STOCHASTIC BANDIT CONVEX OPTIMIZATION WITH BLOCKING CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study stochastic bandit convex optimization with blocking constraints. After an action is selected, all actions within distance ε become unavailable for the next $p-1$ rounds, where p is the blocking period. Since repeatedly selecting a minimizer is, in general, infeasible, we measure regret against an offline benchmark defined by the infimum average loss over feasible cycles of p actions. We show that the convexity of the loss function and the action set allows the search for this benchmark to be restricted to a ball of radius $2p\varepsilon$ around a global minimizer. This motivates the design and analysis of an algorithm that first learns a point with low loss and then uses lower confidence bound (LCBs) to explore feasible cycles in its neighborhood. By grouping repeated queries and controlling transitions between cycles, our algorithm respects the blocking constraints and accounts for all exploration and transition costs. Under a geometric condition ensuring local feasibility, we establish a regret upper bound of $\widetilde{O}\big(\sqrt{T}[1+(\varepsilon\sqrt{T})^{d/(d+2)}]\big)$ and a minimax lower bound of $\Omega\big(\sqrt{T}[1+(\varepsilon\sqrt{T})^{(d-1)/(d+3)}]\big)$, for fixed action space dimension $d \geq 2$ and blocking period $p > 2$. The lower bound holds on the unit Euclidean ball with Gaussian noise. Both bounds recover the classical $\sqrt{T}$ regret when $\varepsilon\sqrt{T}$ is small. For $p = 2$, a sharper decomposition based on the direction between the two actions improves the upper bound to match the lower bound, establishing minimax optimal dependence on T and ε up to logarithmic factors.

1 INTRODUCTION

In stochastic bandit convex optimization (BCO), a learner seeks to minimize an unknown convex loss over a convex action set using only noisy function evaluations. The standard model assumes that every action remains available at every round. In some applications, however, selecting an action may temporarily make nearby actions unavailable. For example, when mapping the stiffness of articular cartilage, an indentation at one site can alter the response to a subsequent probe at a nearby site before the tissue returns to equilibrium (Berni et al., 2022). In industrial nondestructive testing, laser spot thermography detects defects by locally heating a material's surface and recording its thermal response. Residual heat can interfere with measurements at nearby locations, so scan schedules can visit more distant positions while the heated region cools (Roemer et al., 2023). These settings motivate a temporary exclusion of nearby probe locations in a continuous action space. Multi-armed bandits with blocking constraints are well studied (Basu et al., 2019; Cesa-Bianchi et al., 2026), but their extension to BCO over continuous action spaces remains less explored. We study stochastic BCO with *blocking constraints*.

Specifically, after playing $x_t \in \mathcal{X} \subset \mathbb{R}^d$, the learner cannot select any point in its ε-neighborhood for the next $p-1$ rounds, with the same radius $\varepsilon > 0$ and period $p \geq 2$ at every location. In our stochastic model, the learner observes $f(x_t) + \eta_t$ for a fixed unknown convex loss f. Basu et al. (2019) study stochastic finite-arm blocking with arm-dependent delays, where a play disables only the chosen arm. The equal-delay special case of their model has a combinatorial semi-bandit formulation. Here each play excludes a continuum of points, and availability depends on the previous $p-1$ actions. This changes the optimization task: exploiting the convex loss requires several low-loss points that satisfy pairwise distance constraints. The feasible set of such configurations is nonconvex even though X and f are convex. Since the actions used to learn a good configuration also determine which points remain available, both evaluating candidates and moving between them must follow an

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.

admissible trajectory. Thus a common blocking period and a stationary loss already create geometric and statistical difficulties.

The blocking constraint calls for a comparator that reflects feasible long-run play. Measuring regret against the best fixed action can penalize unavoidable losses: on some instances, every admissible policy incurs regret linear in T against it, even with full knowledge of f. A myopic greedy comparator can be too weak: even when its trajectory remains feasible, its average loss can exceed the optimal feasible long-run value by a positive constant (Appendix B). An ε-packing of p points provides a natural comparator under a common blocking period. Every p consecutive admissible actions form such a packing, while repeating any feasible packing in a fixed order is admissible. Assuming a feasible packing exists, compactness of X and continuity of f ensure that a packing of minimum average loss exists. Its value lower-bounds the average loss of every complete p-round block, and cycling that packing attains the same value in the long run. Thus the packing benchmark equals the optimal long-run average loss achievable with full knowledge of f. We measure regret against T times this value and ask how quickly a learner can approach it using noisy feedback while keeping every action admissible.

