o(T)$.

Theorem 3 (Regret lower bound)**.** *Let $T \geq 4$ be even, $b \in (0, 1/2]$, and $\eta \in (0, 1/2]$. There exist two environments E_+ and E_-, each with one resource, one real arm, $V_1 = \eta$, $V_2 = 0$, and $\Gamma_b = \eta bT/2$, such that any policy—even one that is told V_1 and V_2—satisfies*

$$\max\big\{\mathrm{Reg}_T(E_+), \mathrm{Reg}_T(E_-)\big\} \geq \Gamma_b = \frac{\eta bT}{2}. \tag{6c}$$

The two environments share the same law on the first $T/2$ rounds. On both, $\Delta = \eta$, so the adaptive term in (5) is $\widetilde{O}(\eta^{1/3}T^{2/3})$.

Γ_b is a property of the instance, not a rate. It is zero when local solutions already spend the global budget, and it is linear in T only if the late split η stays a constant. The first half looks the same on both paths. Later the real arm is worth η more, or η less. A planner that knows the path saves the whole budget for the late rounds, or spends it all early. A learner has to choose how much to spend before the split is visible. The two regrets add to at least ηbT. Telling the learner V_1 does not help: both paths have the same V_1. The hard budget is what makes the choice irreversible. If $b = 1$, the same late jump gives $\Gamma_b = 0$, and there is nothing to save. The upper bound therefore keeps Γ_b as its own term. Folding it into the tracking term would claim a rate these two worlds forbid, because there the tracking term is the smaller one. Folding it into W_1 would charge a cost on instances where every round's best decision is already best for the whole budget, and there is nothing to save.

Proposition 4 (Γ_b versus W_1)**.** *There exist environments with $V_1 = O(1)$ and $V_2 = 0$ such that $W_1 = \Theta(T)$ and $\Gamma_b = 0$. On the environments of Theorem 3, $\Gamma_b = bW_1$.*

So W_1 is not the gap. A free arm whose mean jumps once makes W_1 linear, but every local LP is already globally optimal. On the spend-or-save pair the bridge is tight up to the factor b. Liu et al. (2022) already show that some $W_1 + \bar{q}W_2$ term is necessary when the variation is given. That term can overstate the gap; Γ_b is the part that any learner must keep. The proofs are in Appendix F.

Corollary 5 (Replay thinning)**.** *For any fixed $\gamma \in (0, 1]$, replacing each non-mandatory schedule probability by $p^{(\gamma)}_{n,k}$ preserves pathwise feasibility and the same horizon and variation exponents, with at most a γ^{-1} factor in worst-case detection times.*

Thinning by γ only multiplies detection times by γ^{-1}. Exact m, d, b, γ factors are in Appendix E.

6 Experiments

We use $T = 1024$, $b = 0.35$, and three instances. The stationary environment has three real arms and one null arm, with Bernoulli rewards $(0.75, 0.72, 0.50, 0)$ and consumptions $(0.30, 0.24, 0.32, 0)$. The hidden-arm instance is deterministic with two real arms: rewards are $(0.6, 0, 0)$ before $T/4$ and $(0.6, 1, 0)$ afterward, and consumption is identically zero. After a long instance drops arm 2, only a short replay can see that its reward became one. The complementary instance is also deterministic. Arm 0 has reward 0.75 and consumption 0.5 throughout. Arm 1 changes from reward 0 and consumption 1 to reward 1 and consumption 0 at $T/3$. A stale reward memory or a stale consumption memory can each stay wrong after that jump.

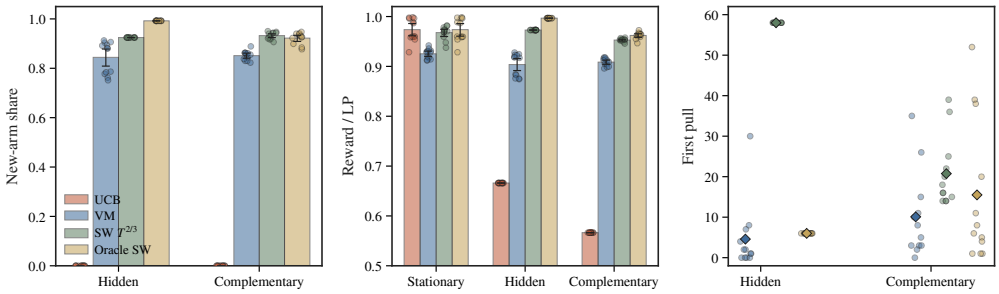


Figure 1: Comparison at $T = 1024$, $b = 0.35$. UCB is stationary UCB-BwK; VM is Vector-MASTER with $\gamma = 1/2$. **(a)** Post-change share of the new arm. **(b)** Pseudo-reward over the global fluid LP. **(c)** Rounds from the change to the first pull of that arm. Points are 12 paired seeds; whiskers in (a)–(b) are bootstrap 95% intervals.

We compare calibrated Vector-MASTER with $\gamma = 1/2$ to a stationary UCB-BwK learner. Both methods are prior-free, so this comparison is fair. We also plot two sliding-window references, and that comparison is not fair. The first reference uses the fixed window $\lceil T^{2/3} \rceil$. The second is an oracle that selects one common reward/consumption window from $\{16, 32, \ldots, T\}$ using the environment identity and four independent tuning seeds; its selected window is then frozen for evaluation. Held-out tuning selects oracle windows 1024, 256, and 64 on the three instances; the oblivious window is 102. So the window references are given a window length, and the oracle is also given the instance type.

The Vector-MASTER thresholds were frozen on diagnostic seeds 0–49 before these evaluation seeds, as recorded in Appendix G. Evaluation uses 12 new paired seeds per environment and method, for 144 episodes. Every method receives the same potential outcomes and the same one-unit reserve. We report pseudo-reward divided by the global fluid-LP upper bound and the post-change share of the newly preferred arm.

On the hidden-arm instance, Vector-MASTER puts 0.844 of post-change rounds on the improved arm, and stationary UCB puts zero. The reward ratio is 0.904, which is 0.238 above stationary UCB. On the complementary instance the share is 0.851 versus zero, and the reward ratio is 0.342 above stationary UCB. The stationary reward ratio is 0.925.

On both change instances, Vector-MASTER first pulls the new arm on every seed. The hidden-arm median delay is 1.5 rounds, and one seed takes 30. The complementary median is 6.5, and the largest delay is 35. Stationary UCB never pulls it. All 144 evaluation runs stay inside the budget. On the complementary instance stationary UCB uses 0.997 of the budget on the old arm; after the jump, Vector-MASTER uses 0.730. Stationary UCB never returns because its UCB fell below the

other arm before the change. Vector-MASTER returns because a short replay still initializes that arm. The tests rarely restart the long instance: T1 fires on two of twelve hidden seeds and three of twelve complementary seeds, and T2R and T2C do not fire. So the new-arm share in Figure 1 is mostly the replay, not a reset of the long instance.

The stationary reward 0.925 versus 0.974 is the cost of that replay when nothing changes. That gap is why we next vary γ. The two window series sit higher because those methods are given a length.

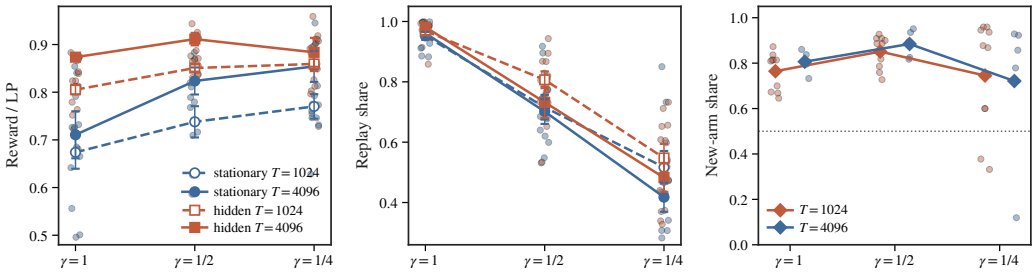


Figure 2: Cost of the random replay schedule. **(a)–(b)** Seeds and paired means, one-standard-error whiskers. **(c)** Terminal share of the improved arm. The dotted line is the recovery threshold $1/2$.

