MEMORY–BATCH TRADE-OFFS IN LIPSCHITZ BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Lipschitz bandits admit near-optimal regret $\widetilde{O}_d(T^{(d+1)/(d+2)})$ with little memory under full adaptivity, or with few batches under unrestricted memory. We characterize the minimax expected pseudo-regret over T rounds with W bits of memory and at most B batches in dimension d. For every memory budget W, we prove a lower bound of order $T^{\frac{d+2}{d+3}}(1+(B-1)W)^{-\frac{1}{d(d+3)}}$. When $W \gtrsim_d \log(eT)$, algorithms with fixed batch boundaries attain, up to logarithmic factors, the larger of this bound and the optimal B-batch regret with unrestricted memory. The analysis separates fine-scale comparisons from the regional information needed to allocate their samples at low regret. Attaining $\widetilde{O}_d(T^{(d+1)/(d+2)})$ regret requires both $B = \Omega_d(\log\log T)$ and $(B-1)W = \widetilde{\Omega}_d(T^{d/(d+2)})$. With one bit of memory, minimax regret is $\widetilde{\Theta}_d(TB^{-1/(d+2)})$ under adaptive batch boundaries while $\Theta_d(T)$ under fixed boundaries.

1 INTRODUCTION

Memory-limited bandits restrict how much information from past actions and rewards a learner can retain (Cover, 1968; Liau et al., 2018); batched bandits restrict when this information can affect its actions (Perchet et al., 2016; Gao et al., 2019).

We study these two constraints jointly in stochastic Lipschitz bandits on $[0,1]^d$ over T rounds (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007; Bubeck et al., 2011). Near-optimal regret $\widetilde{O}_d(T^{(d+1)/(d+2)})$ is achievable with $O_d(\log(eT))$ bits of memory under full adaptivity (Zhu & Huang, 2025), or with $O_d(\log\log T)$ batches under unrestricted memory (Feng et al., 2024). We ask:

How does minimax regret depend jointly on memory and batch constraints in Lipschitz bandits?

In the jointly constrained model, the learner uses at most B batches and maintains at most W bits of reward-dependent mutable state after every observation. Within a batch, rewards may update the state but cannot alter that batch's prescribed action sequence or boundary.

For K-armed bandits, Huang et al. (2026) show that, under suitable conditions on W and T, $\widetilde{\Theta}(K/W)$ batches are necessary and sufficient for $\widetilde{O}(\sqrt{KT})$ regret with W bits of memory. For Lipschitz bandits, directly applying this result to a grid makes the number of actions depend on the desired approximation accuracy. For some memory and batch budgets, our two-scale construction achieves a polynomially smaller regret upper bound than the finite-arm guarantee optimized over regular grid sizes (Section 5.4).

Theorem 1.1 (Informal main theorem; see Theorem 3.1)**.** *Fix an integer $d \geq 1$. For sufficiently large T, $1 \leq B \leq T$, and $W \gtrsim_d \log(eT)$, the minimax expected pseudo-regret in the model above satisfies*

$$\mathfrak{R}_T(B,W) = \widetilde{\Theta}_d\left(T^{\frac{d+1}{d+2}} \vee \frac{T^{\beta_{d,B}}}{B^2} \vee T^{\frac{d+2}{d+3}}\left(1+(B-1)W\right)^{-\frac{1}{d(d+3)}}\right),$$

where $\beta_{d,B} := (d+1)/\{(d+2)[1-(d+2)^{-B}]\}$. The lower bound holds up to a d-dependent constant for all $W \geq 0$, even with adaptively chosen batch boundaries; the upper bound uses fixed boundaries.

The first two terms characterize the unrestricted-memory B-batch regret up to logarithmic factors. The third captures the additional memory constraint and, for fixed $B \geq 2$, decreases with W. Section 3.2 identifies the resulting regret regimes.

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.

Contributions.

(i) **Minimax memory–batch tradeoff.** We characterize the minimax expected pseudo-regret up to logarithmic factors for $W \gtrsim_d \log(eT)$. In particular, logarithmic memory requires $\widetilde{\Omega}_d(T^{d/(d+2)})$ batches for near-optimal regret, polynomially more than the $O_d(\log\log T)$ batches sufficient with unrestricted memory.

(ii) **One-bit regret and adaptive batch boundaries.** Below logarithmic memory, the three-term maximum need not be tight, and fixed boundaries can incur a polynomial loss. For $W = 1$, minimax regret is $\widetilde{\Theta}_d(TB^{-1/(d+2)})$ with adaptive boundaries, but $\Theta_d(T)$ with fixed boundaries, for every $1 \le B \le T$ (Section 6).

1.1 Technique overview

Applying Huang et al. (2026, Theorem 2) to a regular grid of resolution r uses $K \asymp_d r^{-d}$ actions and gives an approximation term $O_d(Tr)$. Under its assumptions and suppressing logarithmic factors, the resource condition $K \lesssim BW$ requires $r \gtrsim_d (BW)^{-1/d}$. Thus, this guarantee ties finer approximation to a larger memory–batch budget.

We use coarse cells of diameter $O_d(s)$ to guide comparisons at resolution $r \le s$. By Lipschitz continuity, if a cell's representative has gap $O_d(s)$, every point in the cell also has gap $O_d(s)$. Fine comparisons therefore cost $O_d(s)$ regret per pull (Figure 1).

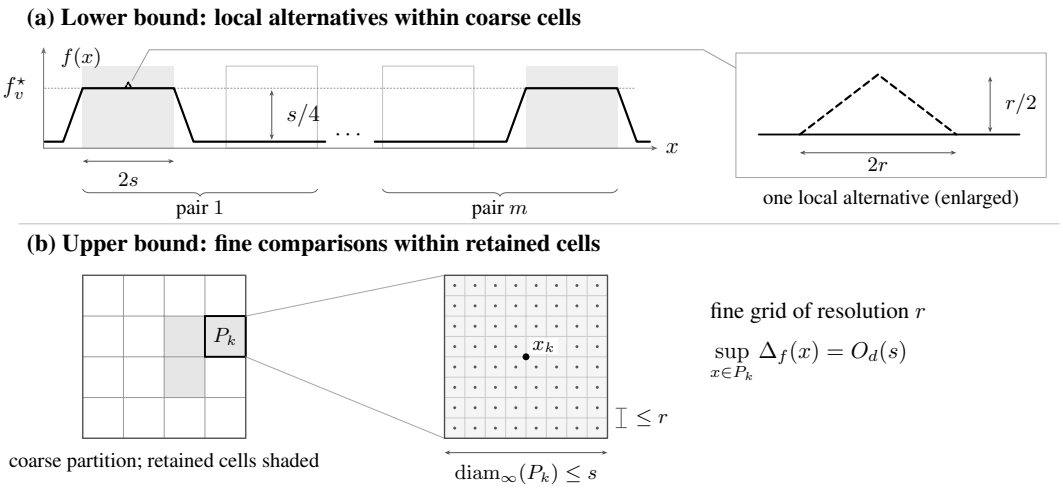

Figure 1: Two-scale geometry. **(a)** One cell in each of $m \asymp_d s^{-d}$ pairs is optimal under the base instance; the other has gap $s/4$. A local alternative adds a bump of height $r/2$ and radius $r \le s/16$ within an optimal cell. **(b)** On the confidence event, retained cells collectively contain a maximizer and have gap $O_d(s)$ throughout; their fine grids have resolution r.

Lower bound: allocating local tests is informative. At admissible scales $r \le s/16$, the hard family has $\asymp_d s^{-d}$ pairs of cells, each with one optimal cell and one with gap $\Theta(s)$. Independent uniform bits select the optimal cells, which contain $\asymp_d r^{-d}$ disjoint radius-r neighborhoods. Each neighborhood supports a local improvement of height $\Theta(r)$. Although one optimal cell suffices on a base instance, regret below $c_d Tr$ on every local alternative forces $\Omega(r^{-2})$ visits to each selected neighborhood with constant probability under the base law. The resulting $\Omega_d(r^{-d-2})$ sample requirement yields the sequential lower bound when it exceeds T.

Under a base instance, testing the unselected cells equally extensively would incur $\Omega_d(sr^{-d-2})$ regret. If expected regret on every base instance is also below $c_d sr^{-d-2}$, visit counts predict the regional choices with a constant average advantage over chance, revealing $\Omega_d(s^{-d})$ bits. Yet the boundary states determine the actions and carry at most $(B-1)W$ bits conditional on public randomness (Huang et al., 2026). Thus, at admissible scales, $s^{-d} \ge C_d(B-1)W$ gives the lower bound

$\Omega_d(\min\{Tr, sr^{-d-2}\})$ (Section 4). The batch-depth term follows from the unrestricted-memory lower bound of Feng et al. (2024, Theorem 8), adapted to our reward model.

Upper bound: reuse memory across regions. For $W \gtrsim_d \log(eT)$, the first construction processes $O_d(s^{-d})$ coarse cells in small groups (Roy Chaudhuri & Kalyanakrishnan, 2020). It reconstructs each group's *active set*, the cells not yet eliminated, through comparisons with a retained reference (Liau et al., 2018). On a confidence event, surviving cells contain a maximizer and have gap $O_d(s)$ throughout. We compare fine actions within these cells and retain the best estimated action across groups. For feasible scales, fine comparisons and exploitation incur $\widetilde{O}_d(sr^{-d-2}+Tr)$ expected regret; active-set reconstruction fits within the same bound. Processing the groups uses $\widetilde{O}_d(1 + s^{-d}/W)$ batches in total. Balancing Tr and sr^{-d-2} yields the joint term in both bounds. A second construction retains active sets across spatial levels to use fewer batches. Both use fixed boundaries and the same final refinement; Section 5 bounds their construction costs.

One-bit memory. For one bit of memory, we additionally bound how well the state retained at a batch boundary can distinguish local alternatives, uniformly of the batch length. The matching policy chooses batch lengths so that comparison outcomes can be recovered from later round and batch indices (Section 6).

2 Related work

Lipschitz bandits and limited adaptivity. Classical continuum-armed and metric bandit results relate regret to the geometry of the action space and its near-optimal regions (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007; Kleinberg et al., 2008; Bubeck et al., 2011; Magureanu et al., 2014). Related partition-based methods appear in tree search and deterministic optimistic optimization (Coquelin & Munos, 2007; Munos, 2011), and adaptive discretization extends to adversarial Lipschitz bandits (Podimata & Slivkins, 2021). Batched feedback has been studied for finite arms, linear bandits, and other structured models (Perchet et al., 2016; Gao et al., 2019; Jin et al., 2021b; Esfandiari et al., 2021; Ruan et al., 2021; Hanna et al., 2023; Sawarni et al., 2024), alongside broader restrictions on interaction in limited-round learning (Agarwal et al., 2017). Feng et al. (2024) study batched Lipschitz bandits without memory constraints, with regret bounds depending on zooming dimension. Our joint characterization concerns the full Lipschitz class on $[0, 1]^d$.

Memory-constrained and streaming bandits. Memory bounds use different units and access rules: state configurations and bits (Cover, 1968; Cover & Hellman, 1970; Zhu & Huang, 2025), stored words (Liau et al., 2018), or retained arms (Roy Chaudhuri & Kalyanakrishnan, 2020; Maiti et al., 2021; Assadi & Wang, 2020; Agarwal et al., 2022; Wang, 2023; Assadi & Wang, 2024). In multi-pass streaming, an evicted arm cannot be sampled again until it reappears in a later pass, although sampling can adapt within a pass (Agarwal et al., 2022). Consequently, pass bounds concern restrictions on revisiting arms and do not directly yield bounds on batched action updates. Communication constraints differ as well: Lau et al. (2026) restrict each batch's returned message to one bit but allow decisions to use the history of received bits, rather than imposing a fixed mutable-state budget. The sequential Lipschitz algorithm of Zhu & Huang (2025) revisits regions through iterative deepening, deciding whether to explore a cell's descendants after sampling it. Here that decision can affect actions only in a later batch. Other related formulations study memory-constrained adversarial bandits (Xu & Zhao, 2021), bounded-space online learning (Srinivas et al., 2022; Peng & Zhang, 2023; Peng & Rubinstein, 2023), and streaming tradeoffs (Jin et al., 2021a; Li et al., 2023).

Joint memory and adaptivity constraints. We build on the finite-arm analysis of Huang et al. (2026), whose hard family already separates a base gap from a smaller local improvement. Our lower bound combines their boundary-state argument with localized change-of-measure comparisons based on classical likelihood-ratio tools (Mannor & Tsitsiklis, 2004; Kaufmann et al., 2016). Our upper bound draws on repeated comparisons with a retained reference (Liau et al., 2018) and processing candidates in small groups (Roy Chaudhuri & Kalyanakrishnan, 2020). Related blockwise comparisons under joint memory and batch constraints appear in Huang et al. (2026, Section 5). Our construction keeps the reference fixed across an entire pass and uses the final selections to guide finer comparisons within Lipschitz cells.

3 MODEL AND MINIMAX REGRET

3.1 BATCHED POLICIES WITH LIMITED MEMORY

Throughout, d, T, B, W are integers satisfying $d, T \geq 1$, $1 \leq B \leq T$, and $W \geq 0$.

Lipschitz Bandits. Let $\mathcal{X} = [0,1]^d$ with the sup norm, and let $\mathrm{Lip}_1(\mathcal{X})$ denote the class of one-Lipschitz functions from $\mathcal{X}$ to $[0,1]$. For $f \in \mathrm{Lip}_1(\mathcal{X})$, let $\mathfrak{D}(f)$ consist of families of reward distributions $\nu = (\nu_x)_{x\in\mathcal{X}}$ on $[0,1]$, with mean $f(x)$ and measurable dependence on x. At round t, the learner selects an action $A_t \in \mathcal{X}$ and observes a reward Y_t whose conditional distribution, given the past and A_t, is ν_{A_t}. Define

$$f^\star := \sup_{x\in\mathcal{X}} f(x), \qquad \Delta_f(x) := f^\star - f(x), \qquad \mathrm{Reg}_T(f) := \sum_{t=1}^{T} \Delta_f(A_t).$$

Policies. Let ω be an independent public random seed. The complete reward-dependent mutable state satisfies $M_t \in \{0,1\}^W$ for $t = 0, \ldots, T$, with $M_0 = \iota(\omega)$. A policy partitions the T rounds into at most B nonempty batches, with $\tau_0 = 0$. Each batch boundary is chosen adaptively at the start of that batch. The policy is specified by measurable maps satisfying

$$\tau_{b+1} = \psi_b(\omega, \tau_b, M_{\tau_b}), \tag{1}$$
$$A_t = \phi_{b,t}(\omega, \tau_b, M_{\tau_b}), \tag{2}$$
$$M_t = \mathcal{U}_{b,t}(\omega, M_{t-1}, A_t, Y_t), \qquad \tau_b < t \leq \tau_{b+1}.$$

The parameters d, T, B, W, public random seed ω, round and batch indices, and the current batch's fixed boundary and action sequence remain available as M_t changes and do not count toward W. Boundaries fixed in advance depend only on ω. The model follows Huang et al. (2026); Appendix A gives the full specification.

Minimax regret. Let $\mathfrak{A}_{B,W}$ denote the class of policies defined above. The minimax expected pseudo-regret is

$$\mathfrak{R}_T(B,W) := \inf_{\mathcal{A}\in\mathfrak{A}_{B,W}} \sup_{f\in\mathrm{Lip}_1(\mathcal{X}),\, \nu\in\mathfrak{D}(f)} \mathbb{E}_{\nu,\mathcal{A}}\big[\mathrm{Reg}_T(f)\big],$$

where the expectation is over reward and policy randomness.

3.2 MINIMAX REGRET BOUNDS

Write $\ell_T := \log(eT)$, $\alpha_d := (d+1)/(d+2)$, and $\beta_{d,B} := (d+1)/\big((d+2)[1-(d+2)^{-B}]\big)$. We use $[n] = \{1, \ldots, n\}$, $\vee$ for maximum, and $\wedge$ for minimum. Unsubscripted logarithms are natural. The relations $\lesssim_d$, $\gtrsim_d$, and $\asymp_d$ hide constants depending only on d; a tilde additionally hides powers of $\log(eT)$. Unless fixed in a statement, $c_d, C_d > 0$ may change between occurrences. We fix a deterministic action $x_\circ \in \mathcal{X}$ for padding and fallback choices.

Theorem 3.1 (Minimax regret under memory and batch constraints)**.** *There exist constants $c_d, C_d, T_d > 0$, depending only on d, such that, for every $T \geq T_d$ and all B, W,*

$$\mathfrak{R}_T(B,W) \geq c_d \left[T^{\frac{d+1}{d+2}} \vee \frac{T^{\beta_{d,B}}}{B^2} \vee T^{\frac{d+2}{d+3}} \big(1 + (B-1)W\big)^{-\frac{1}{d(d+3)}} \right]. \tag{3}$$

If $W \geq C_d \log(eT)$, then

$$\mathfrak{R}_T(B,W) \leq C_d \log(eT) \left[T^{\frac{d+1}{d+2}} \vee \frac{T^{\beta_{d,B}}}{B^2} \vee T^{\frac{d+2}{d+3}} \big(1 + (B-1)W\big)^{-\frac{1}{d(d+3)}} \right]. \tag{4}$$

The lower bound allows batch boundaries chosen adaptively at batch starts; the upper bound is attained with fixed batch boundaries.

For smaller memory the lower bounds remain valid but need not be tight; Section 6 gives the one-bit rate and its fixed-boundary separation.

Removing the memory constraint gives the policy class $\mathfrak{A}_{B,\infty}$ and minimax regret $\mathfrak{R}_T^{\mathrm{bat}}(B)$, with

$$\mathfrak{R}_T^{\mathrm{bat}}(B) = \widetilde{\Theta}_d\left(T^{\alpha_d} \vee \frac{T^{\beta_{d,B}}}{B^2}\right), \qquad 1 \le B \le T. \tag{5}$$

The exponent $\beta_{d,B}$ matches that in the lower bound of Feng et al. (2024). For sufficiently large C_d, $B \ge C_d \log\log T$ gives $\mathfrak{R}_T^{\mathrm{bat}}(B) = \widetilde{\Theta}_d(T^{\alpha_d})$.