Our key observation is that convexity localizes the packing search: restricting it to within distance $2p\varepsilon$ of a low-loss anchor increases the optimal average loss by at most the anchor's suboptimality (Proposition 2.1). This follows by compressing feasible packings toward the anchor while preserving separation. We therefore first find an anchor using stochastic BCO, then refine a packing in its neighborhood. To implement the anchor phase, we embed batched BCO queries in feasible local cycles. Switching between packings requires additional actions because recent history may block the next packing. We construct transitions in $O(p+p^2/(D/\varepsilon-p+1))$ rounds when $D := \operatorname{diam}(X) \geq p\varepsilon$, and in at most $2p-1$ rounds under an inball condition with sufficiently small ε. The latter condition holds for every full-dimensional compact convex X once ε is sufficiently small. Refinement uses cellwise lower confidence bounds to select feasible packings in the localized continuous region, with each packing played for a batch of complete cycles. In the inball regime, this yields, for fixed p, d and $T \geq p$, a regret upper bound of $\widetilde{O}(\sqrt{T}[1 + (\varepsilon\sqrt{T})^{d/(d+2)}])$, suppressing Lipschitz and logarithmic factors. Our lower bound isolates a statistical difficulty that persists when packings are easy to connect. On the Euclidean unit ball, we hide a better packing among many directions; each informative region has diameter less than ε, so blocking limits how often it can be sampled even though transitions take only $O(p)$ rounds. With $d \geq 2$, Gaussian noise, convex 1-Lipschitz losses, and $0 < \varepsilon \leq 1/[2(2p)^{1/d}]$, this gives a minimax lower bound of $\Omega(\sqrt{T}[1 + (\varepsilon\sqrt{T})^{(d-1)/(d+3)}])$ for fixed p, d. Thus the lower bound captures the cost of identifying a good packing direction under restricted feedback, even when reconfiguration is inexpensive. The general-period upper and lower bounds recover the classical $\sqrt{T}$ scale when $\varepsilon\sqrt{T}$ is small, but their blocking-dependent exponents differ. For $p = 2$, covering only the separation directions of feasible pairs reduces the remaining search to convex subproblems. This yields an upper bound matching the lower rate up to logarithms on the unit ball: the directional structure of pairs suffices to attain the statistical limit without searching the full local packing space cell by cell.

Related work. Basu et al. (2019) study stochastic blocking bandits with finitely many arms. After an arm is played, that arm is unavailable for a fixed number of rounds that depends on the arm. They show that a greedy policy is asymptotically $(1 - 1/e)$-optimal, and they obtain logarithmic regret against it. This model has been extended to contextual information (Basu et al., 2021), combinatorial actions (Atsidakou et al., 2021), matroid constraints (Papadigenopoulos & Caramanis, 2021), and adversarial rewards (Bishop et al., 2020; Cesa-Bianchi et al., 2026). When all arms have the same blocking time in the stochastic finite-arm setting, the problem reduces to a combinatorial semi-bandit Basu et al. (2019); Chen et al. (2013); Kveton et al. (2015). Cesa-Bianchi et al. (2026) study adversarial rewards with the same blocking time, showing that the minimax regret against periodic policies is of order $T^{2/3}$. In our continuous setting, the corresponding object is a feasible p-point packing. The packing space is nonconvex because of the separation constraints, and moving between packings must also satisfy the blocking rule. This changes both the benchmark and the learning problem. The learner must find several separated low-loss actions and must also use feasible actions when moving between candidate solutions.

For stochastic BCO with Lipschitz convex losses, Agarwal et al. (2013) obtain $\widetilde{O}(\operatorname{poly}(d)\sqrt{T})$ regret. Later work improves the dimension dependence and develops other zeroth-order methods (Lattimore

& György, 2021; 2023). These results support our search for a low-loss anchor. Under our blocking constraint, the benchmark is a feasible cycle of separated low-loss actions that must be realized through an admissible trajectory.