Figure 2 varies only the replay probability. At $T = 4096$, changing γ from 1 to $1/4$ reduces stationary replay ownership from 0.961 to 0.418 and raises normalized reward from 0.711 to 0.854. Hidden-arm reward is not monotone: 0.874 at $\gamma = 1$, 0.911 at $\gamma = 1/2$, and 0.884 at $\gamma = 1/4$. Full replay finds the arm but spends most rounds on short instances. A run with no random replay keeps a stationary reward ratio of 0.920 and an adoption of 0.0002: the long instance never samples the neglected arm again.

At $\gamma = 1/4$ the short replay still hits the improved arm near the change, but the long instance does not keep it on $2/10$ seeds at $T = 1024$ and on $1/5$ seeds at $T = 4096$. The two instances in Figure 2 are one-jump probes and do not sweep Δ. Consumption is zero on the hidden-arm screen, so $\Gamma_b = 0$. Periodic probes in Appendix G get stationary reward 0.914 and adoption 0.996 with replay share 0.225, but that schedule is not the one in the theorem. In Figure 1 the short instances own 0.728 of the hidden-arm rounds and 0.782 of the stationary rounds. A later jump at $0.73T$, still at $T = 4096$ and $\gamma = 1/2$, misses on one of five seeds and has mean share 0.718. The main plots use $\gamma = 1/2$ because recovery is then five of five at $T = 4096$.

The plots separate the two costs in the bound. Short runs are what find the new arm, and they spend a visible share of the rounds. Making them rarer returns those rounds when the means do not move, and can miss the arm when they do. On the hidden-arm screen the cost is zero, so $\Gamma_b = 0$: that screen is only about tracking the change, not about saving budget for later. The complementary screen is the other failure. A stale reward or a stale cost can each keep the old arm after the jump, and only a short run sees the arm the long run has dropped. The same split is what the bound keeps: $\widetilde{O}\big(\sqrt{mT} + m^{1/3}(V_1 + dV_2)^{1/3}T^{2/3}\big) + \Gamma_b$, with the second term present when saving budget for a later round can still raise the total reward.

7 Conclusions

We gave a prior-free method for non-stationary BwK with one hard budget. The local term matches a sliding window given the amount of change. The extra gap Γ_b is against a planner that can move budget across time, and it is necessary: any policy already pays it on one of two paths that share the first half, even if it is told the amount of change. The window term on those paths is only $T^{2/3}$. W_1 can be linear while $\Gamma_b = 0$. On the two change instances the method moves to the new arm. The theorem uses $\gamma = 1$; the plots use $\gamma = 1/2$ because a replay spends the same S. The same question extends to contextual arrivals and to concave rewards under one hard budget. The target remains a tracking term that does not need V, plus only the gap forced by a budget that cannot be moved.

REFERENCES

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 12–38. PMLR, 2017. URL https://proceedings.mlr.press/v65/agarwal17b.html.

Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pp. 989–1006. ACM, 2014. doi: 10.1145/2600057.2602844. URL https://doi.org/10.1145/2600057.2602844.

Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 4–18. PMLR, 2016. URL https://proceedings.mlr.press/v49/agrawal16.html.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002a. doi: 10.1023/A:1013689704352.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b. doi: 10.1137/S0097539701398375.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 138–158. PMLR, 2019. URL https://proceedings.mlr.press/v99/auer19a.html.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013. doi: 10.1109/FOCS.2013.30. URL https://doi.org/10.1109/FOCS.2013.30.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM*, 65(3):1–55, 2018. doi: 10.1145/3164539.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Beyond primal-dual methods in bandits with stochastic and adversarial constraints. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8541–8568, 2024a. doi: 10.52202/079017-0273. URL https://openreview.net/forum?id=iJgwd5mWYg.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Bandits with replenishable knapsacks: The best of both worlds. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=yBIJRIYTqa.

Martino Bernasconi, Matteo Castiglioni, and Andrea Celli. No-regret is not enough! bandits with general constraints through adaptive regret minimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3877–3898. PMLR, 2025. URL https://proceedings.mlr.press/v267/bernasconi25a.html.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, volume 27, pp. 199–207, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/91ba7292e5388b90b58d0b839a7f19ec-Abstract.html.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: The best of both worlds. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2767–2783. PMLR, 2022. URL https://proceedings.mlr.press/v162/castiglioni22a.html.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 696–726. PMLR, 2019. URL https://proceedings.mlr.press/v99/chen19b.html.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1079–1087. PMLR, 2019. URL https://proceedings.mlr.press/v89/cheung19b.html.

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. A technical note on non-stationary parametric bandits: Existing mistakes and preliminary solutions. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 619–626. PMLR, 2021. URL https://proceedings.mlr.press/v132/faury21a.html.

Giannis Fikioris and Éva Tardos. Approximately stationary bandits with knapsacks. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3758–3782. PMLR, 2023. URL https://proceedings.mlr.press/v195/fikioris23a.html.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory*, pp. 174–188, 2011. doi: 10.1007/978-3-642-24412-4_16. URL https://doi.org/10.1007/978-3-642-24412-4_16.

Argyrios Gerogiannis, Yu-Han Huang, and Venugopal Veeravalli. Is prior-free black-box non-stationary reinforcement learning feasible? In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2692–2700. PMLR, 2025. URL https://proceedings.mlr.press/v258/gerogiannis25a.html.

Hengquan Guo and Xin Liu. On stochastic contextual bandits with knapsacks in small budget regime. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FCMpUOZkxi.

Qinyang He and Yonatan Mintz. Non-stationary bandits with habituation and recovery dynamics and knapsack constraints. In *2024 IEEE 63rd Conference on Decision and Control*, pp. 1167–1172. IEEE, 2024. doi: 10.1109/CDC56724.2024.10885889. URL https://doi.org/10.1109/CDC56724.2024.10885889.

Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science*, pp. 202–219, 2019. doi: 10.1109/FOCS.2019.00022. URL https://doi.org/10.1109/FOCS.2019.00022.

Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2286–2305. PMLR, 2020. URL https://proceedings.mlr.press/v125/kesselheim20a.html.

Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6483–6492. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21s.html.

Shang Liu, Jiashuo Jiang, and Xiaocheng Li. Non-stationary bandits with knapsacks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16522–16532, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/69469da823348084ca8933368ecbf676-Abstract.html.

Lixing Lyu and Wang Chi Cheung. Bandits with knapsacks: Advice on time-varying demands. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23212–23238. PMLR, 2023. URL https://proceedings.mlr.press/v202/lyu23a.html.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvári. Model selection in contextual stochastic bandit problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10328–10337, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/751d51528afe5e6f7fe95dece4ed32ba-Abstract.html.

Anshuka Rangi, Massimo Franceschetti, and Long Tran-Thanh. Unifying the stochastic and the adversarial bandits with knapsack. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3311–3317, 2019. doi: 10.24963/ijcai.2019/459. URL https://doi.org/10.24963/ijcai.2019/459.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12040–12049, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/263fc48aae39f219b4c71d9d4bb4aed2-Abstract.html.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Bandits with knapsacks beyond the worst case. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23191–23204, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354. PMLR, 2021. URL https://proceedings.mlr.press/v134/wei21b.html.

Xilin Zhang and Wang Chi Cheung. Piecewise-stationary bandits with knapsacks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56001–56040, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/65ae674df2fb642518ae8d2b5435e1b8-Abstract.html.

A CERTIFICATE STATEMENTS

For a calendar interval $I = [u, v]$, write

$$V_1(I) = \sum_{t=u}^{v-1} \|\mu_{t+1} - \mu_t\|_\infty,$$
$$V_{2,j}(I) = \sum_{t=u}^{v-1} \|C_{t+1,j} - C_{t,j}\|_\infty,$$
$$\widetilde{V}_2(I) = \sum_{t=u}^{v-1} \max_j \|C_{t+1,j} - C_{t,j}\|_\infty, \qquad \Delta_I = V_1(I) + \widetilde{V}_2(I)/b'.$$

An active prefix with physical times $s_1 < \cdots < s_n$ is assigned the calendar span $I = [s_1, s_n]$; the same convention defines Δ_P for a calendar interval P.