Dominant terms and resource tradeoffs. Assume $W \ge C_d \ell_T$ and write $Q_{T,B} := T^{\beta_{d,B}-\alpha_d}/B^2$. Relative to T^{α_d}, the three terms in (3) are

$$1, \qquad Q_{T,B}, \qquad \left(\frac{T^{d/(d+2)}}{1+(B-1)W}\right)^{\frac{1}{d(d+3)}}. \tag{6}$$

Table 1: Dominance conditions, including ties. A dash means no additional restriction.

Dominant term	**Batch condition**	**Memory condition**
Sequential	$Q_{T,B} \le 1$	$1+(B-1)W \ge T^{d/(d+2)}$
Batch-depth	$Q_{T,B} \ge 1$	$1+(B-1)W \ge T^{d/(d+2)} Q_{T,B}^{-d(d+3)}$
Joint	—	$1+(B-1)W \le T^{d/(d+2)}(1 \vee Q_{T,B})^{-d(d+3)}$

For fixed $B \ge 2$, once the joint term is at most $T^{\alpha_d}(1 \vee Q_{T,B})$, the maximum no longer depends on W, and further memory can improve minimax regret by at most logarithmic factors. For fixed W, increasing B can move from batch-depth to joint and then sequential dominance, possibly skipping the joint regime. Appendix D.2 derives the boundaries and shows why all three terms are needed.

Batch Complexity Corollary. For $\varrho > 0$, define $\mathcal{B}_T(W;\varrho) := \min\{B \in [T] : \mathfrak{R}_T(B,W) \le \varrho \ell_T T^{\alpha_d}\}$, with $\min \varnothing := +\infty$.

Corollary 3.2 (Batch complexity for near-optimal regret)**.** *For a dimension-dependent constant $\varrho_d > 0$, all sufficiently large T, and $W \gtrsim_d \log(eT)$,*

$$\mathcal{B}_T(W;\varrho_d) = \widetilde{\Theta}_d\left(\log\log T \vee \frac{T^{d/(d+2)}}{W}\right). \tag{7}$$

Appendix D.3 gives the explicit logarithmic factors.

4 Lower Bounds

The regional family in Figure 1(a) yields the sequential and joint lower bounds. A separate batching argument completes (3).

4.1 Two-scale hard family

Fix $0 < s \le s_d$ and $0 < r \le s/16$, with $s_d > 0$ sufficiently small depending only on d. Choose $2m \asymp_d s^{-d}$ centers $x_{j,a} \in [1/4, 3/4]^d$, indexed by $j \in [m]$ and $a \in \{0,1\}$, separated by at least $8s$ in the sup norm, and define the cells $P_{j,a} := \{x \in \mathcal{X} : \|x - x_{j,a}\|_\infty \le s\}$. In each cell, choose the same number $q \asymp_d (s/r)^d$ of centers $z_{j,a,k}$ satisfying $\|z_{j,a,k} - x_{j,a}\|_\infty \le s/4$ and pairwise separation at least $4r$. The neighborhoods $G_{j,a,k} := \{x \in \mathcal{X} : \|x - z_{j,a,k}\|_\infty \le r\}$ are disjoint and contained in their parent cells.

Write $(u)_+ := \max\{u, 0\}$ and $d_\infty(x,S) := \inf_{y \in S} \|x - y\|_\infty$ for nonempty S.

Definition 4.1 (Base and local Bernoulli instances)**.** For $v \in \{0,1\}^m$, define the base mean and its local alternatives by

$$f_v(x) := \frac{1}{4} + \max_{j \in [m]} \left(\frac{s}{4} - \frac{1}{2} d_\infty(x, P_{j,v_j})\right)_+,$$

$$f^+_{v,j,k}(x) := f_v(x) + \frac{1}{2}\left(r - \|x - z_{j,v_j,k}\|_\infty\right)_+, \qquad j \in [m],\ k \in [q]. \tag{8}$$

The instances have Bernoulli rewards with these means.

Under f_v, the bit v_j selects an optimal cell in each pair: $f_v = 1/4 + s/4$ on P_{j,v_j}, whereas $f_v = 1/4$ on $P_{j,1-v_j}$. The local alternative adds a bump of height $r/2$ within $G_{j,v_j,k}$, making $z_{j,v_j,k}$ uniquely optimal. Outside this neighborhood, the reward law is unchanged and the gap under the alternative is at least $r/2$. All means are one-Lipschitz and lie in $[1/4, 3/4]$. Appendix B.1 gives the packing and verifies these properties. The construction separates the local improvement $r/2$ from the regret cost $s/4$ of sampling an unselected cell, with $mq \asymp_d r^{-d}$ selected neighborhoods across $m \asymp_d s^{-d}$ regional choices.

4.2 A TWO-SCALE REGRET LOWER BOUND

Lemma 4.2 (Two-scale lower bound)**.** *There exist dimension-dependent constants $c_d, C_d > 0$ such that, for the scales above with $Tr^2 \geq C_d$, if*

$$(B-1)W \leq c_d s^{-d}, \tag{9}$$

then

$$\mathfrak{R}_T(B, W) \geq c_d \min\{Tr, sr^{-d-2}\}. \tag{10}$$

The bound allows batch boundaries chosen adaptively at batch starts.

Proof sketch. Fix a policy $\mathcal{A} \in \mathfrak{A}_{B,W}$, and let $\mathcal{R}$ be its largest expected regret over the base instances and local alternatives. Suppose $\mathcal{R} < c_d \min\{Tr, sr^{-d-2}\}$. Let V be uniform on $\{0,1\}^m$, independently of ω, and write $\overline{\mathbb{P}}, \overline{\mathbb{E}}$ for probability and expectation under the mixture of base instances f_V.

Since $\mathcal{R} < c_d Tr$, Corollary B.2 gives $n \asymp r^{-2}$ such that each selected neighborhood receives at least n visits with probability at least $5/8$. Set $\rho_{j,a} := q^{-1} \sum_{k=1}^{q} \mathbf{1}\{N_{j,a,k}(T) \geq n\}$, where $N_{j,a,k}(T)$ counts visits to $G_{j,a,k}$, and mark $P_{j,a}$ when $\rho_{j,a} \geq 1/4$. Then $\overline{\mathbb{E}}\rho_{j,V_j} \geq 5/8$, while each marked unselected cell incurs at least $sqn/16$ regret. Thus

$$\overline{\mathbb{P}}(\rho_{j,V_j} \geq 1/4) \geq \frac{1}{2}, \qquad \frac{1}{m}\sum_{j=1}^{m} \overline{\mathbb{P}}(\rho_{j,1-V_j} \geq 1/4) \leq \frac{16\mathcal{R}}{msqn} \leq \frac{1}{8},$$

using $\rho_{j,a} \leq 1$, $msqn \asymp_d sr^{-d-2}$, and $\mathcal{R} < c_d sr^{-d-2}$ with sufficiently small c_d.

Choose the uniquely marked side as $\widehat{V}_j$, with independent fair tie-breaking otherwise. Its average error is at most $5/16$. Conditional data processing and the binary entropy bound give

$$I(V; A_{1:T} \mid \omega) \geq I(V; \widehat{V} \mid \omega) \geq m\big[1 - h_2(5/16)\big] \geq c'_d s^{-d}, \tag{11}$$

where h_2 is binary entropy in bits and $c'_d > 0$ depends only on d. By (19), the boundary-state argument of Huang et al. (2026) gives $I(V; A_{1:T} \mid \omega) \leq (B-1)W \leq c_d s^{-d}$. Choosing $c_d < c'_d$ gives a contradiction. Since the policy was arbitrary, (10) follows. Appendix B.2 gives the full calculation. □

Completing the lower bound. Balancing Tr and sr^{-d-2} gives $r \asymp_d (s/T)^{1/(d+3)}$. With $s \asymp_d (1 + (B-1)W)^{-1/d}$, this yields the joint term in (3) whenever these scales are admissible. Appendix B.2 completes the sequential and joint bounds over all memory and batch budgets.

Even with unrestricted memory, the learner can adapt to rewards only between batches. The adaptive-grid lower bound of Feng et al. (2024, Theorem 8) quantifies this restriction, giving regret of order $T^{\beta_{d,B}}/B^2$. Lemma B.3 transfers this bound to Bernoulli rewards and policies with at most B batches for $B \geq 2$. Appendix B completes the proof of (3), including $B = 1$.

5 UPPER BOUNDS

We prove (4) for $W \geq C_d \ell_T$ using fixed batch boundaries. We first analyze refinement within a selected set of coarse cells, then give two ways to construct this set under different memory and batch budgets.

5.1 Refinement within active cells

Fix $0 < r \le s \le 1$. Let $\mathcal{P}_s = \{P_1, \dots, P_K\}$ be a fixed dyadic partition of $\mathcal{X}$ into $K \asymp_d s^{-d}$ cells of diameter at most s, with midpoint representatives x_k. Subdividing each cell to side length at most r gives an r-net $\mathcal{N}_r(P_k)$ of descendant midpoints, with $O_d((s/r)^d)$ points per cell. Write $\ell_r := \log(e/r)$.

Definition 5.1 (Safe active set)**.** Fix a sufficiently large dimension-dependent constant κ_d. An active set $\mathcal{C} \subseteq \mathcal{P}_s$ is safe for f if

$$\left(\bigcup_{P\in\mathcal{C}} P\right) \cap \arg\max_{x\in\mathcal{X}} f(x) \neq \varnothing, \qquad \sup_{P\in\mathcal{C}} \sup_{x\in P} \Delta_f(x) \le \kappa_d s.$$

We represent $\mathcal{C}$ by its indicator vector $Z = (\mathbf{1}\{P_k \in \mathcal{C}\})_{k=1}^K$.

Given a safe active set, reserve a block of $O_d(r^{-2}\ell_r)$ pulls for each fine-grid action. In blocks belonging to inactive cells, play a coarse representative $\bar{x}$ whose index is available from the stored state and whose gap satisfies $\Delta_f(\bar{x}) \le \kappa_d s$. This keeps the block schedule fixed in advance. Process active-action estimates one at a time, retaining only the largest estimated mean and its action, then exploit that action. This uses $O_d(\ell_r)$ auxiliary bits beyond Z.

Every refinement pull has gap $O_d(s)$, and the retained cells contain an $O_d(r)$-optimal fine-grid action. If $C_d r^{-d-2}\ell_r \le T/2$ and the complete schedule finishes by T, refinement and exploitation therefore have expected regret $O_d(sr^{-d-2}\ell_r)$ and $O_d(Tr)$, respectively, including estimation failures. Cells may be selected and refined one group at a time, provided their union is safe and the recorded action and estimate are retained across groups. Write $\mathrm{Reg}_{\mathrm{build}}(f)$ for construction regret and $\mathrm{Reg}_{\mathrm{ref+exp}}(f)$ for refinement and exploitation regret. Appendix C.1 proves the conditional guarantee used by both constructions below.

5.2 Active-set construction in groups

Storing one group's indicators at a time gives the following counterpart to Lemma 4.2.

Lemma 5.2 (Two-scale upper bound)**.** *There exist constants $C_d, T_d > 0$, depending only on d, such that, for $T \ge T_d$ and scales $0 < r \le s \le 1/16$, the conditions*

$$W \ge C_d \ell_T, \qquad B - 1 \ge C_d\big(1 + s^{-d}/W\big)(1 + \log \ell_T)^2, \qquad C_d r^{-d-2}\ell_T \le T/2, \tag{12}$$

imply the existence of a policy $\mathcal{A} \in \mathfrak{A}_{B,W}$ with fixed batch boundaries satisfying

$$\sup_{f\in \mathrm{Lip}_1(\mathcal{X}),\, \nu\in\mathfrak{D}(f)} \mathbb{E}_{\nu,\mathcal{A}}\big[\mathrm{Reg}_T(f)\big] \le C_d\big[sr^{-d-2}\ell_T + Tr\big]. \tag{13}$$

Construction. Partition the coarse cells into groups whose indicator vectors fit in memory together with $O_d(\ell_T)$ auxiliary bits. Let L_{ser} be the number of passes. Each pass visits every group, initializing its indicators to one and reconstructing its active set from fresh samples.

A group's best representative need not be near-optimal globally. The common reference allows comparisons across groups without storing their indicator vectors together. Choose a fixed initial reference x_{k_0}, with $k_0 \in [K]$. In pass i, use the common reference action $x_{k_{i-1}}$. A prescribed prefix of the first batch estimates its mean; let λ_i be the resulting lower confidence bound, held fixed for the rest of the pass. For an active candidate k, let $\widehat{\mu}_{k,\ell}$ and a_ℓ denote its estimated mean and confidence radius at comparison step $\ell = 1, \dots, i$. Suppressing pass and group indices, update

$$Z_{k,\ell} = Z_{k,\ell-1}\mathbf{1}\{\widehat{\mu}_{k,\ell} + a_\ell \ge \lambda_i - \zeta_i\}, \qquad \zeta_i := s\mathbf{1}\{i = L_{\mathrm{ser}}\}. \tag{14}$$

Inactive entries remain zero, and their prescribed sampling blocks use the reference action. Each nonfinal pass retains the surviving candidate with largest final lower confidence bound as the next reference. In the final pass, refine each group's selected cells before discarding its indicator vector, retaining the largest estimated mean and its action across groups.

Proof sketch. Write $f_K^\star := \max_{k\le K} f(x_k)$. On the joint confidence event $\mathcal{G}$, a best representative survives each pass, and the reference after each nonfinal pass i satisfies $f_K^\star - f(x_{k_i}) \le 2a_i$. Since

$a_{L_{\mathrm{ser}}-1}, a_{L_{\mathrm{ser}}} = O_d(s)$, the final reference lower bound satisfies $f_K^\star - C_d s \le \lambda_{L_{\mathrm{ser}}} \le f_K^\star$. With the final threshold lowered by s, the union $\mathcal{C}$ of selected cells satisfies

$$\{P_k : f_K^\star - f(x_k) \le s\} \subseteq \mathcal{C} \subseteq \{P_k : f_K^\star - f(x_k) \le C_d s\}.$$

The first inclusion preserves a maximizer's cell. The second, together with $f^\star - f_K^\star \le s$ and Lipschitz continuity, bounds the gap throughout each retained cell by $O_d(s)$.

Larger sample sizes are used only for candidates that survive earlier comparisons. The schedule in Appendix C.2 gives $\mathrm{Reg}_{\mathrm{build}}(f) \le C_d s^{-d-1}\ell_T$ on $\mathcal{G}$. Since $r \le s$, this is bounded by the fine-comparison term in (13). Combining (36) with $\mathbb{P}(\mathcal{G}^{\complement}) \le r$ and $\ell_r \le C_d \ell_T$ proves the regret bound. Let $J_s = O_d(1 + s^{-d}/W)$ be the number of groups. With $L_{\mathrm{ser}} = O(1 + \log \ell_T)$ passes, the total number of batches is $J_s\big(L_{\mathrm{ser}}(L_{\mathrm{ser}}+1)/2 + 1\big) + 1$: i comparison batches per group in pass i, one final refinement batch per group, and one exploitation batch. This gives (12); the full proof is in Appendix C.2. □

5.3 Hierarchical active-set construction

Reconstructing active sets saves memory but costs batches. With fewer batches, we retain the active set and refine its cells one spatial level per batch. For $B \ge 3$ and dyadic scales $0 < r \le s \le 1/16$, let $L = L_B(s)$ and $b_{B,r} \asymp_d \log(eB/r)$ be as specified in Appendix C.3. Use L spatial levels with dyadic scales $1 = u_0 \ge u_1 \ge \cdots \ge u_L = s$.

Initialize $\mathcal{C}_0 = \{\mathcal{X}\}$. At level ℓ, let $\mathcal{Q}_\ell \subseteq \mathcal{P}_{u_\ell}$ consist of the cells contained in some $P \in \mathcal{C}_{\ell-1}$. Use two fixed sweeps over the level-ℓ cells, reserving $O_d(b_{B,r}u_\ell^{-2})$ pulls per cell in each sweep. For cells in $\mathcal{Q}_\ell$, sample their midpoints; other blocks use a representative of the preceding active set.

Let $\widehat{\mu}_{P,1}$ and $\widehat{\mu}_{P,2}$ be the rounded means from the two sweeps, with confidence radius $a_\ell = u_\ell/32$. For nonempty $\mathcal{Q}_\ell$, define

$$\lambda_\ell := \max_{P \in \mathcal{Q}_\ell} (\widehat{\mu}_{P,1} - a_\ell), \qquad \mathcal{C}_\ell := \{P \in \mathcal{Q}_\ell : \widehat{\mu}_{P,2} + a_\ell \ge \lambda_\ell - u_\ell\}. \tag{15}$$

If $\mathcal{Q}_\ell$ is empty, set $\mathcal{C}_\ell = \varnothing$ and use a fixed fallback action. Of the first-sweep estimates, retain only their largest lower confidence bound λ_ℓ. The second sweep computes the new indicator vector one entry at a time. Both sampling sequences depend only on $\mathcal{C}_{\ell-1}$, so the two sweeps form one batch. On the common confidence event, $\mathcal{C}_\ell$ is safe at scale u_ℓ.

At level ℓ, there are $O_d(u_\ell^{-d})$ blocks, each with $O_d(b_{B,r}u_\ell^{-2})$ pulls of gap $O_d(u_{\ell-1})$. Choose the intermediate scales so that $u_{\ell-1}u_\ell^{-d-2} \asymp_d s^{-\gamma_d(L)}$, where $\gamma_d(L) := (d+1)/[1-(d+2)^{-L}]$.

Lemma 5.3 (Hierarchical upper bound)**.** *For some $C_d > 0$ and all sufficiently large T, if*

$$W \ge C_d(s^{-d} + b_{B,r}), \qquad C_d L b_{B,r} r^{-d-2} \le T/2, \tag{16}$$

then a policy $\mathcal{A} \in \mathfrak{A}_{B,W}$ with fixed batch boundaries and at most $L + 2 \le B$ batches satisfies

$$\sup_{f \in \mathrm{Lip}_1(\mathcal{X}),\, \nu \in \mathfrak{D}(f)} \mathbb{E}_{\nu,\mathcal{A}}\big[\mathrm{Reg}_T(f)\big] \le C_d \left[L b_{B,r} s^{-\gamma_d(L)} + b_{B,r} s r^{-d-2} + Tr \right]. \tag{17}$$

When $L = B - 2$, balancing these costs gives exponent $\beta_{d,B}$ if the indicator vector fits in memory; otherwise, we increase s.