Continuum-armed and metric bandits exploit similarity between nearby actions in large or infinite spaces (Agrawal, 1995; Kleinberg et al., 2008; Bubeck et al., 2011a). Their complexity depends on the geometry of near-optimal regions, often described by covering, packing, zooming, or near-optimality dimensions. For globally L-Lipschitz functions on a d-dimensional Euclidean domain, minimax regret scales as $L^{d/(d+2)}T^{(d+1)/(d+2)}$ up to logarithms (Bubeck et al., 2011b). The same horizon exponent appears in the blocking-dependent term of our general-period upper bound, reflecting a related local covering–estimation tradeoff. Here convexity first localizes the search around a global minimizer, and local estimates compare separated packings rather than individual actions. Further related work is given in Appendix A.

2 PRELIMINARIES

Let $\mathcal{X} \subset \mathbb{R}^d$ be compact, convex, and full-dimensional, with diameter $D := \operatorname{diam}(\mathcal{X}) > 0$ and a known inball $\overline{B}(c, r_0) \subseteq \mathcal{X}$ of radius $r_0 > 0$. All norms are Euclidean. The unknown loss $f : \mathcal{X} \to [0, 1]$ is convex and L-Lipschitz. At each round t, the learner chooses $x_t \in \mathcal{X}$ and observes $Y_t = f(x_t) + \eta_t$. The noise variables $\{\eta_t\}_{t\geq 1}$ are i.i.d., mean-zero, and σ-sub-Gaussian, i.e., $\mathbb{E}[e^{\lambda\eta_t}] \leq e^{\lambda^2\sigma^2/2}$ for every $\lambda \in \mathbb{R}$. Write $f^\star := \min_{x\in\mathcal{X}} f(x)$, which exists by compactness.

For a blocking radius $\varepsilon > 0$ and a period $p \geq 2$, choosing x_t blocks all points at distance strictly less than ε from x_t for the next $p - 1$ rounds. An action sequence is therefore *admissible* if

$$\|x_t - x_{t-s}\| \geq \varepsilon, \qquad 1 \leq s \leq \min\{p - 1, t - 1\}, \quad t \geq 1.$$

Equivalently, every block of p consecutive actions consists of mutually ε-separated points. For any $K \subseteq \mathcal{X}$, let $\mathcal{P}_p^\varepsilon(K)$ denote the *space of ordered ε-packings*:

$$\mathcal{P}_p^\varepsilon(K) := \big\{(u_1, \ldots, u_p) \in K^p : \|u_i - u_j\| \geq \varepsilon \text{ for all } 1 \leq i < j \leq p\big\}.$$

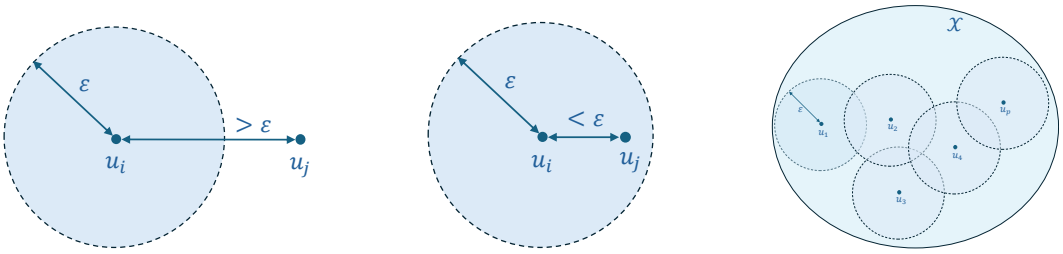
For $P = (u_1, \ldots, u_p) \in \mathcal{P}_p^\varepsilon(K)$, write $B(P) := \{u_1, \ldots, u_p\}$ for its underlying unordered *packing set*. Figure 1 illustrates the separation constraint for two consecutive actions and for a p-point packing set. For a packing $P = (u_1, \ldots, u_p) \in \mathcal{P}_p^\varepsilon(K)$, define its average loss and the packing benchmark by

$$F_f(P) := \frac{1}{p}\sum_{i=1}^{p} f(u_i), \qquad v_\varepsilon(K) := \inf_{P\in\mathcal{P}_p^\varepsilon(K)} F_f(P).$$

It's direct to see that the average loss is independent of the ordering. We assume $\mathcal{P}_p^\varepsilon(\mathcal{X}) \neq \varnothing$. Since $\mathcal{X}$ is compact and the separation constraints are closed, $\mathcal{P}_p^\varepsilon(\mathcal{X})$ is compact and the infimum is attained. The same holds for any compact K with $\mathcal{P}_p^\varepsilon(K) \neq \varnothing$.