The owner map uses the planned dyadic order of each candidate. If two candidates have the same planned length, they do not overlap. If a final block is cut short and its base has the same interval as a shorter cell, the planned order still picks a unique owner. If a last tie remains, we use a fixed identifier. A candidate completes only if the epoch lasts through e_a and the outcome at e_a is seen.

Lemma 6 (Paused BwK interface)**.** *On the joint event constructed in Appendix B, simultaneously for every scheduled instance and active prefix, predictable confidence widths obey*

$$\sum_{q \le n} \sum_i x_{q,i} \beta_{q,i} = \widetilde{O}(\sqrt{mn} + m). \tag{7}$$

The base instance also satisfies

$$\widetilde{g}_n \ge \min_{q \le n} \phi_{s_q} - O(\Delta_I), \tag{8}$$
$$\sum_{q \le n} (\widetilde{g}_q - R_q) \le \widetilde{O}(\sqrt{mn} + m) + O(n\Delta_I), \tag{9}$$
$$\sum_{q \le n} \frac{Z_{q,j} - \widetilde{h}_{q,j}}{b'} \le \widetilde{O}((\sqrt{mn} + m)/b') + O(nV_{2,j}(I)/b'). \tag{10}$$

Equivalently, after increasing the universal coefficients in C, there are fixed universal constants $c_B, c_O > 0$ such that, for $q \in \{R, C_1, \ldots, C_d\}$,

$$\sum_{r \le n} D_r^q \le C(n) + c_B n \Delta_I, \qquad \widetilde{g}_n \ge \min_{r \le n} \phi_{s_r} - c_O \Delta_I. \tag{10a}$$

For any calendar interval P contained in one live block, fixing an optimal dual q_u at one reference time $u \in P$ and using the actual owner on every round gives

$$\sum_{t \in P} (R_t - \phi_t) \le \sum_j b' q_{u,j} \left[\sum_{t \in P} \frac{Z_{t,j} - \widetilde{h}_{t,j}}{b'} \right] + c_F |P| \Delta_P + c_M C(|P|). \tag{11}$$

Here $c_F, c_M > 0$ are fixed universal constants. Thus every constant used by the replay proof is named before the multi-scale reduction.

Lemma 7 (Stopping conversion)**.** *If resource $j^\star$ crosses the reserve at time τ, then*

$$T - \tau < \sum_{t \le \tau} \frac{Z_{t,j^\star} - \widetilde{h}_{t,j^\star}}{b'}. \tag{12}$$

The following sections establish these statements uniformly over the adaptively paused instances and then aggregate them.

B UNIFORM EVENTS FOR PREDICTABLY PAUSED INSTANCES

Lemma 8 (Uniform optional-skipping event)**.** *With probability at least $1-\delta$, the concentration and schedule-count inequalities used below hold simultaneously for every scheduled instance, active prefix, arm, resource, tested interval, fixed-dual martingale, and stopped-trial sequence.*

Proof. Let $\widehat{n} = \lfloor \log_2 T \rfloor + 1$ and

$$\ell = \log \frac{24mdT^4\widehat{n}^2}{\delta}, \qquad \Lambda_T = \log \frac{12T^3\widehat{n}^2}{\delta}.$$

We use four reveal stages. First, $\mathcal{H}_{t^-}$ contains all feedback through $t-1$, the current epoch and instance states, and coins of candidates starting strictly before t. Attach an independent uniform variable U_a to each candidate a starting at t, and reveal these variables to obtain $\mathcal{G}_{t^-}$. The owner o_t and mixture x_t are $\mathcal{G}_{t^-}$-measurable. Next reveal an independent action coin and draw $i_t \sim x_t$, obtaining $\mathcal{F}_{t^a}$. Finally draw the round-t potential-outcome table, conditionally independent of $\mathcal{F}_{t^a}$ with the means in Section 3, and reveal its selected entry to obtain $\mathcal{F}_t$.

Although the implementation samples schedule variables at block start, deferred reveal is distributionally identical: a candidate's coin has no effect before its nominal start and all schedule coins are mutually independent and independent of action and outcome randomness. In particular, candidate eligibility and block liveness at its start are $\mathcal{H}_{t^-}$-measurable, while $\mathbf{1}\{U_a \le p_{n,k}^{(\gamma)}\}$ is conditionally Bernoulli with parameter $p_{n,k}^{(\gamma)}$ given $\mathcal{H}_{t^-}$.

Let a scheduled instance be active at physical times $s_1 < \cdots < s_n$. The preceding filtration makes each active indicator and action distribution predictable before the corresponding outcome table. For arm i, optional skipping gives, uniformly over active sample count N,

$$\left| \sum_{q \le n} \mathbf{1}\{i_q = i\}(R_q - \mu_{s_q,i}) \right| \le O(\sqrt{N\ell} + \ell), \tag{13}$$

and the same event holds for each resource. There are at most $T(\lfloor \log_2 T \rfloor + 1)$ nominal candidates and at most T^2 calendar intervals. Union bounds over candidates, active counts, arms, resources and fixed-dual interval martingales are covered by ℓ. The independent schedule-count and stopped-trial bounds are covered by Λ_T. Safety stopping only selects prefixes already present in these uniform events. Calendar gaps contribute path variation over an instance's physical-time span, not fictitious samples.

For predictable widths $\alpha_{q,i} = \min\{1, c\sqrt{\ell/(N_{q,i}+1)}\}$, the action-sampling martingale and the elliptical count sum give

$$\sum_{q \le n} \sum_{i} x_{q,i}\alpha_{q,i} \le O(\sqrt{mn\ell} + m\ell + \sqrt{n\ell}). \tag{14}$$

For completeness, retaining constants in Hoeffding–Azuma, using $\beta_{q,i} = \min\{1, \sqrt{2\ell/(N_{q,i}+1)}\}$, and applying $\sum_{r=1}^{N} r^{-1/2} \le 2\sqrt{N}$ bounds the sum of confidence widths, action-sampling noise, outcome noise, and one initialization pass by

$$32(1 + b'^{-1}) \min\{\sqrt{mn\ell} + m\ell,\ n\}. \tag{14a}$$

The same bound covers one reward coordinate, one scaled resource coordinate, and the fixed-dual martingale because $b'\|q_u\|_1 \le 1$. This is the numerical origin of $C_1 = C_2 = C_3 = 64$ in (4b), with a factor two reserved for the clipped linear branch. Forced initialization contributes $O(m)$ reward discrepancy and $O(m/b')$ scaled resource discrepancy. □

C BASE CERTIFICATES

Proof of Lemma 6. For an active prefix I, let $D_I^r = V_1(I)$ and $D_I^c = \widetilde{V}_2(I)$. Equation (7) follows from (14), absorbing the logarithmic factors and explicit initialization. On the joint event,

$$U_t \ge \mu_t - D_I^r\mathbf{1}, \qquad L_t \le C_t + D_I^c\mathbf{1}\mathbf{1}^\top.$$

Mixing a true optimizer with the null arm by weight $b'/(b' + D_I^c)$ proves (8). For (9), decompose

$$\widetilde{g}_t - R_t = (U_t - \mu_t)^\top x_t + (\mu_t^\top x_t - R_t)$$

and apply (14) and a reward martingale. Equation (10) follows from

$$Z_{t,j} - \widetilde{h}_{t,j} = (Z_{t,j} - C_{t,j}^\top x_t) + (C_{t,j} - L_{t,j})^\top x_t.$$

The first terms in these two decompositions are martingale sums; (14) bounds the confidence-width terms, and changing a historical mean to the current mean costs at most $n\Delta_I$. Since $C(n)$ dominates all stationary terms in (9)–(10), taking c_B to be the largest coefficient of these variation remainders and c_O to be the coefficient in (8) proves (10a).

LP target variation. For any s, t,

$$|\phi_t(b') - \phi_s(b')| \le \|\mu_t - \mu_s\|_\infty + b'^{-1} \max_j \|C_{t,j} - C_{s,j}\|_\infty. \tag{15}$$

The same null-arm mixing proves this statement in both directions.