Completing the upper bound. Use a fixed action for $B = 1$ and refinement from the full domain for $B = 2$. For $B \ge 3$, Appendix D.1 selects a construction according to its batch requirement and chooses feasible spatial scales to prove (4). The construction need not change when the dominant regret term changes. For example, at $B = 3$, the hierarchy attains the joint term with logarithmic memory and the batch-depth term with sufficiently large memory, up to logarithmic factors.

5.4 Comparison with the finite-arm guarantee

We compare our upper bound with the guarantee obtained by applying Huang et al. (2026, Theorem 2) to a regular grid and optimizing its size under the same budgets. Under the conditions in Appendix D.4,

when the joint term dominates, the two bounds have polynomial orders

$$\text{Direct grid:} \quad T^{\alpha_d}\left(\frac{T^{d/(d+2)}}{BW}\right)^{1/d}, \qquad \text{Our bound:} \quad T^{\alpha_d}\left(\frac{T^{d/(d+2)}}{BW}\right)^{1/[d(d+3)]}.$$

Thus our bound is polynomially smaller when $T^{d/(d+2)}/(BW)$ grows polynomially with T.

6 Minimax regret with one bit of memory

For $W \gtrsim_d \log(eT)$, fixed batch boundaries attain minimax regret up to logarithmic factors. We now consider $W = 1$, the smallest nonzero memory budget, to determine the attainable regret and whether fixed boundaries remain sufficient. Let $\mathfrak{R}_T^{\mathrm{fix}}(B, W)$ denote minimax regret over policies with batch boundaries fixed in advance.

Theorem 6.1 (One-bit minimax regret)**.** *There exist constants $c_d, C_d > 0$, depending only on d, such that, for every $T \geq 2$ and all B,*

$$c_d T B^{-1/(d+2)} \leq \mathfrak{R}_T(B, 1) \leq C_d T \left(\frac{\log(eB)}{B}\right)^{1/(d+2)}. \tag{18}$$

In contrast, $\mathfrak{R}_T^{\mathrm{fix}}(B, 1) = \Theta_d(T)$ for every B.

For fixed boundaries, conditioning on the public seed leaves at most two possible actions at each round when $W = 1$; Lemma A.2 therefore gives linear regret.

At $B = T$, adaptive boundaries attain near-optimal regret $\widetilde{\Theta}_d(T^{\alpha_d})$, whereas fixed boundaries still incur linear regret. Even with adaptive boundaries, the three-term lower bound need not be tight below logarithmic memory. For $d = 1$ and $B \asymp T^{1/2}$, its value at $W = 1$ is of order $T^{2/3}$, whereas the one-bit minimax regret is $\widetilde{\Theta}(T^{5/6})$.

Adaptive boundaries. The policy chooses batch lengths so that the comparison statistics and the incumbent's index can be computed from the current round and batch indices. A one-round query starting at time t stores a bit M_{t+1}. The next batch pulls the incumbent for $1 + M_{t+1}$ rounds and resets the state, so the following batch starts at $t + 2 + M_{t+1}$. All these rounds count toward the horizon and regret; the starting times remain determined by the seed and boundary states. For dyadic $0 < \rho \leq 1/2$, Lemma E.2 gives a policy with $O_d(\rho^{-d-2}\log(e/\rho))$ exploration rounds and batches and expected regret $O_d(\rho^{-d-1}\log(e/\rho) + T\rho)$, provided exploration finishes by $T/2$. For sufficiently large B, choose $\rho \asymp_d (\log(eB)/B)^{1/(d+2)}$ with a large enough constant to use at most $B/2 \leq T/2$ exploration rounds and batches. A final exploitation batch gives $O_d(T\rho)$ regret; for bounded B, use a fixed action.

Proof sketch of the lower bound. Fix a one-bit policy. Consider a constant reference mean and $K \asymp_d \delta^{-d}$ disjoint radius-δ bumps of height δ. Let $P_0, P_1, \ldots, P_K$ denote the corresponding laws of the public seed and nonterminal boundary states. Conditional on a common boundary history and seed, the squared Hellinger distances between the reference and alternative laws of the next boundary state sum to $O(\delta^2)$ over the alternatives, uniformly in the batch length. Composing over boundaries gives $\sum_{v=1}^K H^2(P_0, P_v) \lesssim (B-1)\delta^2$. Hence $K^{-1}\sum_{v=1}^K \mathrm{TV}(P_0, P_v) \lesssim_d \sqrt{B\delta^{d+2}}$. Taking $\delta = c_d B^{-1/(d+2)}$ with sufficiently small c_d makes this average distance at most $1/4$. Since the action sequence is determined by the seed and boundary states, disjointness implies that some alternative receives $\Omega(T)$ expected pulls outside its bump, giving $\Omega_d(T\delta)$ regret. See Appendix E. □

7 Conclusion

Lipschitz geometry allows coarse regional information to guide fine action comparisons at low regret. The one-bit separation shows that batch timing is part of the memory model: adaptive boundaries can preserve information through the freely available round index.

Two questions remain open: the sharp regret dependence for $2 \leq W = o(\log T)$ under fixed and adaptive boundaries, and whether restricted Lipschitz classes with fewer near-optimal regions admit sharper tradeoffs.

Reproducibility Statement

This is a theoretical paper. The appendices provide the formal policy specification (Appendix A) and complete proofs of the main and one-bit results (Appendices B–E). The upper-bound proofs specify finite-precision implementations and verify their memory use, batch counts, and horizon constraints.

AI Use Statement

We used generative AI tools to improve the clarity and presentation of the manuscript and to assist with consistency checks. The authors take full responsibility for the final manuscript and supplementary material.

References

Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Proceedings of the 30th Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 39–75. PMLR, 2017. URL https://proceedings.mlr.press/v65/agarwal17c.html.

Arpit Agarwal, Sanjeev Khanna, and Prathamesh Patil. A sharp memory-regret trade-off for multi-pass streaming bandits. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1423–1462. PMLR, 2022. URL https://proceedings.mlr.press/v178/agarwal22a.html.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, 1995. doi: 10.1137/S0363012992237273.

Sepehr Assadi and Chen Wang. Exploration with limited memory: Streaming algorithms for coin tossing, noisy comparisons, and multi-armed bandits. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1237–1250, 2020. doi: 10.1145/3357713.3384341.

Sepehr Assadi and Chen Wang. The best arm evades: Near-optimal multi-pass streaming lower bounds for pure exploration in multi-armed bandits. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 311–358. PMLR, 2024. URL https://proceedings.mlr.press/v247/assadi24a.html.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Conference on Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pp. 454–468. Springer, 2007. doi: 10.1007/978-3-540-72927-3_33.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(46):1655–1695, 2011. URL https://www.jmlr.org/papers/v12/bubeck11a.html.

Pierre-Arnaud Coquelin and Rémi Munos. Bandit algorithms for tree search. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 67–74, 2007. URL https://arxiv.org/abs/cs/0703062.

Thomas M. Cover. A note on the two-armed bandit problem with finite memory. *Information and Control*, 12(5):371–377, 1968. doi: 10.1016/S0019-9958(68)90382-3.

Thomas M. Cover and Martin E. Hellman. The two-armed-bandit problem with time-invariant finite memory. *IEEE Transactions on Information Theory*, 16(2):185–195, 1970. doi: 10.1109/TIT.1970.1054427.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7340–7348, 2021. doi: 10.1609/aaai.v35i8.16901.

Yasong Feng, Zengfeng Huang, and Tianyu Wang. Lipschitz bandits with batched feedback. *IEEE Transactions on Information Theory*, 70(3):2154–2176, 2024. doi: 10.1109/TIT.2023.3312308.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/20f07591c6fcb220ffe637cda29bb3f6-Abstract.html.

Osama Hanna, Lin Yang, and Christina Fragouli. Efficient batched algorithm for contextual linear bandits with large action space via soft elimination. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56772–56783, 2023.

Ruiyuan Huang, Zicheng Lyu, Xiaoyi Zhu, and Zengfeng Huang. Few batches or little memory, but not both: Simultaneous space and adaptivity constraints in stochastic bandits. *arXiv preprint arXiv:2603.13742*, 2026. URL https://arxiv.org/abs/2603.13742v2. Version 2, March 30, 2026.

Tianyuan Jin, Keke Huang, Jing Tang, and Xiaokui Xiao. Optimal streaming algorithms for multi-armed bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5045–5054. PMLR, 2021a. URL https://proceedings.mlr.press/v139/jin21a.html.

Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanquan Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5065–5073. PMLR, 2021b. URL https://proceedings.mlr.press/v139/jin21c.html.

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

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008. doi: 10.1145/1374376.1374475.

Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, volume 17, pp. 697–704, 2004.

Ivan Lau, Daniel McMorrow, Kevin Jamieson, and Jonathan Scarlett. Batched stochastic linear bandits with 1-Bit communication constraints. *arXiv preprint arXiv:2605.30976*, 2026. URL https://arxiv.org/abs/2605.30976.

Shaoang Li, Lan Zhang, Junhao Wang, and Xiang-Yang Li. Tight memory-regret lower bounds for streaming bandits. *arXiv preprint arXiv:2306.07903*, 2023. URL https://arxiv.org/abs/2306.07903.

David Liau, Eric Price, Zhao Song, and Ger Yang. Stochastic multi-armed bandits in constant space. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 386–394. PMLR, 2018. URL https://proceedings.mlr.press/v84/liau18a.html.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 975–999. PMLR, 2014. URL https://proceedings.mlr.press/v35/magureanu14.html.

Arnab Maiti, Vishakha Patil, and Arindam Khan. Multi-armed bandits with bounded arm-memory: Near-optimal guarantees for best-arm identification and regret minimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19553–19565, 2021.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004. URL https://www.jmlr.org/papers/v5/mannor04b.html.

Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems*, volume 24, pp. 783–791, 2011.

Binghui Peng and Aviad Rubinstein. Near optimal memory-regret tradeoff for online learning. In *Proceedings of the 64th IEEE Annual Symposium on Foundations of Computer Science*, pp. 1171–1194. IEEE, 2023. doi: 10.1109/FOCS57990.2023.00069.

Binghui Peng and Fred Zhang. Online prediction in sub-linear space. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1611–1634. SIAM, 2023. doi: 10.1137/1.9781611977554.ch60.

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

Chara Podimata and Alex Slivkins. Adaptive discretization for adversarial Lipschitz bandits. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3788–3805. PMLR, 2021. URL https://proceedings.mlr.press/v134/podimata21a.html.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Regret minimisation in multi-armed bandits using bounded arm memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10085–10092, 2020. doi: 10.1609/aaai.v34i06.6566.

Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 74–87, 2021. doi: 10.1145/3406325.3451004.

Ayush Sawarni, Nirjhar Das, Siddharth Barman, and Gaurav Sinha. Generalized linear bandits with limited adaptivity. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8329–8369, 2024.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Proceedings of the 30th Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1794–1834. PMLR, 2017. URL https://proceedings.mlr.press/v65/simchowitz17a.html.

Vaidehi Srinivas, David P. Woodruff, Ziyu Xu, and Samson Zhou. Memory bounds for the experts problem. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1158–1171. ACM, 2022. doi: 10.1145/3519935.3520069.

Chen Wang. Tight regret bounds for single-pass streaming multi-armed bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35525–35547. PMLR, 2023. URL https://proceedings.mlr.press/v202/wang23a.html.

Xiao Xu and Qing Zhao. Memory-constrained no-regret learning in adversarial multi-armed bandits. *IEEE Transactions on Signal Processing*, 69:2371–2382, 2021. doi: 10.1109/TSP.2021.3070201.

Xiaoyi Zhu and Zengfeng Huang. Lipschitz bandits in optimal space. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/06b1652d342f3677b412951f79740475-Abstract-Conference.html.

A POLICY MAPS AND BOUNDARY-STATE RECONSTRUCTION

Policy maps. We complete the policy specification in Section 3.1. Let $(\Omega_0, \mathcal{G}_0)$ be the standard Borel space of the public seed ω. For $b \in \{0, \dots, B-1\}$ and $u \in [T]$, the policy maps are Borel on the following full domains:

$$\begin{aligned}
\iota &: \Omega_0 \longrightarrow \{0,1\}^W, \\
\psi_b &: \Omega_0 \times \{0,\dots,T\} \times \{0,1\}^W \longrightarrow \{0,\dots,T\}, \\
\phi_{b,u} &: \Omega_0 \times \{0,\dots,T\} \times \{0,1\}^W \longrightarrow \mathcal{X}, \\
\mathcal{U}_{b,u} &: \Omega_0 \times \{0,1\}^W \times \mathcal{X} \times [0,1] \longrightarrow \{0,1\}^W.
\end{aligned}$$

For every (w, t, m), require

$$\psi_b(w,t,m) \in \begin{cases} \{t+1,\dots,T\}, & t < T, \\ \{T\}, & t = T, \end{cases} \qquad \psi_{B-1}(w,t,m) = T.$$

Thus batches are nonempty until the horizon is reached, and the last permitted batch ends at T. The sequential finite-state subclass has $B = T$ and $\tau_b = b$.

Boundary states. Let $\widehat{B} := \min\{b \in [B] : \tau_b = T\}$ be the number of nonempty batches. For $0 \le b \le B$, set $\bar{\tau}_b := \tau_{\min\{b,\widehat{B}\}}$, and define

$$\bar{M}_b := \begin{cases} M_{\tau_b}, & b < \widehat{B}, \\ 0^W, & b \ge \widehat{B}, \end{cases} \qquad b = 1, \dots, B-1.$$

Write

$$\mathbf{M} := (\bar{M}_1, \dots, \bar{M}_{B-1}), \qquad \mathsf{Tr} := (\bar{\tau}_{1:B-1}, A_{1:T}).$$

The tuple $\mathbf{M}$ contains the states at nonterminal batch boundaries, followed by fixed padding. It is used only in the analysis and is not additional storage available to the learner. The transcript Tr records the batch boundaries and actions.

Lemma A.1 (Boundary reconstruction)**.** *For every policy, there is a policy-dependent Borel map such that*

$$\mathsf{Tr} = \mathsf{Rec}_{\mathcal{A}}(\omega, \mathbf{M}) \quad \textit{a.s.}$$

Conditional on the public seed, the transcript takes at most $2^{(B-1)W}$ values. More generally, the boundaries and actions through batch j are determined by $\omega, \bar{M}_1, \dots, \bar{M}_{j-1}$ and take at most $2^{(j-1)W}$ values.

Proof. For an input $(w, m_1, \dots, m_{B-1})$, initialize $t_0 = 0$ and $m_0 = \iota(w)$, and recursively set

$$t_{b+1} = \psi_b(w, t_b, m_b), \qquad a_u = \phi_{b,u}(w, t_b, m_b), \quad t_b < u \le t_{b+1}, \quad b = 0, \dots, B-1.$$

The boundary conventions ensure $t_B = T$. Finite composition of the policy maps defines the Borel map $\mathsf{Rec}_{\mathcal{A}}(w, m_{1:B-1}) = (t_{1:B-1}, a_{1:T})$.

For the policy's padded boundary states, induction using (1)–(2) gives $t_b = \bar{\tau}_b$ and $a_u = A_u$. The recursion through batch j uses only the first $j-1$ boundary states, which have at most $2^{(j-1)W}$ possible values. □

Information bound. Let V be independent of ω, and let U be generated from Tr using fresh randomness independent of $(V, \omega, \mathbf{M}, \mathsf{Tr})$. Conditional data processing gives

$$I(V; U \mid \omega) \le I(V; \mathsf{Tr} \mid \omega) \le H(\mathbf{M} \mid \omega) \le (B-1)W. \tag{19}$$

Entropy and mutual information are measured in bits (Cover & Thomas, 2006, Chapters 2–3). For $B = 1$, the tuple $\mathbf{M}$ is empty and the bound is zero.

Finite action ranges.

Lemma A.2 (Lower bound from finite action ranges)**.** *Let $S \geq 1$ be an integer. Suppose a policy $\mathcal{A}$ admits measurable sets $\mathcal{C}_t(\omega) \subseteq \mathcal{X}$, determined by its public seed, such that $|\mathcal{C}_t(\omega)| \leq S$ and $A_t \in \mathcal{C}_t(\omega)$ almost surely for every $t \in [T]$ and every admissible instance. Then*

$$\sup_{f \in \mathrm{Lip}_1(\mathcal{X}),\, \nu \in \mathfrak{D}(f)} \mathbb{E}_{\nu,\mathcal{A}}\big[\mathrm{Reg}_T(f)\big] \geq c_d T S^{-1/d}.$$

Proof. Choose $N = 2S$ centers $z_1, \dots, z_N \in [1/4, 3/4]^d$ separated by $4r_0$, where $r_0 = c_d S^{-1/d}$ for a sufficiently small constant $c_d > 0$. Consider the Bernoulli instances with means $g_v(x) = 1/4 + (r_0 - \|x - z_v\|_\infty)_+$. These means belong to $\mathrm{Lip}_1(\mathcal{X})$, and their disjoint regions $G_v = \{x \in \mathcal{X} : \|x - z_v\|_\infty \leq r_0/2\}$ satisfy $\Delta_{g_v}(x) \geq (r_0/2)\mathbf{1}\{x \notin G_v\}$.