Every block of p consecutive admissible actions forms a packing and therefore has average loss at least $v_\varepsilon(\mathcal{X})$. Conversely, from an empty history, repeating any fixed ordering of an optimal packing gives an admissible period-p policy with long-run average loss $v_\varepsilon(\mathcal{X})$. Thus, for stationary losses, this benchmark is the best long-run average loss achievable with full knowledge of f. Following the use of feasible periodic comparators in blocking bandits (Cesa-Bianchi et al., 2026), we define the expected pseudo-regret of a policy π as

$$R_T^\varepsilon(\pi, f) := \mathbb{E}_f^\pi\left[\sum_{t=1}^{T} f(x_t)\right] - Tv_\varepsilon(\mathcal{X}). \tag{1}$$



(a) Admissible pair (b) Inadmissible pair (c) ε-packing

Figure 1: Geometry of blocking. The dashed circle around a point marks the boundary of its open radius-ε blocked neighborhood. In (a), u_j is at least ε from u_i and is not blocked by it; in (b), it lies inside that neighborhood and is blocked. In (c), every pair of marked points is at least ε apart, so the points form a feasible packing set. Their blocked neighborhoods may overlap.

Although the benchmark optimizes over several mutually separated points, convexity allows us to localize the search around a single low-loss point.

Proposition 2.1 (Localization of the optimal packing)**.** *For any $z \in \mathcal{X}$, let $K_z := \mathcal{X} \cap \overline{B}(z, 2p\varepsilon)$. Then*

$$0 \le v_\varepsilon(K_z) - v_\varepsilon(\mathcal{X}) \le f(z) - f^\star. \tag{2}$$

In particular, every $x^\star \in \arg\min_{\mathcal{X}} f$ satisfies $v_\varepsilon(\mathcal{X} \cap \overline{B}(x^\star, 2p\varepsilon)) = v_\varepsilon(\mathcal{X})$.

The proof uses the packing-compression argument in Lemma G.7: points of a feasible packing are moved toward z along line segments while preserving feasibility, and convexity controls the increase in loss. Thus an anchor satisfying $f(z) - f^\star \le \delta$ restricts the packing search to K_z at a cost of at most δ in the benchmark value.

3 General Period: Upper Bounds

Proposition 2.1 shows that, for $K_z := \mathcal{X} \cap \overline{B}(z, 2p\varepsilon)$, the best packing in K_z has average loss at most $f(z) - f^\star$ above the global benchmark. This suggests a two-phase policy: first learn a low-loss *anchor* z by stochastic convex-bandit optimization, then learn a feasible packing inside K_z. Both phases must be implemented with admissible physical actions: each anchor query is embedded in a feasible cycle, and switching target packing sets may require transfers. We first establish the transfer results under two geometric conditions, then present the policy and its regret bound.

3.1 Feasible queries and transfers

Let $S \in \mathcal{P}_p^\varepsilon(\mathcal{X})$ be the current *chronological state* (the last p actions, oldest first). A transfer to a target packing $P \in \mathcal{P}_p^\varepsilon(\mathcal{X})$ ends at any state Q with $B(Q) = B(P)$. Repeating Q gives an admissible *complete cycle* of p physical rounds with total loss $pF_f(P)$. For a fresh cycle, stationarity and i.i.d. noise imply that, conditional on the preceding history, the joint law of observations indexed by queried point is independent of arrival order.