Fixed-dual domination. At a reference time u, weak duality gives

$$\mu_u^\top x \le \phi_u(b') + q_u^\top (C_u x - b'\mathbf{1}), \qquad b'\|q_u\|_1 \le 1.$$

For the actual mixture x_t, add and subtract the time-u quantities. Equation (15) and the definition of Δ_P give

$$\mu_t^\top x_t - \phi_t \le q_u^\top (C_t x_t - b'\mathbf{1}) + c_F \Delta_P$$

for a universal c_F. Since $q_u^\top (\widetilde{h}_t - b'\mathbf{1}) \le 0$,

$$q_u^\top (C_t x_t - b'\mathbf{1}) \le q_u^\top (Z_t - \widetilde{h}_t) + q_u^\top (C_t x_t - Z_t).$$

Adding $R_t - \mu_t^\top x_t$ to the weak-duality inequality therefore gives

$$R_t - \phi_t \le q_u^\top (Z_t - \widetilde{h}_t) + M_t + c_F \Delta_P,$$

where

$$M_t = (R_t - \mu_t^\top x_t) + q_u^\top (C_t x_t - Z_t)$$

is a martingale difference. Summing over P, writing $q_u^\top (Z_t - \widetilde{h}_t) = \sum_j b' q_{u,j} D_{t,j}^C$, and using the minimum of its Hoeffding bound and deterministic range bound gives

$$\sum_{t \in P} M_t \le (1 + b'^{-1}) \min\{\sqrt{|P|\ell}, |P|\} \le c_M C(|P|).$$

This proves (11). The same fixed q_u multiplies the whole signed sum; replacing it by time-varying q_t would destroy the martingale. □

D VECTOR MULTI-SCALE REDUCTION

The simultaneous test threshold used in Section 4 is

$$\widehat{\rho}(s) := 6(\lfloor \log_2 T \rfloor + 1) \log \frac{12T^3(\lfloor \log_2 T \rfloor + 1)^2}{\delta}\, \rho(s), \tag{4a}$$

Take $C_1 = C_2 = C_3 = 64$ in Appendix E. With

$$c_\star = c_B + c_F + c_O + 1,$$

one sufficient symbolic choice of test constants is

$$K_R = K_C = 4, \qquad K_1 = 3 + c_M/6, \qquad K_\tau = K_1 + 2 + (c_B + c_O + 1)/6, \tag{4b}$$

and

$$\kappa = \min\{1,\ 3/c_B,\ 3/c_\star\}. \tag{4c}$$

These are fixed universal constants because c_B, c_O, c_F, c_M are fixed in (10a)–(11). The reference implementation exposes concrete finite-horizon constants for reproducibility; the proof certifies the inequalities (4b)–(4c), not an unproved numerical substitution for the named constants.

Define

$$D_t^R = \widetilde{g}_t - R_t, \qquad D_{t,j}^C = (Z_{t,j} - \widetilde{h}_{t,j})/b'.$$

Definition 9 (Admissible calendar interval)**.** Within a live block, P is *admissible* if it is either (i) a prefix beginning at the block start, or (ii) the nominal support J_a of a completed candidate. This uses only the block-anchored laminar dyadic schedule; shifted periodic probes are not part of the theorem algorithm.

Lemma 10 (Laminar ownership and schedule packing)**.** *Let P be admissible and $A_a(P) = \{t \in P : o_t = a\}$. The nonempty sets $A_a(P)$ partition P, and each is the first $k_a = |A_a(P)|$ active rounds of a fresh instance a. On the schedule-count event, simultaneously for every admissible P of length s,*

$$\sum_{a:A_a(P)\neq\emptyset} C(k_a) \le 6(\lfloor \log_2 T \rfloor + 1)C(s)\Lambda_T. \tag{16a}$$

Proof. For a block prefix, every owner starts inside the block and all its active rounds so far lie in P. For a completed support $P = J_a$, an owner is either a or a lower-order dyadic cell whose nominal start lies in J_a: higher-order instances cannot own while a covers the round, distinct equal-order cells are disjoint, and lower-order cells are contained in J_a. The containment may be equality only when a is a truncated mandatory base. Thus no owner has an earlier active round outside P. All candidates may be allocated at block start; "fresh" means that their statistics contain no observations before their first owned round.

Fix an order-n block. For optional order $r < n$, let $\mathcal{S}_r(P)$ be the scheduled cells that can own in P. Block-anchored order-r cells are disjoint, and the simultaneous Bernstein event gives

$$|\mathcal{S}_r(P)| \le 2\left(p_{n,r}^{(\gamma)}\frac{s}{2^r} + 2\right)\Lambda_T.$$

When $P = J_a$, conditioning on a being scheduled contributes one deterministic cell, absorbed by the additive 2; all descendant coins remain independent. The event is union-bounded over block starts, orders, block prefixes, and nominal candidate supports. If $2^r \le s$, then $k_a \le 2^r$, C is nondecreasing, and

$$p_{n,r}^{(\gamma)}\frac{s}{2^r}C(2^r) = \gamma\rho(2^n)s \le C(s).$$

Hence $\sum_{a\in\mathcal{S}_r(P)} C(k_a) \le 6C(s)\Lambda_T$. If $2^r > s$, at most one order-r cell can own in P, and its contribution is at most $C(s)$. The mandatory base has the same one-owner bound. Summing over orders proves (16a). □

Lemma 11 (Aggregate multi-scale certificate)**.** *For every admissible P of length s satisfying $\Delta_P \le \rho(s)$, simultaneously for $q \in \{R, C_1, \ldots, C_d\}$,*

$$\sum_{t\in P} D_t^q \le 6(\lfloor \log_2 T \rfloor + 1)C(s)\Lambda_T + c_B s\Delta_P. \tag{16}$$

Moreover, on every $t \in P$,

$$\widetilde{g}_t \ge \min_{u\in P}\phi_u - c_O\Delta_P. \tag{16b}$$

Proof. By Lemma 10, $A_a(P)$ is an active prefix whose physical-time span lies in P. Since $k_a \le s$ and ρ is nonincreasing,

$$\Delta_{A_a(P)} \le \Delta_P \le \rho(s) \le \rho(k_a).$$

The paused-instance certificate therefore gives

$$\sum_{t\in A_a(P)} D_t^q \le C(k_a) + c_B k_a\Delta_P.$$

Sum over owners, use $\sum_a k_a = s$, and apply (16a). Applying the active-prefix optimism bound to the owner at t, whose entire active span is contained in P, proves (16b). No owner-only sum is divided by the nominal length s; aggregation first reconstructs the full calendar interval. □

Lemma 12 (No false restart and restart charge)**.** *For a sufficiently small universal $\kappa > 0$, an epoch prefix I with $\Delta_I \le \kappa\rho(|I|)$ triggers none of T1, T2R, or T2C. Consequently every non-final epoch satisfies (16e).*

Proof. Suppose an epoch prefix I satisfies $\Delta_I \le \kappa\rho(|I|)$. For any tested block prefix Q of length $s \le |I|$, Q is admissible and (16) implies

$$\frac{1}{s}\sum_{t\in Q} D_t^q \le \widehat{\rho}(s) + c_B\Delta_I. \tag{16c}$$

Since $s \le |I|$, monotonicity of ρ, (4a), and $\widehat{n}\Lambda_T \ge 1$ give

$$\Delta_I \le \kappa\rho(|I|) \le \kappa\rho(s) \le \frac{\kappa}{6}\widehat{\rho}(s).$$

Therefore the right-hand side of (16c) is at most $(1 + c_B\kappa/6)\widehat{\rho}(s) \le \frac{3}{2}\widehat{\rho}(s) < K_R\widehat{\rho}(s)$, and the same calculation with $q = C_j$ is below $K_C\widehat{\rho}(s)$. Thus neither T2 test fires.