Write $\mathbb{P}_v^\circ, \mathbb{E}_v^\circ$ for the corresponding interaction laws and expectations. For each t and almost every seed w, at most S regions intersect $\mathcal{C}_t(w)$, so

$$\frac{1}{N}\sum_{v=1}^{N} \mathbb{P}_v^\circ(A_t \in G_v \mid \omega = w) \leq \frac{1}{N}\sum_{v=1}^{N} \mathbf{1}\{\mathcal{C}_t(w) \cap G_v \neq \varnothing\} \leq \frac{1}{2}.$$

Integrating over the common seed law and summing over t gives

$$\frac{1}{N}\sum_{v=1}^{N} \mathbb{E}_v^\circ \mathrm{Reg}_T(g_v) \geq \frac{r_0 T}{4} \geq c_d T S^{-1/d},$$

which lower-bounds the supremum. □

By Lemma A.1, for each seed the reconstruction map provides an instance-independent set of at most $2^{(B-1)W}$ action transcripts. Projecting onto the time-t action and applying Lemma A.2 gives

$$\mathfrak{R}_T(B, W) \geq c_d T 2^{-(B-1)W/d}. \tag{20}$$

B Lower-bound proofs

We use the model and notation of Section 3, together with the entropy conventions of Appendix A. Relative entropy is measured in nats. Write $B_\infty(x, a) := \{y : \|x - y\|_\infty \leq a\}$. For measurable $G \subseteq \mathcal{X}$, write $N_G(t) := \sum_{u=1}^{t} \mathbf{1}\{A_u \in G\}$.

B.1 Hard instances and local testing

Regional packing. We realize the packing used in Definition 4.1 as follows. Choose an even number $2m \asymp_d s^{-d}$ of centers from the first of the following grids, pair them as $(x_{j,0}, x_{j,1})_{j \in [m]}$, and use the second grid for the neighborhood centers in each cell:

$$\mathcal{U}_s = \tfrac{1}{4}\mathbf{1}_d + 8s\{0, \dots, \lfloor (16s)^{-1} \rfloor\}^d,$$
$$\mathcal{Z}(x) = x - \tfrac{s}{4}\mathbf{1}_d + 4r\{0, \dots, \lfloor s/(8r) \rfloor\}^d.$$

Here $\mathbf{1}_d$ is the all-ones vector. The coarse centers lie in $[1/4, 3/4]^d$ and are separated by at least $8s$. Each $\mathcal{Z}(x)$ contains the same number $q \asymp_d (s/r)^d$ of points, lies in $B_\infty(x, s/4)$, and has separation at least $4r$. Enumerate these points as $z_{j,a,k}$. The cells lie in $\mathcal{X}$, and all neighborhoods are pairwise disjoint and contained in their parent cells.

Geometric properties. The base function and each added bump have Lipschitz constant at most $1/2$, so all means belong to $\mathrm{Lip}_1(\mathcal{X})$ and lie in $[1/4, 3/4]$ for small enough s_d. Every selected cell is an optimal plateau, with $f_v^\star = 1/4 + s/4$. The cells have mutual distance at least $6s$, so every unselected cell has mean $1/4$. The bump has height $r/2$ and a unique peak in a selected cell; thus $(f_{v,j,k}^+)^\star = 1/4 + s/4 + r/2$, attained uniquely at $z_{j,v_j,k}$, and

$$\Delta_{f_v}(x) = s/4, \qquad x \in P_{j,1-v_j}, \tag{21}$$
$$\Delta_{f_{v,j,k}^+}(x) \geq r/2, \qquad x \notin G_{j,v_j,k}. \tag{22}$$

Moreover, $0 \le f^+_{v,j,k} - f_v \le r/2$, and

$$f^+_{v,j,k}(x) = f_v(x) \qquad (x \notin G_{j,v_j,k}). \tag{23}$$

For a fixed policy, write $\mathbb{P}_v, \mathbb{E}_v$ and $\mathbb{P}^+_{v,j,k}, \mathbb{E}^+_{v,j,k}$ for the interaction laws and expectations under the base and local instances, and abbreviate $N_{j,a,k}(t) := N_{G_{j,a,k}}(t)$.

Change of measure. We use a change-of-measure argument stopped after a prescribed number of visits to the perturbed region (Mannor & Tsitsiklis, 2004; Kaufmann et al., 2016); related formulations appear in Simchowitz et al. (2017); Huang et al. (2026).

Lemma B.1 (Localized change-of-measure bound)**.** *Fix a policy and two instances $\nu_i \in \mathfrak{D}(f_i)$, $i = 0, 1$, with induced laws $\mathbb{P}_i$ and expectations $\mathbb{E}_i$. Let $G \subseteq \mathcal{X}$ be measurable and suppose, for some $\Delta > 0$ and $\kappa \ge 0$, that*

$$\nu_{1,x} = \nu_{0,x} \quad (x \notin G), \qquad \sup_{x \in G} D(\nu_{1,x} \| \nu_{0,x}) \le \kappa, \qquad \inf_{x \notin G} \Delta_{f_1}(x) \ge \Delta.$$

For every integer $1 \le n < T$,

$$\mathbb{P}_0(N_G(T) \ge n) \ge 1 - \frac{\mathbb{E}_1 \mathrm{Reg}_T(f_1)}{\Delta(T-n)} - \sqrt{\frac{n\kappa}{2}}. \tag{24}$$

Proof. For $t = 0, \dots, T$, let $\mathcal{F}_t := \sigma(\omega, (A_u, Y_u)_{1 \le u \le t})$ and set $\tau_n := \inf\{t \in [T] : N_G(t) \ge n\} \wedge T$, with $\inf \varnothing = \infty$. The policy maps in Section 3.1 make A_t $\mathcal{F}_{t-1}$-measurable, so the stopped chain rule gives

$$D\Big(\mathbb{P}_1|_{\mathcal{F}_{\tau_n}} \,\Big\|\, \mathbb{P}_0|_{\mathcal{F}_{\tau_n}}\Big) = \mathbb{E}_1 \sum_{t=1}^{\tau_n} D(\nu_{1,A_t} \| \nu_{0,A_t}) \le \kappa \mathbb{E}_1 N_G(\tau_n) \le n\kappa. \tag{25}$$

The event $E_n := \{N_G(T) \ge n\}$ equals $\{N_G(\tau_n) \ge n\}$ and therefore belongs to $\mathcal{F}_{\tau_n}$. Pinsker's inequality yields $\mathbb{P}_0(E_n) \ge \mathbb{P}_1(E_n) - \sqrt{n\kappa/2}$.

On $E_n^{\complement}$, $\mathrm{Reg}_T(f_1) \ge \Delta(T - N_G(T)) \ge \Delta(T-n)$. Hence $\mathbb{P}_1(E_n^{\complement}) \le \mathbb{E}_1 \mathrm{Reg}_T(f_1)/[\Delta(T-n)]$. Combining these inequalities proves the claim. □

Corollary B.2 (Testing selected neighborhoods)**.** *For sufficiently small s_d, there exist constants $c_0, c_d, C_d > 0$ such that, whenever $0 < s \le s_d$, $0 < r \le s/16$, and $Tr^2 \ge C_d$, the threshold $n = \lfloor c_0 r^{-2} \rfloor$ satisfies $2 \le n \le T/4$ and $n \asymp r^{-2}$. For every fixed policy and all v, j, k, if $\mathbb{E}^+_{v,j,k} \mathrm{Reg}_T(f^+_{v,j,k}) \le c_d T r$, then*

$$\mathbb{P}_v\big(N_{j,v_j,k}(T) \ge n\big) \ge \frac{5}{8}.$$

Proof. For $u, v \in [1/4, 3/4]$, $D(\mathrm{Ber}(u) \| \mathrm{Ber}(v)) \le 16(u-v)^2/3$. The local means differ by at most $r/2$ and agree outside the selected neighborhood, whose complement has alternative gap at least $r/2$. Lemma B.1, setting $\kappa = 4r^2/3$ and $\Delta = r/2$ gives:

$$\mathbb{P}_v(N_{j,v_j,k}(T) \ge n) \ge 1 - \frac{2\mathbb{E}^+_{v,j,k} \mathrm{Reg}_T(f^+_{v,j,k})}{r(T-n)} - \sqrt{2nr^2/3} \ge \frac{5}{8}. \tag{26}$$

Choose c_0, c_d sufficiently small for the last inequality; decreasing s_d and increasing C_d ensures the stated bounds on n. □

B.2 Sequential and joint lower bounds

Sequential minimax lower bound. Fix $s = s_d/2$ and $r = a_d T^{-1/(d+2)}$, where $a_d > 0$ is a sufficiently small constant. For sufficiently large T, these scales satisfy the assumptions of Corollary B.2. If a policy has expected regret below $c_d T r$ on every base instance and local alternative, that corollary

forces $\Omega(r^{-2})$ expected visits to each selected neighborhood. Since these neighborhoods are disjoint, under any base instance f_v,

$$T \geq \sum_{j=1}^{m} \sum_{k=1}^{q} \mathbb{E}_v N_{j,v_j,k}(T) \gtrsim_d mqr^{-2} \asymp_d r^{-d-2} = a_d^{-(d+2)} T.$$

Choosing a_d sufficiently small gives a contradiction. Thus, for all sufficiently large T and every admissible B, W,

$$\mathfrak{R}_T(B, W) \geq c_d T^{\alpha_d}. \tag{27}$$

Proof of Lemma 4.2. Fix a policy and use the notation and estimator $\widehat{V}$ from the proof sketch in Section 4.2, with n specified in Corollary B.2. Suppose

$$\mathcal{R} < c_d \min\{Tr, sr^{-d-2}\} \tag{28}$$

for sufficiently small c_d. The tie-breaking bits are independent of $(V, \omega, A_{1:T})$. Corollary B.2 gives $\mathbb{E}_v \rho_{j,v_j} \geq 5/8$ for every v, j. Since $0 \leq \rho_{j,v_j} \leq 1$, this implies $\mathbb{P}_v(\rho_{j,v_j} \geq 1/4) \geq 1/2$. On the other hand, marking an unselected cell requires at least $qn/4$ pulls of gap $s/4$, and these pulls are disjoint across cells. Thus, pathwise,

$$\mathrm{Reg}_T(f_v) \geq \frac{sqn}{16} \sum_{j=1}^{m} \mathbf{1}\{\rho_{j,1-v_j} \geq 1/4\}.$$

Averaging and using $msqn \asymp_d sr^{-d-2}$ yields

$$\frac{1}{m} \sum_j \overline{\mathbb{P}}(\rho_{j,V_j} \geq 1/4) \geq \frac{1}{2}, \qquad \overline{\mathbb{E}} \sum_j \mathbf{1}\{\rho_{j,1-V_j} \geq 1/4\} \leq \frac{16\mathcal{R}}{sqn} \leq \frac{m}{8}. \tag{29}$$

The independent tie-breaking coins give

$$\overline{\mathbb{P}}(\widehat{V}_j \neq V_j) = \tfrac{1}{2}\big[1 - \overline{\mathbb{P}}(\rho_{j,V_j} \geq 1/4) + \overline{\mathbb{P}}(\rho_{j,1-V_j} \geq 1/4)\big],$$

so the average coordinate error is at most $5/16$. Conditioning reduces entropy, and the binary error bound and concavity of h_2 gives

$$\begin{aligned} H(V \mid \widehat{V}, \omega) &\leq \sum_j H(V_j \mid \widehat{V}_j) \\ &\leq \sum_j h_2\Big(\overline{\mathbb{P}}(\widehat{V}_j \neq V_j)\Big) \leq m h_2(5/16). \end{aligned}$$

Since V is uniform and independent of ω, $H(V \mid \omega) = m$. The estimator $\widehat{V}$ depends only on the action sequence and independent randomness, so conditional data processing gives

$$[1 - h_2(5/16)]m \leq I(V; \widehat{V} \mid \omega) \leq I(V; A_{1:T} \mid \omega). \tag{30}$$

Since $m \asymp_d s^{-d}$, this proves (11).

Under (9), (19) gives $I(V; A_{1:T} \mid \omega) \leq (B-1)W \leq c_d s^{-d}$, contradicting (11) for sufficiently small c_d. Thus every policy has $\mathcal{R} \geq c_d \min\{Tr, sr^{-d-2}\}$, which proves (10). □

Completion of the sequential and joint lower bounds. The sequential lower bound (27) holds for all B, W. For bounded $(B-1)W$, (20) gives linear regret and dominates the joint term. In the intermediate range $K_d < (B-1)W < \eta_d T^{d/(d+2)}$, choose $s \asymp_d (1 + (B-1)W)^{-1/d}$ and $r \asymp_d (s/T)^{1/(d+3)}$, allowing dyadic rounding. A small coefficient in s and large K_d ensure $s \leq s_d$ and (9). Choosing η_d small makes Ts^{d+2} sufficiently large, giving $r/s \asymp_d (Ts^{d+2})^{-1/(d+3)} \leq 1/16$. Also $Tr^2 \gtrsim_d T^{d/(d+2)}$, so the testing assumption holds for sufficiently large T. Applying (10) gives

$$\mathfrak{R}_T(B, W) \geq c_d \min\{Tr, sr^{-d-2}\} \asymp_d T^{\frac{d+2}{d+3}} (1 + (B-1)W)^{-\frac{1}{d(d+3)}}.$$

For $(B-1)W \geq \eta_d T^{d/(d+2)}$, the last expression is $O_d(T^{\alpha_d})$ and is covered by the sequential lower bound. Thus, for sufficiently large T and every admissible B, W,

$$\mathfrak{R}_T(B, W) \geq c_d \left[T^{\alpha_d} \vee T^{\frac{d+2}{d+3}} (1 + (B-1)W)^{-\frac{1}{d(d+3)}}\right].$$

□

B.3 LOWER BOUND FROM BATCHING

Lemma B.3 (Bernoulli transfer of the adaptive-grid batch lower bound)**.** *Fix $d \geq 1$. There are constants $c_d, T_d > 0$ such that, for $T \geq T_d$ and $2 \leq B \leq T$,*

$$\mathfrak{R}_T^{\mathrm{bat}}(B) \geq c_d \frac{T^{\beta_{d,B}}}{B^2}.$$

The same bound holds for $\mathfrak{R}_T(B, W)$ for every integer $W \geq 0$.

Proof. Fix $\mathcal{A} \in \mathfrak{A}_{B,\infty}$. Feed the binary observations $\widetilde{Y}_t = \mathbf{1}\{Y_t^{\mathrm{G}} \geq 0\}$ from the Gaussian experiment to $\mathcal{A}$. We obtain a policy $\mathcal{A}^\sharp$ with exactly B nonempty batches, as required by Feng et al. (2024, Section 1.2), by subdividing the original batches without changing their actions.

If new batch b starts at t_{b-1} inside an original batch ending at τ, set

$$t_b = \min\{\tau, T - (B - b)\}.$$

At each t_b, process the newly revealed binary rewards in order. Invoke the original decision rule if $t_b = \tau$; otherwise retain the original boundary and remaining actions. This preserves the simulated states at original boundaries and hence the entire action sequence.

Each t_b is chosen at t_{b-1} and satisfies $t_{b-1} < t_b \leq T - (B - b)$. A strict subdivision leaves exactly $B - b$ rounds, which form singleton batches. If no strict subdivision occurs, each new batch completes an original batch. Since the original policy uses at most B batches, the construction yields exactly B nonempty batches.

Applying Feng et al. (2024, Theorem 8) to $\mathcal{A}^\sharp$ on the full-dimensional Gaussian family $\mathcal{F}_{d,T,B}$ gives a mean μ such that

$$\mathbb{E}_\mu^{\mathcal{A}^\sharp} \sum_{t=1}^{T} (\mu^\star - \mu(A_t)) \geq \frac{T^{\beta_{d,B}}}{256 B^2}. \tag{31}$$

The source radii satisfy $0 < r_j \leq r_1 = 1/B \leq 1/2$. Its base means are maxima of the floor $r_1/2$ and cones with peak heights at most $r_1/2 + r_j/16 + r_B/16$ (Feng et al., 2024, equations (25)–(26)); the alternatives in its Lemma 6 increase one peak by at most $3r_j/16$. Consequently, uniformly over T, B and all source means,

$$0 \leq \mu(x) \leq \frac{r_1}{2} + \max_j \left(\frac{r_j}{16} + \frac{r_B}{16} + \frac{3r_j}{16} \right) \leq \frac{13 r_1}{16} \leq \frac{13}{32}.$$

Let Φ, ϕ be the standard normal distribution function and density, and put $f_\mu := \Phi \circ \mu$ and $c_\Phi := \min_{z \in [0, 13/32]} \phi(z) > 0$. Since $\|\phi\|_\infty < 1$, $f_\mu \in \mathrm{Lip}_1(\mathcal{X})$ and

$$f_\mu^\star - f_\mu(x) \geq c_\Phi(\mu^\star - \mu(x)). \tag{32}$$

Under the source observation $Y_t^{\mathrm{G}} \sim \mathcal{N}(\mu(A_t), 1)$, the transformed reward has the Bernoulli kernel with mean $f_\mu(A_t)$. Thus the simulated original states, boundaries, and actions have the same joint law as under $\mathcal{A}$ on f_μ; only the additional subdivision boundaries differ. Combining (31) and (32) gives

$$\mathbb{E}_{f_\mu}^{\mathcal{A}} \mathrm{Reg}_T(f_\mu) \geq c_\Phi \mathbb{E}_\mu^{\mathcal{A}^\sharp} \sum_{t=1}^{T} (\mu^\star - \mu(A_t)) \geq \frac{c_\Phi}{256 B^2} T^{\beta_{d,B}}.$$

Take the supremum over admissible instances and the infimum over $\mathcal{A}$. The finite-memory claim follows from $\mathfrak{A}_{B,W} \subseteq \mathfrak{A}_{B,\infty}$. □

Proof of the lower bound in Theorem 3.1. For $2 \leq B \leq T$, the sequential and joint bounds established in Appendix B.2, together with Lemma B.3, give the three terms in (3). Each bound holds for every $W \geq 0$ in the same policy class $\mathfrak{A}_{B,W}$. Taking the minimum of their constants therefore yields (3).

For $B = 1$, (20) gives $\mathfrak{R}_T(1, W) \geq c_d T$ for every $W \geq 0$. Since $\beta_{d,1} = 1$, the displayed maximum becomes

$$T^{\alpha_d} \vee T \vee T^{\frac{d+2}{d+3}} = T,$$

so the same conclusion holds.

Finally, choose T_d as the maximum of the horizon thresholds required by the preceding bounds. Both c_d and T_d depend only on d, completing the proof. □

C UPPER-BOUND CONSTRUCTIONS

We specify and verify the constructions of Section 5, using its notation for safe active sets and refinement. Appendix D selects their scales and completes the upper-bound proof.

C.1 REFINEMENT WITHIN ACTIVE CELLS

Throughout this appendix, use the fallback action $x_\circ$ from Section 3.2; ties in estimated means or confidence bounds retain the first action in the prescribed order. Write $\mathrm{bl}(z) := 2 + \lceil \log_2(1 \vee z) \rceil$ for $z > 0$, and let $\lceil z \rceil_2$ be the smallest dyadic number at least $z \in (0, 1]$. Thus $\mathrm{bl}(z) \asymp \log(ez)$ for $z \geq 1$ and $z \leq \lceil z \rceil_2 < 2z$. Geometry, traversal orders, and numerical schedules are fixed from the problem parameters and public seed, and regenerated without stored tables. Retained addresses and statistics are charged to the state budget.

Dyadic grids. Partition each coordinate into half-open intervals of length 2^{-j}, including the right endpoint in the last interval, and use midpoint representatives. The level-j' ancestor of a level-j cell, $j' \leq j$, has address

$$\left(j', \left\lfloor k_1/2^{j-j'} \right\rfloor, \ldots, \left\lfloor k_d/2^{j-j'} \right\rfloor\right). \tag{33}$$

To realize the partition and nets of Section 5.1, take $j_s = \lceil \log_2(1/s) \rceil$ and $j_r = \lceil \log_2(1/r) \rceil$. The level-$j_s$ cells form $\mathcal{P}_s$, with diameter at most s and $|\mathcal{P}_s| \leq (2/s)^d$. Their level-$j_r$ descendant midpoints form r-nets $\mathcal{N}_r(P)$ of size at most $(2s/r)^d$ per parent. A level and d coordinate indices give each cell an $O_d(\log(e/r))$-bit address and allow lexicographic traversal without a stored geometric table.

Rounded means. For a prescribed block length $n \geq 1$ and mesh $0 < \varepsilon \leq 1$, store $S_t = \sum_{i \leq t} \lfloor Y_i/\varepsilon \rfloor$ and return $\widehat{\mu} = \varepsilon S_n/n$. Since $0 \leq S_t \leq n\lfloor 1/\varepsilon \rfloor$, the accumulator uses $\lceil \log_2(n\lfloor 1/\varepsilon \rfloor + 1) \rceil$ bits, and averaging the rounding errors gives

$$\left| \widehat{\mu} - \frac{1}{n} \sum_{i=1}^{n} Y_i \right| < \varepsilon. \tag{34}$$

If the block samples an action x measurable at its starting history $\mathcal{H}$, the fresh bounded rewards have conditional mean $f(x)$. Thus Hoeffding's inequality gives, for $a > \varepsilon$, almost surely,

$$\mathbb{P}\big(|\widehat{\mu} - f(x)| > a \mid \mathcal{H}\big) \leq 2\exp\{-2n(a - \varepsilon)^2\}. \tag{35}$$

Confidence screening. Both constructions use the following two inequalities. For $\eta, a, \zeta \geq 0$, suppose $\mu_\star$ bounds the candidate means, $\mu_\star - \eta \leq \lambda \leq \mu_\star$, and $f(x) \leq U(x) \leq f(x) + 2a$. The rule $U(x) \geq \lambda - \zeta$ retains every candidate with $\mu_\star - f(x) \leq \zeta$, since $U(x) \geq f(x) \geq \mu_\star - \zeta \geq \lambda - \zeta$. Conversely, every retained candidate satisfies

$$\mu_\star - f(x) \leq \mu_\star - U(x) + 2a \leq \eta + \zeta + 2a.$$

Refinement across groups. For $\mathcal{P}_s$, fix pairwise disjoint, nonempty groups $\mathcal{I}_1, \ldots, \mathcal{I}_{J_0} \subseteq [K]$. Before refining group j, let t_j be the current boundary and $\mathcal{H}_j = \sigma(\omega, (A_t, Y_t)_{t \leq t_j})$. At t_j, the indicator vector $Z^{(j)}$ is stored in the mutable state, and the filler $\bar{x}_j \in \{x_\circ, x_1, \ldots, x_K\}$ is determined by that state, the public seed, and the round and batch indices. These requirements hold on every sample path. Let $\mathcal{G}_j$ be an $\mathcal{H}_j$-measurable event. The selected cells are $\mathcal{C}_j = \{P_k : k \in \mathcal{I}_j,\ Z_k^{(j)} = 1\}$. On $\mathcal{G}_j$, assume

$$\Delta_f(x) \leq \kappa_d s \qquad \text{for every } x \in \{\bar{x}_j\} \cup \bigcup_{P \in \mathcal{C}_j} P.$$

On $\mathcal{G} := \bigcap_j \mathcal{G}_j$, the collection $\bigcup_j \mathcal{C}_j$ contains a maximizer's cell.

Construction phases may occur between refinement groups, provided they preserve the fine-action record. The auxiliary memory bound below includes the filler index and refinement workspace; any other state retained for later construction is counted separately.

Lemma C.1 (Finite-memory refinement). *Suppose $C_d r^{-d-2}\ell_r \le T/2$ and all construction and refinement batches finish by round T on every sample path. Under the assumptions above, refinement uses one batch per group, at most $C_d r^{-d-2}\ell_r$ pulls, and at most $C_d\ell_r$ auxiliary bits beyond the stored indicator vector after every observation. The resulting policy $\mathcal{A}$ satisfies*

$$\mathbb{E}_{\nu,\mathcal{A}}\big[\mathrm{Reg}_{\mathrm{ref+exp}}(f)\mathbf{1}_{\mathcal{G}}\big] \le C_d\big[sr^{-d-2}\ell_r + Tr\big]. \tag{36}$$

Proof. Sampling and memory. Set $n_r = \lceil A_d^{\mathrm{ref}} r^{-2}\ell_r\rceil$ and $\varepsilon = r/512$. Reserve one n_r-pull block for each $x \in \mathcal{N}_r(P_k)$, $k \in \mathcal{I}_j$. Sample x if $Z_k^{(j)} = 1$ and $\bar{x}_j$ otherwise. The indicator vector and filler index determine all actions at the batch start.

Process the active-block rounded means one at a time, retaining only the largest mean and its action. Discard each accumulator after comparison and preserve this record across groups and intervening construction phases. Finally, exploit the recorded action $\widehat{x}$, or $x_\circ$ if the record is empty. The number of blocks is

$$N_{\mathrm{block}} := \sum_j \sum_{k\in\mathcal{I}_j} |\mathcal{N}_r(P_k)| \le C_d r^{-d}, \qquad N_{\mathrm{block}} n_r \le C_d r^{-d-2}\ell_r.$$

The filler index uses at most $\lceil \log_2(K+1)\rceil = O_d(\ell_r)$ bits. Together with the grid addresses, control indices, one accumulator, and fine-action record, the refinement workspace uses $O_d(\ell_r)$ auxiliary bits.

Accuracy and regret. Let $\mathcal{E}$ be the event that every active-block estimate has error at most $r/32$. Conditional on its starting history, each active block satisfies (35). Taking expectations and a union bound gives

$$\mathbb{P}(\mathcal{E}^{\complement}) \le 2N_{\mathrm{block}} \exp\{-2n_r(r/32-\varepsilon)^2\} \le r$$

for sufficiently large A_d^{ref}.

On $\mathcal{G}$, every refinement pull has gap at most $\kappa_d s$, and a maximizer's cell supplies a sampled fine action with gap at most r. On $\mathcal{G}\cap\mathcal{E}$, the recorded action therefore satisfies $\Delta_f(\widehat{x}) \le r + 2(r/32) \le 2r$. Hence

$$\begin{aligned}\mathbb{E}\big[\mathrm{Reg}_{\mathrm{ref+exp}}(f)\mathbf{1}_{\mathcal{G}}\big] &\le \kappa_d s N_{\mathrm{block}} n_r + 2Tr + T\mathbb{P}(\mathcal{G}\cap\mathcal{E}^{\complement}) \\ &\le C_d\big[sr^{-d-2}\ell_r + Tr\big].\end{aligned}$$

□

Taking every $\mathcal{G}_j$ to be the whole sample space gives the unconditional refinement guarantee in Section 5.1. Under the same hypotheses, using $\mathrm{Reg}_T(f) \le T$ on $\mathcal{G}^{\complement}$ gives

$$\mathbb{E}\mathrm{Reg}_T(f) \le \mathbb{E}[\mathrm{Reg}_{\mathrm{build}}(f)\mathbf{1}_{\mathcal{G}}] + C_d\big[sr^{-d-2}\ell_r + Tr\big] + T\mathbb{P}(\mathcal{G}^{\complement}). \tag{37}$$

C.2 Active-set construction in groups

We prove Lemma 5.2 using the following sampling schedule.

Parameters and updates. Use $f_K^\star$ from Section 5.2. Prescribe

$$\begin{aligned} n_s = \lceil A_d^{\mathrm{ser}} s^{-2}\ell_T\rceil, \qquad L_{\mathrm{ser}} = \lceil \log_2\log_2(4n_s)\rceil \vee 1, \\ w_{\mathrm{ctl}} = \lceil C_d^{\mathrm{ctl}}\ell_T\rceil. \end{aligned} \tag{38}$$

Let $w_s := \min\{K, W - w_{\mathrm{ctl}}\} \ge 1$, $J_s := \lceil K/w_s\rceil$, and partition $[K]$ into fixed groups $\mathcal{I}_1,\dots,\mathcal{I}_{J_s}$ of size at most w_s. Each group uses i batches in pass i and one final refinement batch. Set $D_{\mathrm{ser}} := L_{\mathrm{ser}}(L_{\mathrm{ser}}+1)/2 + 1$; including exploitation, the batch constraint is

$$J_s D_{\mathrm{ser}} + 1 \le B, \qquad K \le w_s\left\lfloor \frac{B-1}{D_{\mathrm{ser}}}\right\rfloor. \tag{39}$$

For $c_\diamond \geq 64$, use

$$n_0 = 1, \qquad c_\ell = 2^{\ell - L_{\mathrm{ser}} + 1},$$
$$n_\ell = \left\lceil c_\ell \sqrt{n_s n_{\ell-1}} \right\rceil, \qquad 1 \leq \ell \leq L_{\mathrm{ser}},$$
$$\varepsilon = 2^{-10-\lceil \frac{1}{2}\log_2 n_s \rceil}, \qquad a_\ell = \sqrt{2\log(c_\diamond K L_{\mathrm{ser}}^2/r)/n_\ell} + 2\varepsilon.$$

In pass i, the reference action is $x_{\bar{k}}$ with $\bar{k} = k_{i-1}$. Let $\bar{\mu}_i$ be the rounded mean of its n_i-pull prefix in the first batch, and set $\lambda_i = \bar{\mu}_i - a_i$. The reference action is fixed throughout the pass, and λ_i is fixed after this prefix.

At step ℓ, the current indicator vector prescribes n_ℓ pulls of x_k for each active index k, and of $x_{\bar{k}}$ otherwise. For each active candidate, compute $\widehat{\mu}_{k,\ell}$ from its n_ℓ rewards using mesh ε, and set $\mathrm{LCB}_{k,\ell} = \widehat{\mu}_{k,\ell} - a_\ell$ and $\mathrm{UCB}_{k,\ell} = \widehat{\mu}_{k,\ell} + a_\ell$. Update the indicator vector by (14). Filler estimates are ignored; the first batch's candidate actions are fixed before the reference prefix is observed. In each nonfinal pass i, among candidates surviving step i, store the largest lower confidence bound and its index in $\mathrm{ch}_i = (k_i^{\mathrm{ch}}, \lambda_i^{\mathrm{ch}})$. In the final pass, $\mathrm{best} = (x^{\mathrm{best}}, \widehat{\mu}^{\mathrm{best}})$ retains the fine-action record from Appendix C.1. The sentinel $\lambda_i = \perp$ triggers the incumbent prefix once; $\varnothing$ denotes an uninitialized action address. Algorithm C.1 specifies the traversal and initialization.

Algorithm C.1 Policy processing groups at fixed scales

Input: Scales $r \leq s$, groups $(\mathcal{I}_j)_{j \leq J_s}$, and the prescribed sampling schedule.

Choose a public initial incumbent $k_0 \in [K]$.
for $i = 1, \ldots, L_{\mathrm{ser}}$ **do**
  $(\bar{k}, \lambda_i) \leftarrow (k_{i-1}, \perp)$.
  if $i < L_{\mathrm{ser}}$ **then**
    $\mathrm{ch}_i \leftarrow (\varnothing, -\infty)$.
  else
    Replace the pass record by $\mathrm{best} \leftarrow (\varnothing, -\infty)$.
  end if
  for $j = 1, \ldots, J_s$ **do**
    Initialize the group indicator vector to all ones.
    Perform steps $1, \ldots, i$ using (14), keeping the reference action fixed and updating the group indicators and record as above; denote the resulting indicator vector by $Z^{(j)}$.
    if $i = L_{\mathrm{ser}}$ **then**
      Refine group $\mathcal{I}_j$ using $Z^{(j)}$ and filler $x_{\bar{k}}$, updating best as in Appendix C.1.
    end if
    Erase $Z^{(j)}$.
  end for
  if $i < L_{\mathrm{ser}}$ **then**
    $k_i \leftarrow k_i^{\mathrm{ch}}$ if initialized; otherwise $k_i \leftarrow k_{i-1}$.
  end if
end for
Use x^{best} for all remaining pulls if initialized, and $x_\circ$ otherwise.

Proof of Lemma 5.2. Fix the constants in the prescribed sampling schedule before choosing C_d in (12). Its memory condition then ensures $w_{\mathrm{ctl}} \leq W/2$ and $J_s \leq C_d(1 + s^{-d}/W)$. The horizon condition and $r \leq s$ imply $\log n_s \leq C_d \ell_T$, hence $D_{\mathrm{ser}} \leq C_d(1 + \log \ell_T)^2$. After increasing the lemma's constant, its batch condition gives $J_s D_{\mathrm{ser}} + 1 \leq B$.

Sampling and memory costs. Choose A_d^{ser} so that $L_{\mathrm{ser}} \geq 2$, $\log_2 n_s \geq 2L_{\mathrm{ser}}$, and $L_{\mathrm{ser}}^2 \leq \sqrt{n_s}$. Removing ceilings from the recurrence gives $v_\ell = n_s^{1-2^{-\ell}} 2^{-2(L_{\mathrm{ser}}-\ell)+L_{\mathrm{ser}}2^{1-\ell}}$. Its consecutive ratios are at least four, and $v_\ell \leq n_\ell \leq 2v_\ell$, yielding $n_\ell \geq 2n_{\ell-1}$ and

$$n_s/2 \leq n_{L_{\mathrm{ser}}} \leq 2n_s, \qquad n_{L_{\mathrm{ser}}-1} \geq n_s/16, \qquad n_\ell \leq 2n_s 4^{-(L_{\mathrm{ser}}-\ell)}. \tag{40}$$

Consequently,

$$\sum_{i=1}^{L_{\mathrm{ser}}} \sum_{\ell \le i} n_\ell \le C n_s, \tag{41}$$

$$L_{\mathrm{ser}} n_1 + \sum_{i=1}^{L_{\mathrm{ser}}} \sum_{\ell=2}^{i} \frac{n_\ell}{\sqrt{n_{\ell-1}}} \le C\sqrt{n_s}. \tag{42}$$

For the second bound, use $n_\ell/\sqrt{n_{\ell-1}} \le c_\ell \sqrt{n_s} + 1$, $\sum_{\ell=2}^{L_{\mathrm{ser}}} (L_{\mathrm{ser}} - \ell + 1) c_\ell \le 8$, and $L_{\mathrm{ser}}^2 \le \sqrt{n_s}$. Candidate and filler pulls total $K \sum_i \sum_{\ell \le i} n_\ell$; incumbent prefixes add $\sum_i n_i \le 4n_s$. Thus construction and refinement use

$$T_{\mathrm{build}} \le C_d K n_s \le C_d s^{-d-2} \ell_T, \qquad T_{\mathrm{ref}} \le C_d r^{-d-2} \ell_r. \tag{43}$$

The horizon hypothesis, with sufficiently large constant, bounds the total number of construction and refinement pulls by $T/2$.

The simultaneous state contains one group indicator vector, the fixed incumbent and its LCB, one block accumulator with control indices, and one pass record. In the final pass, the fine-action record replaces the pass record and persists while later indicator vectors are recomputed. Each estimate updates its record before the accumulator is erased. The horizon condition gives $\log K + \log(1/r) \le C_d \ell_T$; confidence offsets are prescribed, so the rounded-mean and address bounds give at most

$$w_s + w_{\mathrm{ctl}} \le W \tag{44}$$

bits after every observation, for sufficiently large C_d^{ctl}. Refinement rewards update only refinement state. They do not change the reference action and its lower confidence bound or subsequent comparison updates. Every later comparison block therefore retains its prescribed action and sample size conditional on its starting history.