Now let a be a completed candidate, $P = J_a$, $h = |P|$, and let Q be its block prefix through e_a. Both P and Q are admissible, with variation at most Δ_I. In (11), $b'\|q_u\|_1 \le 1$; applying (16) to every resource coordinate and then taking the $b'q_u$-weighted sum gives

$$\frac{1}{h}\sum_{t\in P}(R_t - \phi_t) \le \widehat{\rho}(h) + (c_B + c_F)\Delta_I + c_M\rho(h).$$

For each $s \in Q$, (16b) gives $\widetilde{g}_s \ge \min_{u\in Q}\phi_u - c_O\Delta_I$; taking the minimum over s yields $\mathcal{U}_{e_a} \ge \min_{u\in Q}\phi_u - c_O\Delta_I$. Target sensitivity (15), and $P \subseteq Q$, then imply

$$\frac{1}{h}\sum_{t\in P}\phi_t \le \min_{u\in Q}\phi_u + \Delta_I \le \mathcal{U}_{e_a} + (c_O + 1)\Delta_I.$$

Adding the last two displays and using $\rho(h) \le \widehat{\rho}(h)/6$ gives

$$\frac{1}{h}\sum_{t\in J_a} R_t \le \mathcal{U}_{e_a} + \left(1 + \frac{c_M}{6}\right)\widehat{\rho}(h) + c_\star\Delta_I. \tag{16d}$$

Finally $h \le |I|$ implies $\Delta_I \le \kappa\widehat{\rho}(h)/6$. By (4c), $c_\star\kappa/6 \le 1/2$, so the right-hand side of (16d) is strictly below $\mathcal{U}_{e_a} + K_1\widehat{\rho}(h)$ for the choice in (4b). Thus T1 does not fire.

Therefore every non-final epoch E obeys

$$\Delta_E > \kappa\rho(|E|). \tag{16e}$$

□

Lemma 13 (Adaptive replay optimism)**.** *For every realized order-n block, the learner-independent nominal partition and the deferred-reveal schedule satisfy (17) simultaneously over all adaptive block endpoints.*

Proof. Let t_n be the predictable start of an order-n nominal block $\bar{J} = [t_n, t_n + 2^n - 1] \cap [1, T]$. Before revealing any schedule coin, greedily partition $\bar{J}$ into maximal consecutive intervals $\mathcal{I}_1, \dots, \mathcal{I}_{\bar{L}}$ satisfying $\Delta_{\mathcal{I}_\ell} \le \rho(|\mathcal{I}_\ell|)$. This partition depends only on the fixed environment path. Write $e_\ell = \max\mathcal{I}_\ell$.

Let E_n be the earliest of the nominal block end, a restart-test time, a reserve-crossing time, and T. Intersect the nominal partition with the realized block $J = [t_n, E_n]$, discard empty intersections, and relabel them $\mathcal{I}'_1, \dots, \mathcal{I}'_L$, where $e'_\ell = \max\mathcal{I}'_\ell = \min\{e_\ell, E_n\}$. Truncation preserves

$$\Delta_{\mathcal{I}'_\ell} \le \Delta_{\mathcal{I}_\ell} \le \rho(|\mathcal{I}_\ell|) \le \rho(|\mathcal{I}'_\ell|),$$

where the last inequality uses that ρ is nonincreasing.

For $k \in \{0, \dots, n\}$, abbreviate

$$\alpha_k = \rho(2^k), \qquad \widehat{\alpha}_k = \widehat{\rho}(2^k).$$

Define $\tau_\ell(k)$ as the first $t \in \mathcal{I}'_\ell$ such that $\phi_t - \widetilde{g}_t > K_\tau\widehat{\alpha}_k$, and set $\tau_\ell(k) = \infty$ if no such time exists. Let

$$\xi_\ell(k) = \begin{cases} e'_\ell - \tau_\ell(k) + 1, & \tau_\ell(k) < \infty, \\ 0, & \tau_\ell(k) = \infty. \end{cases}$$

Because $\widehat{\alpha}_k \le 2\widehat{\alpha}_{k+1}$ (check the $s^{-1/2}$, s^{-1}, and clipped branches separately), grouping each positive gap $\phi_t - \widetilde{g}_t$ by the largest crossed threshold gives

$$\sum_{t\in\mathcal{I}'_\ell}(\phi_t - \widetilde{g}_t) \le K_\tau \left(|\mathcal{I}'_\ell|\widehat{\alpha}_n + 2\sum_{k=0}^{n} \widehat{\alpha}_k \xi_\ell(k) \right). \tag{17i}$$

This is the discrete layer-cake identity: the baseline pays $\widehat{\alpha}_n$ for every point, and a point crossing level k is counted in the suffix beginning at the first crossing of that level.

Split each suffix as

$$\xi_\ell(k) = \min\{\xi_\ell(k), 4\cdot 2^k\} + [\xi_\ell(k) - 4\cdot 2^k]_+.$$

For the short part, put $r = |\mathcal{I}'_\ell|$. If $r \le h = 2^k$, then $\widehat{\rho}(h) \le \widehat{\rho}(r)$. If $r > h$, monotonicity of $\widehat{C}(s) = s\widehat{\rho}(s)$ gives $\widehat{C}(h) \le \widehat{C}(r)$, while $\min\{r, 4h\}/h \le 4$. These two cases give

$$\sum_{\ell} \widehat{\alpha}_k \min\{\xi_\ell(k), 4\cdot 2^k\} \le 4\sum_{\ell} |\mathcal{I}'_\ell|\widehat{\rho}(|\mathcal{I}'_\ell|). \tag{17j}$$

The order-n suffix is always short because $|\mathcal{I}'_\ell| \le 2^n$. It remains to control long suffixes for $k < n$.

Fix such a k and put $h = 2^k$. On the deterministic nominal piece $\mathcal{I}_\ell$, define the online crossing time

$$\sigma_\ell(k) = \inf\{s \in \mathcal{I}_\ell : \text{the block is alive through } s \text{ and } \phi_s - \widetilde{g}_s > K_\tau \widehat{\alpha}_k\}.$$

Whenever the block is alive through $t-1$, $\{\tau_\ell(k) \le t-1\} = \{\sigma_\ell(k) \le t-1\}$. An aligned start t is eligible when

$$\sigma_\ell(k) \le t-1, \qquad t \le e_\ell - 2h, \qquad (t - t_n) \bmod h = 0,$$

and the block is alive through $t-1$. Every item in this definition is $\mathcal{H}_{t^-}$-measurable: it uses only gaps through $t-1$, the deterministic endpoint e_ℓ, and past survival. It does not use the future random endpoint E_n.

Let $B_{t,k}$ be the order-k schedule coin at an aligned start. By the deferred-reveal construction,

$$\Pr(B_{t,k} = 1 \mid \mathcal{H}_{t^-}) = p^{(\gamma)}_{n,k} = \gamma\frac{\alpha_n}{\alpha_k}.$$

Let Q_k count eligible failures before the first eligible success. Order all aligned starts deterministically. Conditional on the event that the r-th start is the next eligible trial and all previous eligible trials failed, its coin is still unrevealed and fails with probability $1 - p^{(\gamma)}_{n,k}$. Iterating this conditional identity q times gives

$$\Pr(Q_k \ge q) \le (1 - p^{(\gamma)}_{n,k})^q.$$

If $p^{(\gamma)}_{n,k} = 1$, then $Q_k = 0$. Otherwise, taking

$$q_k = \left\lceil \frac{\log(T^3(\log_2 T + 1)/\delta)}{-\log(1 - p^{(\gamma)}_{n,k})} \right\rceil \le 1 + \frac{\alpha_k}{\gamma\alpha_n}\log\frac{T^3(\log_2 T + 1)}{\delta}$$

and union-bounding over block–scale pairs yields, simultaneously,

$$Q_k \le 1 + \frac{\rho(2^k)}{\rho(2^n)}\frac{1}{\gamma}\log\frac{T^3(\log_2 T + 1)}{\delta}. \tag{17a}$$

For a realized piece define

$$A_\ell(k) = \#\{t : \tau_\ell(k) + 1 \le t \le e'_\ell - 2h,\ (t - t_n) \bmod h = 0\}.$$