Confidence and safe active sets. There are at most $2KL_{\mathrm{ser}}^2$ candidate and incumbent comparison blocks. Apply (35) with mesh ε, each block's n_ℓ, a_ℓ (or n_i, a_i for a prefix), and failure budget $r/(2KL_{\mathrm{ser}}^2)$. Their joint accuracy event $\mathcal{G}$ has probability at least $1 - r$ and satisfies

$$f(x_{\bar{k}}) - 2a_i \le \lambda_i \le f(x_{\bar{k}}), \tag{45}$$

$$f(x_k) - 2a_\ell \le \mathrm{LCB}_{k,\ell} \le f(x_k) \le \mathrm{UCB}_{k,\ell} \le f(x_k) + 2a_\ell. \tag{46}$$

A representative attaining $f_K^\star$ survives every comparison, and the pass record's LCB is at least its LCB. Hence

$$f_K^\star - f(x_{k_i}) \le 2a_i, \qquad 1 \le i < L_{\mathrm{ser}}. \tag{47}$$

The terminal sample bounds and horizon condition give

$$a_{L_{\mathrm{ser}}-1} \le C_d \sqrt{\log(c_\diamond K L_{\mathrm{ser}}^2/r)/n_s} \le C_d s. \tag{48}$$

For each final group j, let $\mathcal{G}_j$ be the accuracy event for all comparison estimates through that group's screening. It is measurable before the group's refinement batch. On $\mathcal{G}_j$, screening with $\eta = 2a_{L_{\mathrm{ser}}-1} + 2a_{L_{\mathrm{ser}}}$ and slack s gives, for $k \in \mathcal{I}_j$,

$$\mathbf{1}\{f_K^\star - f(x_k) \le s\} \le Z_k^{(j)} \le \mathbf{1}\{f_K^\star - f(x_k) \le s + 6a_{L_{\mathrm{ser}}-1}\}, \qquad k \in \mathcal{I}_j. \tag{49}$$

The covering property gives $f^\star - f_K^\star \le s$, so every point in a retained cell has gap $O_d(s)$. The incumbent used for filler pulls satisfies the same bound. A cell containing a maximizer has representative gap at most s relative to $f_K^\star$ and therefore survives its screening. Thus $\mathcal{G} = \bigcap_j \mathcal{G}_j$ preserves a maximizer's cell, while each $\mathcal{G}_j$ supplies the measurable safety condition required by Appendix C.1.

Regret. Let $\mathrm{Reg}_{\mathrm{build}}^K(f)$ denote construction regret relative to $f_K^\star$. On $\mathcal{G}$, first-step pulls cost at most $KL_{\mathrm{ser}}n_1$. For $i, \ell \ge 2$, a sampled candidate survived step $\ell - 1$, so screening with $\eta = 2a_{i-1} + 2a_i$ gives gap $C(a_{\ell-1} + \zeta_i)$. Incumbent prefixes and fillers in pass $i \ge 2$ have gap at most $2a_{i-1}$; the first prefix is absorbed by the first-step cost. Therefore (41)– (42) yield

$$\begin{aligned}\mathrm{Reg}_{\mathrm{build}}^K(f) &\le C_d K \left[L_{\mathrm{ser}} n_1 + \sum_{i=2}^{L_{\mathrm{ser}}} \sum_{\ell=2}^{i} n_\ell a_{\ell-1} + s \sum_{\ell} n_\ell + \varepsilon n_s \right] \\ &\le C_d K \left[\sqrt{n_s \log(c_\diamond K L_{\mathrm{ser}}^2/r)} + s n_s \right].\end{aligned}$$

Converting from $f_K^\star$ to $f^\star$ adds at most sT_{build}, so

$$\mathrm{Reg}_{\mathrm{build}}(f) \le sT_{\mathrm{build}} + \mathrm{Reg}_{\mathrm{build}}^K(f) \le C_d s^{-d-1}\ell_T. \tag{50}$$

Substitute (50) into (37). Since $\mathbb{P}(\mathcal{G}^{\complement}) \le r$, $s^{-d-1} \le sr^{-d-2}$, and $\ell_r \le C_d \ell_T$, we obtain (13). □

C.3 Hierarchical active-set construction

We use $\gamma_d(L)$ from Section 5.3. For $3 \le B \le T$ and dyadic $0 < r \le s \le 1/16$, prescribe

$$L_B(s) = \min\left\{B-2, \left\lfloor \frac{(d+1)\log(1/s)}{\log(d+2)} \right\rfloor\right\}, \qquad b_{B,r} = \mathrm{bl}(A_d^{\mathrm{hier}} B r^{-d-1}),$$

with A_d^{hier} sufficiently large.

Proof of Lemma 5.3. **Spatial levels.** Set $L = L_B(s) \ge 1$ and $s = 2^{-j_s}$. Choose $\theta_\ell = (1-(d+2)^{-\ell})/(1-(d+2)^{-L})$, $j_\ell = \lceil \theta_\ell j_s \rceil$, and $u_\ell = 2^{-j_\ell}$, $0 \le \ell \le L$. Then $1 = u_0 \ge \cdots \ge u_L = s$. The ideal radii $\widetilde{u}_\ell = s^{\theta_\ell}$ satisfy $\widetilde{u}_{\ell-1}\widetilde{u}_\ell^{-d-2} = s^{-\gamma_d(L)}$; dyadic rounding changes each product by at most 2^{d+2}. Consequently,

$$\sum_{\ell=1}^{L} u_{\ell-1} u_\ell^{-d-2} \le C_d L s^{-\gamma_d(L)}, \qquad \sum_{\ell=1}^{L} u_\ell^{-d-2} \le L s^{-d-2}. \tag{51}$$

Initialize the root cell as active.

One batch per spatial level. Implement the two sweeps and update (15) with $n_\ell = \lceil A_d^{\mathrm{samp}} u_\ell^{-2} b_{B,r} \rceil$ reserved pulls per cell per sweep, rounding mesh $u_\ell/512$, and confidence radius $a_\ell = u_\ell/32$. Compute preceding-level ancestors by (33). Blocks outside $\mathcal{Q}_\ell$ use the first representative of the preceding active set, or $x_\circ$ if it is empty. If $\mathcal{Q}_\ell = \varnothing$, set $\lambda_\ell = -1$ and the new indicator vector to zero. Both sweeps' actions are fixed by the old indicator vector; first-sweep rewards affect only the second sweep's state updates. Retain the old vector until the new vector is complete.

Validity and active-set construction cost. There are at most $C_d L r^{-d} \le C_d B r^{-d}$ potential active-candidate blocks. Apply (35) with the above sample sizes and meshes, tolerance $u_\ell/32$, and conditional failure budget $r/(2C_d B r^{-d})$ per block. For sufficiently large A_d^{samp}, their joint accuracy event $\mathcal{E}_0$ has probability at least $1 - r/2$ and is measurable at the end of active-set construction.

On $\mathcal{E}_0$, induction preserves a maximizer's cell and bounds every retained cell's gap at every point by $\kappa_d u_\ell$. Indeed, if $P_\ell^\star$ contains a maximizer and its ancestor is active, then $f^\star - u_\ell - 2a_\ell \le \lambda_\ell \le f^\star$. The screening rule, with $\mu_\star = f^\star$, $\eta = u_\ell + 2a_\ell$, and $\zeta = u_\ell$, retains $P_\ell^\star$ and gives representative gap at most $2u_\ell + 4a_\ell$ for every survivor. Adding the diameter gives gap at every point at most $3u_\ell + 4a_\ell \le \kappa_d u_\ell$. Thus every candidate and filler action at level ℓ lies in an active preceding-level cell and has gap $O_d(u_{\ell-1})$. Using (51), the construction regret and pull count obey

$$\mathrm{Reg}_{\mathrm{build}}(f)\mathbf{1}_{\mathcal{E}_0} \le C_d L b_{B,r} s^{-\gamma_d(L)}, \qquad T_{\mathrm{build}} \le C_d L b_{B,r} s^{-d-2}. \tag{52}$$

Refinement and resources. Apply Lemma C.1 to the terminal indicator vector as a single group $[K]$, with validity event $\mathcal{E}_0$. Choose the first active representative as filler, or $x_\circ$ if the active set is empty. Its index is determined by the stored vector, and its safety on $\mathcal{E}_0$ follows from the preceding argument. Refinement and exploitation add two batches, so the count is $L + 2 \le B$. The simultaneously stored old and new indicator vectors require $C_d s^{-d}$ bits; addresses, one accumulator, the reference, and the fine-action record require $C_d b_{B,r}$ bits. The accumulator is erased after each estimate is processed. Since $\ell_r \le C_d b_{B,r}$, (16) bounds all construction and refinement pulls by $T/2$ and ensures the state cap after every observation. Finally, substitute (52) and $\mathbb{P}(\mathcal{E}_0^{\complement}) \le r/2$ into (37) to obtain (17). □

D Proof of the upper bound and consequences

We select feasible scales for the fixed-scale constructions of Sections C.2 and C.3 directly in the proof of the joint upper bound, retaining its explicit $C_d \ell_T$ factor.

D.1 Proof of the joint upper bound

Define a horizon-dependent bound on the pass count and the associated batch overhead:

$$L_T^{\mathrm{ser}} = \lceil \log_2(8\ell_T) \rceil \vee 1, \qquad \Gamma_T = L_T^{\mathrm{ser}}(L_T^{\mathrm{ser}}+1)/2 + 1 \asymp (\log\log(eT))^2. \tag{53}$$

Proof of the upper bound in Theorem 3.1. Throughout this proof, write $p := d+2$. Set $b_{B,T} := \mathrm{bl}(8BT)$ and $\upsilon_d := 1/[d(d+3)]$. For $B=1$, a fixed action has regret at most $T = T^{\beta_{d,1}}$. For the remaining cases, set $B_0 := \lceil A_d^{\mathrm{switch}} \Gamma_T \rceil$. We refine the full domain for $B=2$, use the hierarchy for $3 \le B < B_0$, and process cells in groups for $B \ge B_0$. The threshold B_0 accommodates the batch overhead of processing cells in groups. It need not coincide with the dominance boundaries in Appendix D.2, which depend jointly on B and W.

Fix the constants in Lemmas 5.2 and 5.3, and let $H_d \ge 1$ dominate the resource coefficients in (16). There is $J_d \ge 1$, independent of the remaining choices, such that for $B \le T$ and $r \ge T^{-1/p}$,

$$b_{B,r} \le J_d b_{B,T}, \qquad b_{B,r} \le J_d \ell_T.$$

These inequalities follow from the definition of bl and $r^{-d-1} \le T^{(d+1)/p}$. Choose A_d^{opt} large enough for the resource bounds of the construction in groups, then $A_d^{\mathrm{switch}} \ge 1$ large enough to absorb the additive batch overhead, and finally

$$A_d^{\mathrm{scale}} \ge 4H_d \max\{1, J_d(A_d^{\mathrm{switch}}+1)\}.$$

Choose the coefficient in $W \ge C_d \ell_T$ at least $4H_dJ_d$ and above the other memory thresholds, then increase T_d as needed. All constants depend only on d; subsequent unlabelled constants may absorb these fixed choices. The branch and its scales depend only on the known parameters and public seed.

Two batches. We first establish a bound valid for every integer $W \ge 0$: for sufficiently large T,

$$\mathfrak{R}_T(2,W) \le C_d T \min\left\{1, \max\left\{\left(\frac{\log(4T)}{T}\right)^{1/(d+3)}, 2^{-c_d W}\right\}\right\}. \tag{54}$$

Use the safe scale-1 active set $\{\mathcal{X}\}$. For a sufficiently large A_d^{root} and sufficiently small $\xi_d > 0$, let r be the smallest dyadic radius at least

$$\widetilde{r} := \max\left\{\left(\frac{A_d^{\mathrm{root}}\log(4T)}{T}\right)^{1/(d+3)}, 2^{-\xi_d W}\right\}.$$

If r exceeds a fixed small cutoff, a fixed action proves (54). Otherwise, $r < 2\widetilde{r}$ and $\ell_r \le C_d \log(4T)$. Choosing ξ_d and the cutoff sufficiently small, and A_d^{root} sufficiently large, gives

$$C_d \ell_r \le W, \qquad C_d r^{-d-2}\ell_r \le \frac{C_d}{A_d^{\mathrm{root}}} Tr \le T/2.$$

Lemma C.1 applies with one refinement batch followed by one exploitation batch, yielding

$$\mathfrak{R}_T(2,W) \le C_d\left[r^{-d-2}\ell_r + Tr\right] \le C_d Tr.$$

Substituting $r < 2\widetilde{r}$ proves (54).

For $W \ge C_d\ell_T$, increasing the logarithmic-memory constant ensures $T2^{-c_dW} \le T^{\alpha_d}$. Since $\beta_{d,2} = (d+2)/(d+3)$, (54) therefore gives

$$\mathfrak{R}_T(2,W) \le C_d T^{\beta_{d,2}} \ell_T^{1/(d+3)} \le C_d \ell_T \frac{T^{\beta_{d,2}}}{2^2},$$

which proves (4) for $B=2$.

The case $3 \le B < B_0$.

Suppose $3 \le B < B_0$. Define

$$\eta_T := \frac{A_d^{\mathrm{scale}}\Gamma_T b_{B,T}}{T}, \qquad \gamma_\star := \gamma_d(B-2), \qquad \vartheta := \frac{p}{(p+1)\gamma_\star + 1}.$$

Then

$$\gamma_\star \vartheta = \beta_{d,B}, \qquad 1 - \frac{1+\vartheta}{p+1} = \beta_{d,B}, \qquad \vartheta \le \frac{1}{p}.$$

Here s_B is the radius dictated by active-set construction over $B-2$ spatial levels, s_W is the smallest radius whose complete indicator vector fits in memory, and $\widetilde{r}$ balances fine refinement with exploitation. Set

$$s_B := \eta_T^{\vartheta}, \qquad s_W := \left(\frac{A_d^{\mathrm{scale}}}{W}\right)^{1/d}, \qquad \widetilde{s} := s_B \vee s_W, \qquad \widetilde{r} := (\eta_T \widetilde{s})^{1/(p+1)}.$$

For T large enough, $\eta_T \le 1$. Since $p\vartheta \le 1$ and $\widetilde{s} \ge s_B = \eta_T^{\vartheta}$,

$$\eta_T \le \widetilde{s}^{\,p}, \qquad \widetilde{r} = (\eta_T \widetilde{s})^{1/(p+1)} \le \widetilde{s}. \tag{55}$$

After increasing the logarithmic-memory constant and T_d, uniformly over $3 \le B < B_0$ we also have $\widetilde{s} \le 1/32$. Take the smallest dyadic $s \ge \widetilde{s}$ and $r \ge \widetilde{r}$. Dyadic rounding is monotone, so $0 < r \le s \le 1/16$.

By (55), $r \ge \widetilde{r} \ge \eta_T^{1/p} \ge T^{-1/p}$, so the fixed bounds with coefficient J_d apply. Since $s \ge s_W$ and $W \ge C_d \ell_T$,

$$H_d(s^{-d} + b_{B,r}) \le \frac{H_d}{A_d^{\mathrm{scale}}} W + H_d J_d \ell_T \le \frac{W}{4} + \frac{W}{4} \le W.$$

For the horizon condition, keep the switching coefficient explicit: $L = L_B(s) \le B < B_0 \le (A_d^{\mathrm{switch}} + 1)\Gamma_T$ and $r^{-p} \le \eta_T^{-1}$. Consequently,

$$\begin{aligned} H_d L b_{B,r} r^{-p} &\le H_d J_d (A_d^{\mathrm{switch}} + 1)\Gamma_T b_{B,T} \eta_T^{-1} \\ &= \frac{H_d J_d (A_d^{\mathrm{switch}} + 1)}{A_d^{\mathrm{scale}}} T \le T/4. \end{aligned}$$

Both resource inequalities in (16) now hold uniformly over $3 \le B < B_0$, so Lemma 5.3 applies. Since $r^{p+1} \asymp \eta_T s$ and $b_{B,r} \le C_d b_{B,T}$, its terminal refinement term satisfies

$$b_{B,r} s r^{-p} \asymp \frac{b_{B,r}}{\eta_T} r \le C_d T r.$$

If $L = B - 2$, then $\gamma_d(L) = \gamma_\star$ and $s \ge s_B$, so

$$L b_{B,r} s^{-\gamma_d(L)} \le C_d \Gamma_T b_{B,T} \eta_T^{-\gamma_\star \vartheta} \le C_d T^{\beta_{d,B}} (\Gamma_T b_{B,T})^{1-\beta_{d,B}}.$$

If $L < B - 2$, then $L = \lfloor (p-1)\log(1/s)/\log p \rfloor$. For $q = p^{-L} \le \min\{1/p, p s^{p-1}\}$,

$$s^{-\gamma_d(L)} = s^{-(p-1)} \exp\left(\frac{(p-1)q\log(1/s)}{1-q}\right) \le C_d s^{-(p-1)},$$

since $s^{p-1}\log(1/s)$ is bounded. Using $\eta_T \le s^p$ from (55),

$$\begin{aligned} L b_{B,r} s^{-\gamma_d(L)} &\le C_d \Gamma_T b_{B,T} s^{-(p-1)} \le C_d T \eta_T s^{-(p-1)} \\ &\le C_d T (\eta_T s)^{1/(p+1)} \le C_d T r. \end{aligned}$$

Finally, $s = s_B \vee s_W$ up to a factor two and $r \asymp (\eta_T s)^{1/(p+1)}$ imply

$$T r \le C_d \max\left\{ T^{\beta_{d,B}} (\Gamma_T b_{B,T})^{1-\beta_{d,B}}, T^{\frac{d+2}{d+3}} (\Gamma_T b_{B,T})^{1/(p+1)} (1 \vee W)^{-\upsilon_d} \right\}.$$

Since $B \le C_d \Gamma_T$, $b_{B,T} \le C_d \ell_T$, $\Gamma_T = O((\log \ell_T)^2)$, and $1 - \beta_{d,B} \le 1 - \alpha_d = 1/p$,

$$B^2 (\Gamma_T b_{B,T})^{1-\beta_{d,B}} \le C_d \ell_T, \qquad (\Gamma_T b_{B,T})^{1/(p+1)} B^{\upsilon_d} \le C_d \ell_T$$

for $T \ge T_d$. Using $1 + (B-1)W \le B(1 \vee W)$ proves that the hierarchical policy's expected regret is bounded by (4).

The case $B \geq B_0$. Suppose $B \geq B_0$. Define

$$\sigma_T = \left(\frac{A_d^{\mathrm{opt}} \ell_T}{T}\right)^{1/p}, \qquad \Xi = \left(\frac{T}{\ell_T}\right)^{d/p} \wedge \left(1 + \frac{(B-1)W}{A_d^{\mathrm{opt}} \Gamma_T}\right).$$

Since $B \geq B_0$ and $W \gtrsim_d \ell_T$,

$$\Xi \gtrsim_d \min\left\{\left(\frac{T}{\ell_T}\right)^{d/p}, \ell_T\right\} \longrightarrow \infty$$

uniformly over this regime. Choose

$$\begin{aligned} \widetilde{s} &= \max\left\{\sigma_T, \left(1 + \frac{(B-1)W}{A_d^{\mathrm{opt}} \Gamma_T}\right)^{-1/d}\right\}, \\ \widetilde{r} &= \left(\frac{\widetilde{s} \ell_T}{T}\right)^{1/(p+1)}, \qquad s = \lceil \widetilde{s} \rceil_2, \qquad r = \lceil \widetilde{r} \rceil_2. \end{aligned} \tag{56}$$

Because $A_d^{\mathrm{opt}} \geq 1$, this choice satisfies $\widetilde{s} = \max\{\sigma_T, \Xi^{-1/d}\}$. For sufficiently large T_d, the rounded scales lie in $(0, 1/16]$ and satisfy $r \leq s$.

The memory condition of Lemma 5.2 holds by our choice of the coefficient in $W \geq C_d \ell_T$. Since $s^{-d} \leq \Xi$ and $(1 + \log \ell_T)^2 \asymp \Gamma_T$, its batch requirement follows from

$$\begin{aligned} C_d(1 + s^{-d}/W)(1 + \log \ell_T)^2 &\leq C_d' \Gamma_T + \frac{C_d' \Xi \Gamma_T}{W} \\ &\leq C_d' \Gamma_T + \frac{C_d' \Gamma_T}{W} + \frac{C_d'(B-1)}{A_d^{\mathrm{opt}}} \leq B - 1. \end{aligned}$$

The last inequality uses the fixed choices of A_d^{opt} and A_d^{switch}.

The choice $s \geq \sigma_T$ and $r^{p+1} \asymp s \ell_T / T$ give

$$r^{-p} \ell_T \leq C_d T \big(A_d^{\mathrm{opt}}\big)^{-1/(p+1)}.$$

Thus the horizon condition in Lemma 5.2 holds by the choice of A_d^{opt}, and the lemma applies. Moreover,

$$s r^{-p} \ell_T \asymp T r. \tag{57}$$

For the resulting policy $\mathcal{A}$, substituting (56) into Tr gives

$$\sup_{f \in \mathrm{Lip}_1(\mathcal{X}),\, \nu \in \mathfrak{D}(f)} \mathbb{E}_{\nu, \mathcal{A}}\big[\mathrm{Reg}_T(f)\big] \leq C_d T^{\frac{d+2}{d+3}} \ell_T^{\frac{1}{d+3}} \Xi^{-\upsilon_d}. \tag{58}$$

The two terms defining Ξ correspond to statistical saturation and the memory–batch budget. Since A_d^{opt} is fixed,

$$\begin{aligned} T^{\frac{d+2}{d+3}} \ell_T^{\frac{1}{d+3}} \Xi^{-\upsilon_d} &\leq C_d \left[T^{\alpha_d} \ell_T^{1/(d+2)} \vee T^{\frac{d+2}{d+3}} (1 + (B-1)W)^{-\upsilon_d} \ell_T^{1/(d+3)} \Gamma_T^{\upsilon_d}\right] \\ &\leq C_d \ell_T \left[T^{\alpha_d} \vee T^{\frac{d+2}{d+3}} (1 + (B-1)W)^{-\upsilon_d}\right], \end{aligned}$$

where the last step uses $\Gamma_T = O((\log \ell_T)^2)$. This proves (4) in the large-batch regime. The three constructions have batch boundaries fixed in advance and together cover every $1 \leq B \leq T$. □

D.2 Dominant terms and resource boundaries

We compare the three displayed terms in Theorem 3.1, using $Q_{T,B}$ from (6). The algebraic comparisons hold for all admissible budgets. They characterize minimax regret within the theorem's logarithmic factor only when $W \geq C_d \ell_T$ and T is sufficiently large.

Dominance and memory saturation. The conditions stated in the main text follow by comparing the normalized terms in (6). Equality between two terms gives a dominance boundary only when their common value is at least the third term.

For example, the sequential and joint terms are equal at $1 + (B-1)W = T^{d/(d+2)}$, but this is a boundary between dominant terms only when $Q_{T,B} \le 1$. Similarly, the batch-depth and joint terms are equal at

$$1 + (B-1)W = T^{d/(d+2)} Q_{T,B}^{-d(d+3)} = B^{2d(d+3)} T^{d(d+2)-d(d+3)\beta_{d,B}}, \tag{59}$$

and this crossing is visible above the sequential term only when $Q_{T,B} \ge 1$. The sequential and batch-depth terms coincide when $Q_{T,B} = 1$ and are maximal there exactly when $1+(B-1)W \ge T^{d/(d+2)}$. All three terms coincide when $Q_{T,B} = 1$ and $1 + (B-1)W = T^{d/(d+2)}$. An exact intersection need not occur at admissible integer budgets.

If the memory threshold in Section 3.2 lies below $C_d \ell_T$, the joint term is never largest in the matching-memory regime.

Increasing the batch budget. Direct subtraction of exponents gives

$$\log Q_{T,B} = \frac{\alpha_d \log T}{(d+2)^B - 1} - 2 \log B.$$

For this comparison, extend B to real values. Both terms on the right decrease strictly with B. For $T > 1$, the sequential and batch-depth terms therefore have a unique crossing in $(1, T)$, specified by

$$\frac{\alpha_d \log T}{(d+2)^B - 1} = 2 \log B. \tag{60}$$

For fixed d, its location satisfies $B \sim \log\log T / \log(d+2)$. Indeed, at $B = (1-\epsilon)\log\log T/\log(d+2)$, the first term in $\log Q_{T,B}$ has order $(\log T)^{\epsilon}$ and dominates $2\log B$; at the corresponding $(1+\epsilon)$ scale, it tends to zero. Integer budgets use the adjacent integers around this crossing.

For fixed $W > 0$, the joint term also decreases with B. Moreover, the ratio of the batch-depth term to the joint term decreases strictly for real $B = b \ge 2$. Its logarithmic derivative is

$$\beta'_{d,b} \log T - \frac{2}{b} + \frac{W}{d(d+3)(1+(b-1)W)} \le \beta'_{d,b} \log T - \frac{2}{b} + \frac{1}{d(d+3)(b-1)} < 0,$$

because $\beta'_{d,b} < 0$ and $d(d+3) \ge 4$. At $B = 1$, the batch-depth term equals T and is largest. At $B = T$ and integer $W \ge 1$, the sequential term is largest. Thus, ignoring ties, increasing an integer batch budget can visit batch-depth, joint, and sequential dominance in that order, or skip the joint regime. It cannot return from joint to batch-depth dominance. The sequential term dominates precisely when both $Q_{T,B} \le 1$ and $1 + (B-1)W \ge T^{d/(d+2)}$.

These exact crossings concern the displayed expressions. Terms differing only by logarithmic factors cannot be distinguished at $\widetilde{\Theta}_d$ precision. In particular, the first budget at which the sequential term is largest need not equal the minimum budget for the near-optimal regret target in Corollary 3.2.

Necessity of the three regret terms. Each term can exceed the other two by a polynomial factor within the matching-memory regime. For sufficiently large T, consider the following integer budgets.

Sequential dominance. Take $B = T$ and $W = \lceil C_d \ell_T \rceil$. Relative to T^{α_d}, the batch-depth and joint terms have orders

$$T^{-2} \qquad \text{and} \qquad T^{-\frac{2}{d(d+2)(d+3)}} \ell_T^{-\frac{1}{d(d+3)}},$$

respectively. Both are polynomially smaller than the sequential term.

Batch-depth dominance. Take $B = 2$ and $W = \lceil T^{d/(d+2)} \rceil$. The joint term has order T^{α_d}, while $\beta_{d,2} = (d+2)/(d+3)$ gives the batch-depth term $T^{(d+2)/(d+3)}/4$. Its ratio to either of the other terms has order $T^{1/((d+2)(d+3))}$. More generally, (59) at $B = 2$ reduces to $1 + W = 4^{d(d+3)}$. Thus, throughout $W \ge C_d \ell_T$ and for sufficiently large T, $\mathfrak{R}_T(2, W) = \widetilde{\Theta}_d(T^{(d+2)/(d+3)})$.

Joint dominance. Take $B = 3$ and $W = \lceil C_d \ell_T \rceil$. The joint term has order

$$T^{\frac{d+2}{d+3}} \ell_T^{-\frac{1}{d(d+3)}}.$$

Since $\alpha_d < \beta_{d,3} < (d+2)/(d+3)$, it exceeds both the sequential and batch-depth terms by polynomial factors. For these budgets, the upper-bound proof uses the hierarchy. Joint dominance therefore does not require using the construction in groups.

These examples and the matching upper bound show that all three terms in (3) are needed to characterize minimax regret uniformly up to logarithmic factors.

D.3 Proof of the batch-complexity corollary

Proof of Corollary 3.2.

$$c_d \left[\log\log T \vee \frac{T^{d/(d+2)}}{W \ell_T^{d(d+3)}} \right] \le \mathcal{B}_T(W; \varrho_d) \le C_d \left[\log\log T \vee \frac{T^{d/(d+2)}}{W} \right]. \tag{61}$$

Choose ϱ_d larger than the constant in (4), and set

$$B_+ := \left\lceil C_d \left[\log\log T \vee \frac{T^{d/(d+2)}}{W} \right] \right\rceil.$$

Since $W \gtrsim_d \ell_T$, we have $B_+ \le T$ for sufficiently large T. After increasing C_d,

$$(B_+ - 1)W \ge c_d T^{d/(d+2)}, \qquad T^{\beta_{d,B_+} - \alpha_d} = \exp\left(\frac{\alpha_d \log T}{(d+2)^{B_+} - 1} \right) \le C_d.$$

Substitution into (4) gives $\mathfrak{R}_T(B_+, W) \le \varrho_d \ell_T T^{\alpha_d}$ and therefore the upper bound in (61).

Conversely, the joint term in (3) gives, for every $\Lambda > 0$,

$$\mathfrak{R}_T(B, W) \le \Lambda T^{\alpha_d} \quad \Longrightarrow \quad 1 + (B-1)W \ge c_d T^{d/(d+2)} \Lambda^{-d(d+3)}.$$

If $\mathfrak{R}_T(B, W) \le \varrho_d \ell_T T^{\alpha_d}$, take $\Lambda = \varrho_d \ell_T$. For sufficiently large T, subtracting one and dividing by W yields

$$B \ge c_d \frac{T^{d/(d+2)}}{W \ell_T^{d(d+3)}}.$$

The batch-depth term also requires $B^{-2} e^{u_B} \le C_d \ell_T$, where $u_B := \alpha_d \log T / [(d+2)^B - 1]$. If $B \le c_d \log\log T$, with c_d sufficiently small, then $(d+2)^B \le (\log T)^{1/2}$, so $u_B \ge c_d \sqrt{\log T}$ and the requirement fails. Hence $B \ge c_d \log\log T$. Combining the two necessary conditions proves (61); (7) follows after suppressing powers of ℓ_T. □

D.4 Finite-arm comparison

We derive the comparison of regret upper bounds in Section 5.4. Recall $\ell_T = \log(eT)$ and the logarithmic batch overhead Γ_T from (53).

Admissible grids. For a regular K-point grid, Lipschitz continuity bounds the approximation contribution to regret by $C_d T K^{-1/d}$. The formal version of Huang et al. (2026, Theorem 2) uses a block size $S \in [K]$. For constants $c_0, C_0 > 0$ and $T \ge c_0 K$, its memory, batch count, and finite-arm regret are bounded by

$$C_0 S \ell_T, \qquad C_0 \frac{K}{S} \Gamma_T, \qquad C_0 \sqrt{K T \ell_T}\, \Gamma_T,$$

respectively, after increasing C_0 if necessary.

Choose $S = \min\{K, \lfloor W/(C_0 \ell_T) \rfloor\}$. For $W \ge 2 C_0 \ell_T$, this gives

$$\frac{K}{S} \le \max\left\{ 1, \frac{2 C_0 K \ell_T}{W} \right\}.$$

Consequently, for sufficiently large C and sufficiently small $c > 0$, the conditions

$$W \geq C\ell_T, \qquad B \geq C\Gamma_T, \qquad 2 \leq K \leq K_{\max} := \left\lfloor \frac{cBW}{\ell_T \Gamma_T} \right\rfloor \wedge \left\lfloor \frac{T}{c_0} \right\rfloor$$

ensure that this application fits the budgets. The resulting continuum regret bound is

$$C_d T K^{-1/d} + C_0 \sqrt{KT\ell_T}\, \Gamma_T.$$

Rounding the grid size to a dth power changes only dimension-dependent constants.

Optimizing the direct guarantee. Suppress logarithmic factors for the following comparison of polynomial orders. Balancing $TK^{-1/d}$ and $\sqrt{KT}$ gives $K \asymp_d T^{d/(d+2)}$. If $K_{\max} \leq T^{d/(d+2)}$, then $\sqrt{KT} \leq TK^{-1/d}$ throughout the permitted range, so choosing K of order $K_{\max}$ is optimal up to constants. If $K_{\max} \geq T^{d/(d+2)}$, the balancing choice is available. Thus

$$\min_{2 \leq K \leq K_{\max}} \left\{ TK^{-1/d} + \sqrt{KT} \right\} \asymp_d T^{\alpha_d} \vee TK_{\max}^{-1/d}.$$

Restoring logarithmic factors, the optimized stated bound has polynomial order $T^{\alpha_d} \vee T(BW)^{-1/d}$. Comparing this expression with the joint term gives the regret penalties stated in Section 5.4.

The one-dimensional example. For $d = 1$, $W \asymp \log T$ with a sufficiently large constant, and $B \asymp T^{1/4}/\log T$, the preceding applicability conditions hold and $K_{\max} = \widetilde{\Theta}(T^{1/4})$. At the optimizing grid size $K \asymp K_{\max}$,

$$TK^{-1} = \widetilde{\Theta}(T^{3/4}), \qquad \sqrt{KT} = \widetilde{\Theta}(T^{5/8}), \qquad T^{3/4}(1 + (B-1)W)^{-1/4} \asymp T^{11/16}.$$

The direct-grid bound is therefore of polynomial order $T^{3/4}$. Since $B \gg \log\log T$, our batch-depth term is bounded by the statistical term of order $T^{2/3}$, which is smaller than $T^{11/16}$. Theorem 3.1 thus gives $\widetilde{O}(T^{11/16})$, a separation of polynomial order $T^{1/16}$ between the stated guarantees.

E Proof of the one-bit regret bounds

We prove Theorem 6.1 using the policy maps, boundary states, and padding conventions of Appendix A.

Fixed boundaries. Conditional on the public seed, fixed boundaries restrict each round's action to an instance-independent set of at most 2^W points. Lemma A.2 therefore gives

$$\mathfrak{R}_T^{\mathrm{fix}}(B, W) \geq c_d T 2^{-W/d}. \tag{62}$$

For $W = 1$, combining this bound with $\mathrm{Reg}_T(f) \leq T$ gives $\mathfrak{R}_T^{\mathrm{fix}}(B, 1) = \Theta_d(T)$; compare Zhu & Huang (2025, Section 4). It remains to prove the matching bounds with adaptive boundaries.

E.1 Lower bound with adaptive boundaries

Two-state contraction. The entropy bound (19) controls the number of bits carried across boundaries. For one bit, we additionally bound how well the final state of each batch distinguishes disjoint local alternatives. The following estimate is uniform in the batch length and can then be composed across nonterminal boundaries.

Use $H^2(P, Q) = \int (\sqrt{dP} - \sqrt{dQ})^2$, with respect to any common dominating measure, and $\mathrm{TV}(P, Q) \leq H(P, Q)$. For deterministic maps $\kappa_t : \{0,1\}^2 \to \{0,1\}$, put

$$K_t^{(u)}(i, j) = (\tfrac{1}{2} - u)\mathbf{1}\{\kappa_t(i, 0) = j\} + (\tfrac{1}{2} + u)\mathbf{1}\{\kappa_t(i, 1) = j\}.$$

Inputs are independent under each law. The reference has $u_{0,t} = 0$, and alternatives $v \in [K]$ have biases satisfying

$$|u_{v,t}| \leq \delta \leq \tfrac{1}{4}, \qquad \#\{v : u_{v,t} \neq 0\} \leq 1 \quad \text{for every } t. \tag{63}$$

Starting from a common state $i_\circ$, let $Q_{v,0} = \delta_{i_\circ}$ and $Q_{v,t} = Q_{v,t-1} K_t^{(u_{v,t})}$.

Lemma E.1 (Two-state spatial separation)**.** *Uniformly in n, K and the transitions,*

$$\sum_{v=1}^{K} H^2(Q_{0,n}, Q_{v,n}) \le 48\delta^2. \tag{64}$$

Proof. Let $\mathcal{I} = \{t : \exists i,\ \kappa_t(i,0) \neq \kappa_t(i,1)\}$. Outside $\mathcal{I}$, all laws use the same kernel. For $t \in \mathcal{I}$, the reference rows have output-one probabilities in $\{0, 1/2, 1\}$, at least one equal to $1/2$. Their common mass $a_t = \sum_j \min_i K_t^{(0)}(i,j)$ is therefore at least $1/2$. The kernels $J_t(i,j) = \min_i K_t^{(0)}(i,j)/a_t$ and $L_t = 2K_t^{(0)} - J_t$ are Markov, and J_t has identical rows. Joint convexity and data processing give

$$H^2(PK_t^{(0)}, QK_t^{(0)}) \le \tfrac{1}{2} H^2(P,Q),$$
$$H^2(QK_t^{(0)}, QK_t^{(u)}) \le H^2(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2+u)) \le 4u^2.$$

Apply the triangle inequality and $(a+b)^2 \le \frac{3}{2}a^2 + 3b^2$. For $D_t = \sum_v H^2(Q_{0,t}, Q_{v,t})$, disjointness gives

$$D_t \le \begin{cases} D_{t-1}, & t \notin \mathcal{I}, \\ \frac{3}{4} D_{t-1} + 12\delta^2, & t \in \mathcal{I}. \end{cases}$$

Since $D_0 = 0$, summing the geometric series proves the bound. □

Composition across boundaries. For laws P, Q of a finite sequential record $\mathsf{Z}_{1:k}$ with common support and conditional kernels P_b, Q_b,

$$H^2(P,Q) \le 4\mathbb{E}_P \sum_{b=1}^{k} H^2\big(P_b(\cdot \mid \mathsf{Z}_{<b}), Q_b(\cdot \mid \mathsf{Z}_{<b})\big). \tag{65}$$

To prove (65) with expectation under P, set $L_b = dQ_{1:b}/dP_{1:b}$, $L_0 = 1$, $\xi_b = L_b/L_{b-1}$, and $\vartheta(x) = (x-1)^2/(x+1)$. For $x > 0$ and $u \ge 0$, the identities

$$\vartheta(xu) - \vartheta(x) - x\vartheta'(x)(u-1) = \frac{4x^2(u-1)^2}{(1+x)^2(1+xu)}, \qquad (\sqrt{u}+1)^2 \le (1+x^{-1})(1+xu)$$

imply, since $\mathbb{E}_P[\xi_b \mid \mathsf{Z}_{<b}] = 1$, $\mathbb{E}_P[\vartheta(L_b) - \vartheta(L_{b-1}) \mid \mathsf{Z}_{<b}] \le 4\mathbb{E}_P[(\sqrt{\xi_b}-1)^2 \mid \mathsf{Z}_{<b}]$. Sum over b and use $(\sqrt{x}-1)^2 \le \vartheta(x)$. For a common independent seed law μ, the same inequality holds conditionally and integrates by $H^2(\mu(dw)P^w, \mu(dw)Q^w) = \int H^2(P^w, Q^w)\,\mu(dw)$.