Starting from the suffix of $\xi_\ell(k)$ points, exclude its crossing point, at most $h-1$ points before the first aligned start, and the final $2h$ points needed for completion. Fewer than $4h$ points are discarded, and the remaining aligned starts are spaced by h. Therefore

$$A_\ell(k) \ge \frac{[\xi_\ell(k) - 4h]_+}{h}. \tag{17b}$$

Every counted start is eligible. If one of its coins succeeds, its replay $a = [t, t+h-1]$ lies in $\mathcal{I}'_\ell$ and

$$e_a = t + h - 1 \le e'_\ell - h - 1 \le E_n - h - 1 < E_n. \tag{17k}$$

Thus a completes before the realized block endpoint. Since $\Delta_{J_a} \le \Delta_{\mathcal{I}_\ell} \le \rho(|\mathcal{I}_\ell|) \le \rho(h)$, applying (16) and (16b) to J_a gives

$$\begin{aligned}\frac{1}{h}\sum_{s\in J_a} R_s &= \frac{1}{h}\sum_{s\in J_a}\widetilde{g}_s - \frac{1}{h}\sum_{s\in J_a} D_s^R \\ &\ge \min_{s\in J_a}\phi_s - \widehat{\rho}(h) - (c_B + c_O)\Delta_{\mathcal{I}_\ell} \\ &\ge \phi_{\tau_\ell(k)} - \widehat{\rho}(h) - (c_B + c_O + 1)\Delta_{\mathcal{I}_\ell} \\ &> \widetilde{g}_{\tau_\ell(k)} + \left[K_\tau - 1 - \frac{c_B + c_O + 1}{6}\right]\widehat{\rho}(h) \\ &\ge \mathcal{U}_{e_a} + K_1\widehat{\rho}(h).\end{aligned} \tag{17l}$$

The second inequality uses target sensitivity within $\mathcal{I}_\ell$. The third uses the crossing definition and $\Delta_{\mathcal{I}_\ell} \le \rho(h) \le \widehat{\rho}(h)/6$. The final inequality uses $\mathcal{U}_{e_a} \le \widetilde{g}_{\tau_\ell(k)}$ and (4b). Hence T1 would fire at $e_a < E_n$, contradicting the definition of E_n. No counted coin succeeds, so $\sum_\ell A_\ell(k) \le Q_k$: for every full nominal piece, eligibility after its crossing is exactly counted by $A_\ell(k)$; only the final truncated piece can have uncounted eligible starts, and those satisfy $t > E_n - 2h$, whereas every counted start satisfies $t \le E_n - 2h$. Thus every counted failure precedes the first possible uncounted eligible success and is included in the stopped prefix defining Q_k.

Multiplying (17b) by $\widehat{\alpha}_k h$, using (17a), and summing over $k < n$ controls all long suffixes. Combining this bound with (17i)–(17j), and absorbing the explicit factors $\widehat{n}, \Lambda_T$ into $\widetilde{O}$, yields

$$\sum_{t\in J}(\phi_t - \widetilde{g}_t) \le \widetilde{O}\left(\sum_{\ell=1}^{L} C(|\mathcal{I}'_\ell|) + \frac{1}{\gamma}\sum_{k=0}^{n-1}\frac{\rho(2^k)}{\rho(2^n)}C(2^k)\right). \tag{17}$$

Test T2R bounds the sum of D^R, T2C bounds every D_j^C, and (17) bounds the target–predictor component. □

E Full rate assembly

Write

$$C(s) = \min\{c_1\sqrt{s} + c_2, c_3 s\},$$

where

$$c_1 = C_1(1 + b'^{-1})\sqrt{m\ell}, \qquad c_2 = C_2(1 + b'^{-1})m\ell, \qquad c_3 = C_3\max\{1, b'^{-1}\}.$$

Let

$$\Delta = V_1 + \widetilde{V}_2/b', \qquad N_\Delta = 1 + \kappa^{-2/3}c_1^{-2/3}\Delta^{2/3}T^{1/3} + (\kappa c_3)^{-1}\Delta.$$

Lemma 14 (Nominal partition charging)**.** *Let J be a realized prefix of a nominal block and let $\mathcal{I}'_1, \dots, \mathcal{I}'_{L_J}$ be the nonempty truncations of the greedy nominal partition used in Lemma 13. Then*

$$L_J \le 1 + 2c_1^{-2/3}\Delta_J^{2/3}|J|^{1/3} + c_3^{-1}\Delta_J, \tag{17c}$$

and

$$\sum_{\ell=1}^{L_J} C(|\mathcal{I}'_\ell|) \le c_1\sqrt{L_J|J|} + c_2 L_J. \tag{17d}$$

Proof. Only the last nonempty piece can be truncated by the adaptive endpoint. For every earlier piece $I_\ell = [u_\ell, v_\ell]$, greedy maximality implies, with $I_\ell^+ = [u_\ell, v_\ell + 1]$ and $s_\ell = |I_\ell^+|$,

$$\Delta_{I_\ell^+} > \rho(s_\ell) = \min\{c_1 s_\ell^{-1/2} + c_2 s_\ell^{-1}, c_3\}.$$

Classify the boundary by the active branch of this minimum. On the statistical branch,

$$1 < c_1^{-2/3} \Delta_{I_\ell^+}^{2/3} s_\ell^{1/3};$$

on the linear branch, $1 < c_3^{-1} \Delta_{I_\ell^+}$. The edge sets defining the extensions are disjoint and contained in J. Hölder's inequality, together with $\sum_\ell s_\ell \le 2|J|$ and $\sum_\ell \Delta_{I_\ell^+} \le \Delta_J$, gives (17c), including one unchargeable final piece. Finally, $C(s) \le c_1\sqrt{s} + c_2$ and Cauchy–Schwarz give (17d). □

Proposition 15 (Global aggregation)**.** *On the joint event, the number of epochs is at most N_Δ, and both pre-stopping target regret and every scaled resource discrepancy are bounded by $\widetilde{O}(\mathfrak{R}_\gamma)$, with $\mathfrak{R}_\gamma$ given in (18).*

Proof. To derive this count, assign each non-final epoch E_j, of length s_j, to the statistical branch if $\Delta_j > \kappa c_1 s_j^{-1/2}$, and otherwise to the linear branch, where $\Delta_j > \kappa c_3$. Then

$$\sum_{j \in \mathcal{A}} 1 \le \kappa^{-2/3} c_1^{-2/3} \sum_{j \in \mathcal{A}} \Delta_j^{2/3} s_j^{1/3} \le \kappa^{-2/3} c_1^{-2/3} \Delta^{2/3} T^{1/3},$$

by Hölder, while $|\mathcal{B}| \le (\kappa c_3)^{-1} \Delta$.

We next make the blockwise interface explicit. For any realized block $J = [u, e]$, including a reserve-truncated final block, write

$$\sum_{t \in J} (\phi_t - R_t) = \sum_{t \in J} (\phi_t - \widetilde{g}_t) + \sum_{t \in J} D_t^R. \tag{17g}$$

If $J^- = [u, e-1]$ is nonempty, both T2 tests were evaluated and did not fire on J^-; otherwise their sums are zero. The final round obeys $D_e^R \le 1$ and $D_{e,j}^C \le b'^{-1}$. Consequently,

$$\sum_{t \in J} D_t^R \le K_R |J^-| \widehat{\rho}(\max\{1, |J^-|\}) + 1, \qquad \sum_{t \in J} D_{t,j}^C \le K_C |J^-| \widehat{\rho}(\max\{1, |J^-|\}) + b'^{-1}. \tag{17h}$$

The clipped definition of C absorbs these one-round terms, so (17h) is $\widetilde{O}(C(|J|))$. Lemma 13 applies to the first sum in (17g) for any of the stopping-time endpoints defining e; its deferred-reveal proof never conditions on which endpoint occurs. Thus (17), (17g), and (17h) control target regret and every resource coordinate through the reserve-crossing round.