Proof of the lower bound in Theorem 6.1. For $0 < \delta \le 1/32$, choose $\{z_v : v \in [K]\} = \{2\delta + 4\delta j : 0 \le j < \lfloor (4\delta)^{-1} \rfloor\}^d$. Then $K \ge (8\delta)^{-d}$ and the radius-δ balls are disjoint. Using the notation of Appendix B, set

$$f_0(x) = \tfrac{1}{2}, \qquad f_v(x) = \tfrac{1}{2} + (\delta - \|x - z_v\|_\infty)_+, \qquad G_v = B_\infty(z_v, \delta/2). \tag{66}$$

These are admissible Bernoulli means, with $\Delta_{f_v} \ge (\delta/2)\mathbf{1}_{G_v^\complement}$.

Fix $\mathcal{A} \in \mathfrak{A}_{B,1}$. At a common boundary history $\mathcal{H}_{b-1} = \sigma(\omega, \bar{M}_1, \dots, \bar{M}_{b-1})$, Lemma A.1 fixes the next batch's initial bit, batch boundary, and actions. The update maps, restricted to binary rewards, therefore satisfy Lemma E.1, with disjoint biases from (66). Let $Q_{v,b}$ be the conditional law of $\bar{M}_b$ given this history under f_v. If the reconstructed batch boundary is T or the horizon was already reached, padding makes all these laws the same point mass; otherwise their summed squared Hellinger distance is at most $48\delta^2$. Interior Bernoulli means give common support conditional on each seed. Applying (65) and then integrating against the common seed law, the laws $P_v = \mathcal{L}_v(\omega, \mathbf{M})$ satisfy

$$\sum_v H^2(P_0, P_v) \le 4\mathbb{E}_0 \sum_{b=1}^{B-1} \sum_v H^2(Q_{0,b}, Q_{v,b}) \le 192(B-1)\delta^2. \tag{67}$$

The same reconstruction makes $F_v = N_{G_v}(T)/T$ a common function of this record, and disjointness gives $\sum_v F_v \le 1$. Thus

$$\frac{1}{K}\sum_v \mathbb{E}_v F_v \le \frac{1}{K} + \frac{1}{K}\sum_v \mathrm{TV}(P_0, P_v) \le \frac{1}{K} + \sqrt{\frac{192(B-1)\delta^2}{K}}. \tag{68}$$

Choose $\delta = c_d B^{-1/(d+2)}$ with $c_d \le 1/32$ and $192\,8^d c_d^{d+2} \le 1/16$. The right side is at most $1/2$, so $\sup_v \mathbb{E}_v \mathrm{Reg}_T(f_v) \ge (\delta T/2)(1 - K^{-1}\sum_v \mathbb{E}_v F_v) \ge \delta T/4$. Taking the infimum proves the bound, including $B = 1$, when the boundary-state tuple is empty. □

E.2 Upper bound with adaptive boundaries

Fix $\rho = 2^{-E} \le 1/2$, $\rho_e = 2^{-e}$, and $q_\rho = \lceil A_d^{\mathrm{conf}} \log(e/\rho) \rceil$. The construction has E phases. Write t_e for the end of phase e, U_e for a deterministic upper bound on t_e, and Q_E for the total number of exploration batches.

Lemma E.2 (One-bit fixed-accuracy implementation)**.** *For sufficiently large A_d^{conf}, the construction below has exactly two mutable configurations and, for every path,*

$$U_e \le C_d q_\rho \rho_e^{-d-2} \quad (e \in [E]), \qquad Q_E \le t_E \le U_E. \tag{69}$$

If $U_E \le T/2$, appending exploitation gives a policy $\mathcal{A}$ satisfying

$$\sup_{f \in \mathrm{Lip}_1(\mathcal{X}),\, \nu \in \mathfrak{D}(f)} \mathbb{E}_{\nu,\mathcal{A}}\big[\mathrm{Reg}_T(f)\big] \le C_d\big(q_\rho \rho^{-d-1} + T\rho\big). \tag{70}$$

The policy belongs to $\mathfrak{A}_{B,1}$ whenever also $Q_E + 1 \le B$.

Comparisons on successive grids. Use the $K_e = 2^{de}$ midpoint representatives of Section C.1, with labels $\mathcal{V}_e = \{J_{e-1}+1, \dots, J_e\}$, where $J_0 = 0$ and $J_e = \sum_{a \le e} K_a \le 2K_e$; set $x_0 = x_\circ$. Then $\max_{v \in \mathcal{V}_e} f(x_v) \ge f^\star - \rho_e$. Prescribe

$$n_\ell = q_\rho 2^\ell, \qquad a_\ell = \sqrt{q_\rho / n_\ell}, \qquad L_e = 6 + 2e. \tag{71}$$

Thus $a_{L_e} = \rho_e/8$ and $n_{L_e} = 64 q_\rho \rho_e^{-2}$. Traverse stages (e, v, ℓ) lexicographically, with $\ell = 0, \dots, L_e$ for each challenger v. The comparison variables are an incumbent label I, a flag $F \in \{0,1\}$ indicating whether the current comparison has finished, and the sum C of the query bits in the current stage. We first specify their updates, then realize them through batch boundary times using a single mutable bit.

Each stage makes $2n_\ell$ queries. In query j, pull x_v if $F = 0$ and $j \le n_\ell$, and x_I otherwise. Set $\sigma_j = \mathbf{1}\{j > n_\ell\}$. If the query batch starts at time t, observe Y_{t+1} and set $M_{t+1} = \sigma_j + (1 - 2\sigma_j)\mathbf{1}\{U \le Y_{t+1}\}$, where U is the unused uniform seed coordinate assigned to this query. Set $\widehat{D}_\ell = (C - n_\ell)/n_\ell$; when $F = 0$, this estimates $f(x_v) - f(x_I)$. At stage end, apply

$$(I^+, F^+) = \begin{cases} (v, 1), & F = 0,\ \widehat{D}_\ell > 2a_\ell, \\ (I, 1), & F = 0,\ \widehat{D}_\ell < -2a_\ell, \\ (I, F), & \text{otherwise}, \end{cases} \qquad I' = I^+, \quad F' = \mathbf{1}\{\ell < L_e\} F^+. \tag{72}$$

The incumbent and flag remain unchanged within a stage. The factor $\mathbf{1}\{\ell < L_e\}$ resets the flag for the next challenger after the last stage. Only the current challenger can replace the incumbent, so the ordered traversal makes I nondecreasing on every path.

Recovering comparison variables from time. Each stage has the fixed batch list $(\mathsf{Q}_1, \mathsf{T}_1, \dots, \mathsf{Q}_{2n_\ell}, \mathsf{T}_{2n_\ell}, \mathsf{C})$: one query and transfer per observation, then one control batch. An alignment batch precedes every phase. Finished comparisons keep their slots, use incumbent actions, and ignore further decisions. Thus Q_E is deterministic and includes query, transfer, control, and alignment batches; exploitation has batch index Q_E.

Put $R_e = 2n_{L_e} + 1$ and

$$\mathrm{enc}_e(I, F, C) = 2R_e I + R_e F + C, \qquad \mathrm{dec}_e(z) = \left(\left\lfloor \frac{z}{2R_e} \right\rfloor, \left\lfloor \frac{z \bmod 2R_e}{R_e} \right\rfloor, z \bmod R_e \right).$$

At query starts the invariant is

$$t = D + z, \qquad z = \mathrm{enc}_e(I, F, C), \qquad 0 \le I \le J_e, \quad 0 \le C < R_e, \quad M_t = 0. \tag{73}$$

The public batch index determines D and all stage indices. The current round t then determines $(I, F, C) = \mathrm{dec}_e(t - D)$. Only M_t is the mutable memory state.

Transferring a bit and updating the comparison. A query starting at time t leaves the state M_{t+1} defined above. At time $t+1$, the transfer batch fixes its length at $1+M_{t+1}$, pulls x_I, and sets $M_u = 0$ after each observation:

$$\begin{aligned}(t, D, z, 0) &\xrightarrow{\mathsf{Q}} (t+1, D+1, z, M_{t+1}) \\ &\xrightarrow{\mathsf{T}} (t+2+M_{t+1}, D+2, z+M_{t+1}, 0).\end{aligned} \tag{74}$$

At the transfer start, $z = (t+1) - (D+1)$ determines the incumbent through dec_e. Both batch lengths are positive integers fixed at their starts. After j pairs, $C \le j \le 2n_\ell < R_e$, so no overflow changes the incumbent or flag.

At stage end, compute (I', F') from (72). For $z' = \mathrm{enc}_e(I', F', 0)$, put $g_{e,\ell} = 2n_\ell + 1$ if $\ell < L_e$, and $g_{e,L_e} = 2R_e$. The control batch pulls $x_{I'}$, sets $M_u = 0$ after each observation, and has length

$$\lambda = g_{e,\ell} + z' - z, \qquad t' = t + \lambda = (D + g_{e,\ell}) + z'. \tag{75}$$

Since $I' \ge I$, for $\ell < L_e$ one has $F' \ge F$ and $\lambda \ge 2n_\ell + 1 - C \ge 1$; at $\ell = L_e$, $\lambda \ge 2R_e - R_e F - C \ge 1$. Thus every control resets C, and the last level resets F for the next challenger.

Phase alignment. Initialize $U_0 = t_0 = I_0 = 0$ and prescribe

$$P_e = K_e \sum_{\ell=0}^{L_e} (4n_\ell + g_{e,\ell}), \qquad a_e^{\mathrm{clk}} = U_{e-1} + 1, \qquad U_e = a_e^{\mathrm{clk}} + P_e + 2R_e J_e.$$

At the end of phase $e-1$, recover I_{e-1} from the phase ending time and choose an alignment batch on $x_{I_{e-1}}$ ending at $t = a_e^{\mathrm{clk}} + 2R_e I_{e-1}$, with $M_u = 0$ after each observation. Its length is $U_{e-1} - t_{e-1} + 1 + 2R_e I_{e-1} \ge 1$. At this boundary, $D = a_e^{\mathrm{clk}}$ and $F = C = M_t = 0$. Each query–transfer pair advances D by two, and each control advances it by $g_{e,\ell}$, giving

$$t_e = a_e^{\mathrm{clk}} + P_e + 2R_e I_e \le U_e, \qquad I_e = \frac{t_e - a_e^{\mathrm{clk}} - P_e}{2R_e}, \qquad F = C = M_{t_e} = 0. \tag{76}$$

The second identity recovers the incumbent label from the phase ending time.

Proof of Lemma E.2. **Resources and policy maps.** For $n_\star = n_{L_e}$, geometric summation gives $\sum_{\ell \le L_e} n_\ell < 2n_\star$ and $L_e + 1 \le n_\star$. Hence $P_e + 2R_e J_e \le C_d K_e n_\star$ and $K_e n_\star = 64 q_\rho 2^{(d+2)e}$, so $U_e = \sum_{a \le e}(1 + P_a + 2R_a J_a) \le C_d q_\rho \rho_e^{-d-2}$ for every prefix. By construction, $t_e \le U_e$ for every e, and positive batch lengths give $Q_E \le t_E$. This proves (69), including all waiting rounds. Query updates leave one bit and all other updates leave zero.

When $U_E \le T/2$ and $Q_E + 1 \le B$, the recursions define $(\psi_b, \phi_{b,u})$ on the finite set of schedule-reachable time/bit pairs, with $t < \psi_b \le U_E < T$ for $b < Q_E$. At $b = Q_E$, decode I_E using (76) and exploit to T. On all other inputs use batch boundary T and action $x_\circ$, with padding at T as in Appendix A; initialize $M_0 = 0$. Updates are the query map or zero, selected by the public batch index, and use only the arguments of $\mathcal{U}_{b,u}$. Thus the maps are total and Borel, $\psi_{B-1} = T$, and every batch's actions and boundary are fixed at its start.

Concentration. Fix $f \in \mathrm{Lip}_1(\mathcal{X})$ and $\nu \in \mathfrak{D}(f)$. At stage $\zeta = (e, v, \ell)$, let $\mathcal{H}_\zeta$ be its starting history with only previously used seed coordinates exposed, and let (I_ζ, F_ζ) be its starting pair. Query times are predictable in this filtration, and each query uses a fresh uniform independent of its reward. On $\{F_\zeta = 0\}$ the two actions remain fixed throughout the stage; the first n_ℓ written bits have conditional mean $f(x_v)$, and the last n_ℓ have mean $1 - f(x_{I_\zeta})$. Write $\widehat{D}_\zeta$ for this stage's estimate $\widehat{D}_\ell$ and $D_\zeta = f(x_v) - f(x_{I_\zeta})$ for its mean difference. Hoeffding's bound for these bounded increments gives

$$\mathbb{P}\big(F_\zeta = 0,\ |\widehat{D}_\zeta - D_\zeta| > a_\ell \mid \mathcal{H}_\zeta\big) \le 2\mathbf{1}\{F_\zeta = 0\} e^{-n_\ell a_\ell^2} = 2\mathbf{1}\{F_\zeta = 0\} e^{-q_\rho}.$$

In particular, this conditioning does not expose the entire seed. For the event $\mathcal{E}$ that every unfinished stage is accurate,

$$\mathbb{P}(\mathcal{E}^{\complement}) \le 2e^{-q_\rho} \sum_{e=1}^{E} K_e(L_e + 1) \le C_d \rho^{-d} \log(e/\rho) e^{-q_\rho} \le \rho. \tag{77}$$

Incumbents and regret. On $\mathcal{E}$, acceptance strictly improves the incumbent mean, and rejection discards a strictly worse challenger. At an unresolved last level, $D_\zeta \le 3a_{L_e}$. Consequently,

$$f(x_{I_e}) \ge \max_{v\in\mathcal{V}_e} f(x_v) - 3a_{L_e} \ge f^\star - 11\rho_e/8, \qquad \Delta_f(x_{I_e}) \le 2\rho_e. \tag{78}$$

Every incumbent used in phase e, including alignment and control, has gap at most $4\rho_e$: this is trivial for $e = 1$ and follows from $2\rho_{e-1} = 4\rho_e$ and improving means thereafter.

Let $N_{e,v}$ count unfinished challenger queries and $\Delta_v = \Delta_f(x_v)$. For $\Delta_v < 8\rho_e$, the full schedule has at most $\sum_\ell n_\ell \le C_d q_\rho \rho_e^{-2}$ challenger pulls. Otherwise let $\ell_\star = \min\{\ell : a_\ell \le \Delta_v/8\}$. It exists and is positive; an unfinished stage there satisfies

$$\widehat{D}_{\ell_\star} \le -\Delta_v + 4\rho_e + a_{\ell_\star} \le -3\Delta_v/8 < -2a_{\ell_\star}.$$

The challenger is therefore rejected by that level, and $\sum_{\ell\le\ell_\star} n_\ell \le C_d q_\rho/\Delta_v^2$. Together, the two cases give

$$N_{e,v} \le \frac{C_d q_\rho}{(\Delta_v \vee \rho_e)^2}, \qquad \Delta_v N_{e,v} \le \frac{C_d q_\rho}{\rho_e}.$$

All other pulls use an incumbent. Using $t_e - t_{e-1} \le U_e$ and (69), including all waiting rounds,

$$\sum_{t=t_{e-1}+1}^{t_e} \Delta_f(A_t) \le \sum_{v\in\mathcal{V}_e} \Delta_v N_{e,v} + 4\rho_e(t_e - t_{e-1}) \le C_d q_\rho \rho_e^{-d-1}.$$

Sum geometrically, exploit x_{I_E}, and charge $T\mathbb{P}(\mathcal{E}^{\complement}) \le T\rho$ once to obtain (70). □

Proof of the upper bound in Theorem 6.1. For $B = 1$, use a fixed action. For $B \ge 2$, put $\widetilde{\rho}_B = (\log(eB)/B)^{1/(d+2)}$ and choose $A_d^{\mathrm{rad}} \ge 1$ after the confidence and resource constants. If $A_d^{\mathrm{rad}}\widetilde{\rho}_B > 1/4$, a fixed action suffices since $T \le 4A_d^{\mathrm{rad}} T\widetilde{\rho}_B$. Otherwise take $\rho = \lceil A_d^{\mathrm{rad}}\widetilde{\rho}_B\rceil_2 \le 1/2$ using Section C.1. Then

$$q_\rho \le C_d \log(eB), \qquad U_E \le \frac{C_d}{(A_d^{\mathrm{rad}})^{d+2}} B \le B/2, \qquad q_\rho \rho^{-d-1} \le C_d T\rho.$$

Since $Q_E \le t_E \le U_E$, one has $Q_E + 1 \le B$ for $B \ge 2$. Also $t_E \le U_E \le B/2 \le T/2$, leaving a nonempty exploitation batch. Lemma E.2 gives regret at most $C_d T\rho \le C_d' T\widetilde{\rho}_B$. □