For each realized block J, apply Lemma 14 to the same nominal pieces $\mathcal{I}'_\ell$ used in (17). For the replay term in (17), direct expansion of the clipped envelope gives, for $k < n$,

$$\frac{\rho(2^k)}{\rho(2^n)} C(2^k) \le c \left[c_1 2^{n/2} + \frac{c_2 c_3}{c_1} 2^{n/2} + \frac{c_1^2}{c_3} + \frac{c_2^2}{c_3} 2^{-k} \right]. \tag{17e}$$

Substituting (17c)–(17e), summing geometric block lengths within one epoch, and applying Hölder to its block variations gives

$$\begin{aligned} \mathfrak{r}_\gamma(E) = \widetilde{O}\Big(& c_1^{2/3} \Delta_E^{1/3} s_E^{2/3} + \gamma^{-1}(c_1 + c_2 c_3/c_1)\sqrt{s_E} \\ & + c_1 c_3^{-1/2} \sqrt{\Delta_E s_E} + c_2 c_1^{-2/3} \Delta_E^{2/3} s_E^{1/3} \\ & + c_2 + \gamma^{-1} c_2^2/c_3 + (c_2/c_3)\Delta_E \Big). \end{aligned} \tag{17f}$$

Both pre-stopping target regret and every scaled resource discrepancy are at most $\widetilde{O}(\mathfrak{R}_\gamma)$, where

$$\begin{aligned} \mathfrak{R}_\gamma = {} & c_1^{2/3} \Delta^{1/3} T^{2/3} + \gamma^{-1} \left(c_1 + \frac{c_2 c_3}{c_1} \right) \sqrt{N_\Delta T} \\ & + c_1 c_3^{-1/2} \sqrt{\Delta T} + c_2 c_1^{-2/3} \Delta^{2/3} T^{1/3} \\ & + \left(c_2 + \gamma^{-1} \frac{c_2^2}{c_3} \right) N_\Delta + \frac{c_2}{c_3} \Delta. \end{aligned} \tag{18}$$

Indeed, summing (17f) over at most N_Δ epochs uses

$$\sum_E \sqrt{s_E} \le \sqrt{N_\Delta T}, \quad \sum_E \Delta_E^{1/3} s_E^{2/3} \le \Delta^{1/3} T^{2/3},$$
$$\sum_E \sqrt{\Delta_E s_E} \le \sqrt{\Delta T}, \quad \sum_E \Delta_E^{2/3} s_E^{1/3} \le \Delta^{2/3} T^{1/3},$$

while the linear terms add directly. These inequalities produce every term in (18), including the N_Δ initialization contribution. The $\gamma = 1$ specialization of (18) is the quantity $\mathfrak{R}_\Delta$ used in Theorem 1; terms with larger powers are not hidden as "lower order." Under $\Delta \le V_1 + dV_2/b'$, the four variation structures in (18) yield $d^{1/3}$, $d^{1/2}$, $d^{2/3}$, and through N_Δ up to linear-d dependence. The main theorem uses $\mathfrak{R} = \mathfrak{R}_1$; for every fixed $\gamma > 0$, (18) proves Corollary 5 and preserves all T- and variation-exponents. □

Theorem 16 (Shared-ledger Vector-MASTER reduction)**.** *Consider bounded reward and vector feedback with a zero-feedback null action, targets $\phi_t \in [0,1]$, and an edge-additive interval variation Δ_I. Suppose:*

1. *target sensitivity satisfies (15);*
2. *a fresh, predictably paused base instance outputs $(\widetilde{g}_t, \widetilde{h}_t)$ with $\widetilde{h}_t \le b'\mathbf{1}$ and, on one uniform event, satisfies the active-prefix certificates (8)–(10);*
3. *fixed-dual domination (11) holds on every admissible calendar interval; and*
4. *$C(s) = \min\{c_1\sqrt{s} + c_2, c_3 s\}$ is nondecreasing and $\rho(s) = C(s)/s$ is nonincreasing.*

Apply the block-anchored dyadic schedule with replay factor $\gamma > 0$, coordinatewise restart tests, and one non-resetting ledger that switches to the null action after a coordinate exceeds $B-1$, where $b' = (B-1)/T$. Then, on the same event:

1. *total realized consumption is at most $B\mathbf{1}$ on every sample path;*
2. *the number of restart epochs is at most N_Δ; and*
3. *full-horizon target loss and every pre-switch scaled resource discrepancy are at most $\widetilde{O}(\mathfrak{R}_\gamma)$, with $\mathfrak{R}_\gamma$ in (18).*

Proof. Laminar owner aggregation gives (16), while the deferred-reveal replay argument gives (17). Lemma 14 and restart charging yield Proposition 15. Equation (17h) includes a test- or reserve-truncated final round. The shared ledger never resets, so the one-unit reserve proves pathwise feasibility; Lemma 7 converts any null-arm tail into the already controlled resource discrepancy. □

Lemma 17 (Dynamic benchmark bridge)**.** *The hard-budget dynamic comparator satisfies*

$$\mathrm{OPT}(T) \le \mathrm{LP}_{\mathrm{glob}}(B+1) \le \mathrm{LP}_{\mathrm{glob}}(B) + b^{-1} = \sum_{t=1}^{T} \phi_t(b) + \Gamma_b + b^{-1},$$

and

$$\Gamma_b \le W_1 + 2\bar{q} W_2.$$

Proof. Upper-bound the comparator's reward by also including its crossing pull, and let $x_t \in \Delta_m$ be its unconditional action mixture, with the null arm used after stopping. Whether round t is attempted is measurable before the current outcome table. Immediately before the crossing pull every resource consumption is at most B, and that pull adds at most one, so $\sum_t C_t x_t \le (B+1)\mathbf{1}$ after taking expectations. Moreover, $\sum_t \mu_t^\top x_t$ equals the expected reward of the enlarged sequence. This proves $\mathrm{OPT}(T) \le \mathrm{LP}_{\mathrm{glob}}(B+1)$, without requiring reward and consumption to be independent within a round.

Mixing every real action in a solution for budget $B+1$ with the null arm by weight $B/(B+1)$ gives

$$\mathrm{LP}_{\mathrm{glob}}(B+1) \le \frac{B+1}{B}\mathrm{LP}_{\mathrm{glob}}(B) \le \mathrm{LP}_{\mathrm{glob}}(B) + \frac{1}{b}.$$

For a feasible global-LP solution x_t, set $\bar{x} = T^{-1}\sum_t x_t$. Then

$$\begin{aligned}\sum_t \mu_t^\top x_t &\le T\bar{\mu}^\top \bar{x} + W_1 \\ &\le T\mathrm{LP}(\bar{\mu}, \bar{C}; b) + W_1 + \bar{q}W_2.\end{aligned} \tag{19a}$$

The second line is fixed-dual domination with q^*, because global feasibility and the definition of W_2 imply

$$T(q^*)^\top(\bar{C}\bar{x} - b\mathbf{1}) \le (q^*)^\top \sum_t (\bar{C} - C_t)x_t \le \bar{q}W_2.$$

Finally, let $\bar{x}^*$ solve the time-average LP. Roundwise dual domination and $\bar{C}\bar{x}^* \le b\mathbf{1}$ give

$$\begin{aligned}T\mathrm{LP}(\bar{\mu}, \bar{C}; b) &= \sum_t \mu_t^\top \bar{x}^* \\ &\le \sum_t \phi_t(b) + \sum_t (q_t^*)^\top (C_t - \bar{C})\bar{x}^* \\ &\le \sum_t \phi_t(b) + \bar{q}W_2.\end{aligned} \tag{19b}$$

Here $|q^\top(C_t - \bar{C})x| \le \|q\|_\infty \|C_t - \bar{C}\|_1$ for $x \in \Delta_m$. Combining (19a)–(19b), maximizing over globally feasible $\{x_t\}$, and subtracting $\sum_t \phi_t(b)$ proves $\Gamma_b \le W_1 + 2\bar{q}W_2$. □

Proof of Lemma 7. On an LP round, feasibility gives $\widetilde{h}_{t,j} \le b'$. On a forced initialization round the arm has zero samples in that fresh instance, hence its pessimistic consumption estimate is zero and the same inequality holds. At a crossing time τ,

$$b'T < \sum_{t\le\tau} Z_{t,j^\star} \le b'\tau + \sum_{t\le\tau}(Z_{t,j^\star} - \widetilde{h}_{t,j^\star}).$$

Rearranging proves (12). □

If safety stopping occurs, Lemma 7 bounds the null-arm tail by the controlled resource discrepancy plus a $1/b'$ crossing-round remainder. If it does not occur, the tail is empty. Finally,

$$\phi_t(b') \ge \frac{b'}{b}\phi_t(b) \quad\Longrightarrow\quad \sum_t (\phi_t(b) - \phi_t(b')) \le 1/b.$$

On the joint event, combining this reserve cost, (1), and (18) bounds the realized quantity $\mathrm{OPT}(T) - \sum_t R_t$. Taking expectation with $\delta = T^{-2}$ adds at most δT from the failure event.

F Lower bound for the allocation gap

Both environments below have one resource, one real arm, and the null arm. Outcomes are deterministic. The real arm always consumes one. The null arm stays at zero. The horizon $T \ge 4$ is even, $b \in (0, 1/2]$, and $\eta \in (0, 1/2]$, so $B = bT \le T/2$.

On both paths the real-arm reward is $1/2$ for the first $T/2$ rounds. On E_+ it jumps to $1/2 + \eta$; on E_- it jumps to $1/2 - \eta$. The first-half tables are identical, so any policy induces the same law on its first-half actions under both environments. Write N_1 for the number of real-arm pulls in the first half that are counted, i.e. with $t < \tau_B$, and write N_2 for the corresponding late count. Each counted real pull consumes one and the crossing pull is discarded, so $N_1 + N_2 \le B$, $N_1 \le T/2$, and $N_2 \le T/2$.

On E_+ a planner that knows the tables plays the null arm early and the real arm for B late rounds, and $\mathrm{OPT}(E_+) = B(1/2 + \eta)$. On E_- the late rounds are worse, so the same planner spends the whole budget early and $\mathrm{OPT}(E_-) = B/2$.

A policy earns $N_1/2 + (1/2 + \eta)N_2$ on E_+ and $N_1/2 + (1/2 - \eta)N_2$ on E_-. The budget bound gives $N_2 \le B - N_1$, and $1/2 - \eta \ge 0$, so

$$\begin{aligned}N_1/2 + (1/2 + \eta)N_2 &\le B(1/2 + \eta) - \eta N_1, \\ N_1/2 + (1/2 - \eta)N_2 &\le B(1/2 - \eta) + \eta N_1.\end{aligned}$$

Taking expectations and subtracting from the two OPT values yields

$$\mathrm{Reg}_T(E_+) \geq \eta\,\mathbb{E}[N_1], \qquad \mathrm{Reg}_T(E_-) \geq \eta\big(B - \mathbb{E}[N_1]\big).$$

The first-half laws agree, so the two expectations of N_1 are equal. Adding the displays yields

$$\mathrm{Reg}_T(E_+) + \mathrm{Reg}_T(E_-) \geq \eta B = \eta b T,$$

and therefore (6c).

It remains only to check Γ_b. The per-round program with consumption one and right-hand side $b < 1$ has value $\phi_t(b) = b\mu_t$. On E_+

$$\sum_t \phi_t(b) = \frac{bT}{2} + \frac{\eta bT}{2}, \qquad \mathrm{LP}_{\mathrm{glob}}(B) = B\Big(\tfrac{1}{2} + \eta\Big),$$

so $\Gamma_b = \eta bT/2$. On E_- the sum is $bT/2 - \eta bT/2$ and $\mathrm{LP}_{\mathrm{glob}}(B) = B/2$, and the same value of Γ_b follows. One jump of size η gives $V_1 = \eta$ and $V_2 = 0$, so $\Delta = \eta$ and the adaptive term in (5) is $\widetilde{O}(\eta^{1/3}T^{2/3})$.

The same late jump with $b = 1$ lets every round play the real arm. Then $\phi_t(1) = \mu_t$ and $\mathrm{LP}_{\mathrm{glob}}(T) = \sum_t \mu_t$, so $\Gamma_b = 0$. There is no budget to move.

Proof of Proposition 4. Keep one real arm and set every consumption to zero. Rewards are 1 on the first $T/2$ rounds and 0 afterward. Then $\phi_t(b) = \mu_t$, the concatenated local solution is globally optimal, and $\Gamma_b = 0$. The time average is $1/2$, so $W_1 = T/2$, while $V_1 = 1$ and $V_2 = 0$.

On E_+ the time-average reward is $(1+\eta)/2$, hence $W_1 = \eta T/2$ and $\Gamma_b = bW_1$. The same identity holds on E_-. □

G Additional experimental details

The confidence multiplier 0.2 and the T2R/T2C/T1 thresholds 1.5, 3, and 0.35 were frozen on diagnostic seeds 0–49, before the screens, oracle tuning, and focused evaluation. The freeze was manual.

All methods use paired environment seeds and a separately seeded action sampler. One- and two-resource per-round LPs use vertex enumeration, checked against HiGHS; the global fluid LP uses HiGHS.

The focused comparison is in Section 6. The screens below use the same tests with a replay probability γ, and one extra schedule of periodic probes that is not in the theorem. Probe bursts have a random global phase, period $\lceil T^{2/3} \rceil$, and dyadic length equal to the smallest power of two at least $\max\{16, \lceil 4 \log_2 T \rceil\}$.

The stationary screen at $T = 4096$ has four real arms, one null arm, and one resource. For each paired seed, fixed reward means are drawn uniformly from $[0.2, 0.9]$, fixed consumption means from $[0.1, 0.8]$, and outcomes are Bernoulli. The hidden-arm screens are deterministic with two real arms and a null arm: rewards are $(0.6, 0, 0)$ before the change and $(0.6, 1, 0)$ afterward, with zero consumption; changes occur at $0.57T$ and $0.73T$.

Table 2: Replay screens at $T = 4096$, $b = 0.35$, five paired seeds per cell. Reward is pseudo-reward over the global fluid LP. Adoption is the terminal share of the arm that improves at $0.57T$.

Schedule	Stat. reward	Hidden reward	Hidden replay	Adoption
$\gamma = 1$	.711	.874	.982	.806
$\gamma = 1/2$	.824	.911	.732	.884
$\gamma = 1/4$	.854	.884	.482	.721
Periodic probes	.914	.991	.225	.996
No random replay	.920	.774	.000	.0002

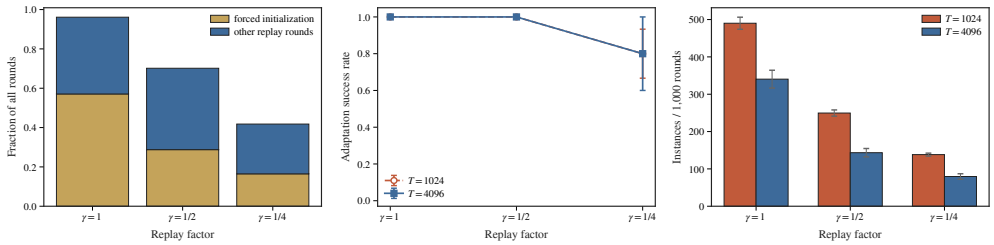


Figure 3: Replay overhead at $T = 4096$: owner-round decomposition, hidden-arm adoption, and scheduled instances. Thinning primarily reduces repeated initialization.

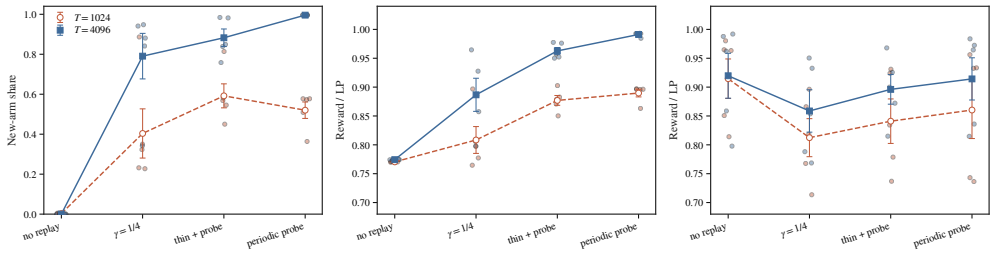


Figure 4: Exploratory alternatives for obtaining fresh data. Periodic probes recover the hidden arm with lower finite-horizon overhead, but are outside the theorem.

H Counterexamples motivating the design

No mandatory fresh replay. The hidden-improving-arm construction in Example 1 gives $\Omega(T)$ regret with $V_1 = 1$ for any specified wrapper that eventually abandons an arm and uses only played-arm certificates.

Independent partial resets. Let $b = 1/2$, arm A have $(\mu, C) = (3/4, 1/2)$, and arm B change from $(0, 1)$ to $(1, 0)$. A reward-only reset inherits the stale high cost of B; a consumption-only reset inherits its stale low reward. Neither selects B. If jointly fresh state is obtained only by intersecting two independent replay events of probabilities $\sqrt{h/H}$, the useful probability is h/H, creating a near-linear waiting time. This rules out the naive separate-scale extension, not all anisotropic algorithms.

I AI use statement

Generative AI tools assisted with editing, manuscript organization, and simulation coding. The authors independently wrote the proofs and take responsibility for the manuscript.