Let $H(S, B(P))$ be the minimum number of additional admissible rounds needed for such a transfer, with value ∞ if none is possible. The worst-case transition time is

$$H(\mathcal{X}; p, \varepsilon) := \sup_{S,P \in \mathcal{P}_p^\varepsilon(\mathcal{X})} H(S, B(P)).$$

Lemma 3.1 (Local reconfiguration under an inball condition)**.** *Assume*

$$\overline{B}(c, r_0) \subseteq \mathcal{X}, \qquad 0 < \varepsilon \le \frac{r_0}{2(2p)^{1/d}}. \tag{3}$$

Any chronological state $S \in \mathcal{P}_p^\varepsilon(\mathcal{X})$ can reach any prescribed ordered target $P \in \mathcal{P}_p^\varepsilon(\mathcal{X})$ within $2p-1$ additional physical rounds, including the p target plays. The worst-case transition time satisfies

$$p \le H(\mathcal{X}; p, \varepsilon) \le 2p - 1.$$

Appendix E proves the lemma. Lemmas E.1 and G.8 also let us fix, for each query $q \in \mathcal{X}$, a completion $P(q) = (q, u_2(q), \dots, u_p(q)) \in \mathcal{P}_p^\varepsilon(K_q)$. Its auxiliary points satisfy $f(u_j(q)) - f(q) \le a_\varepsilon := \min\{2Lp\varepsilon, 1\}$ for $2 \le j \le p$.

Diameter condition. A second sufficient condition uses a long diameter line segment. When $D \ge p\varepsilon$, it contains at least $p+1$ separated parking points that support transfers between packings.

Lemma 3.2 (Diameter-based reconfiguration)**.** *Let $\mathcal{X}$ be compact and convex, not necessarily full-dimensional, with diameter $D \ge p\varepsilon$. For any $S, P \in \mathcal{P}_p^\varepsilon(\mathcal{X})$, an admissible transfer from S reaches some state Q with $B(Q) = B(P)$. The worst-case transition time satisfies*

$$H(\mathcal{X}; p, \varepsilon) \le 3p + \frac{4p^2}{D/\varepsilon - p + 1}. \tag{4}$$

Appendix D proves (4) via parking-point transfers, with worst-case tight dependence on $D/\varepsilon - p + 1$. The parking points and Lemma G.8 yield a local completion $P(q)$ with the same auxiliary-loss bound for each query. With transfer budget $H_D := \lceil 3p + 4p^2/(D/\varepsilon - p + 1) \rceil$, the policy enters $B(P(q))$ and cycles its arrival order Q, which may differ from $P(q)$; Appendix G gives the regret bound.

Neither geometric condition implies the other. For $d = 2$, $p = 50$, and $\varepsilon = 1/20$, the unit disk satisfies the inball condition of Lemma 3.1 but has $D = 2 < 5/2 = p\varepsilon$. Conversely, the thin rectangle $[0,3] \times [0, 1/100]$ satisfies $D \ge p\varepsilon$, but its largest inball has radius $1/200$, below the radius 1 required by that condition.

3.2 Anchor-and-refine learning

Anchor phase. We first learn an anchor with low loss to localize the subsequent packing search. To this end, we run the batched BCO procedure of Lemma F.1 on $\mathcal{X}$ for n virtual queries $q_1, \dots, q_n$, with n chosen below and failure probability T^{-2}. With high probability, the cumulative excess loss is at most $\mathcal{A}_T\sqrt{n}$, and they form at most $\mathcal{J}_T$ *query segments* (maximal consecutive runs of identical queries), where

$$\mathcal{A}_T := \max\{1, A_d(1+\sigma)[\log(eT^3)]^4\}, \qquad \mathcal{J}_T := \max\{1, \lceil B_d[\log(eT^3)]^3 \rceil\}.$$

Here A_d, B_d are the constants in Lemma F.1. Set $H := 2p - 1$. Each segment at q starts with one transfer to $B(P(q))$, followed by repetitions of the arrival state Q (with $Q = P(q)$ under the inball construction). Each fresh cycle supplies only its observation at q to BCO and has loss at most $pf(q) + b_{\mathrm{anc}}$, where $b_{\mathrm{anc}} := (p-1)a_\varepsilon$. The initial transfer plays $P(q)$ once in $p \le H$ rounds; all transfer observations are discarded. If BCO requests a $(\mathcal{J}_T + 1)$st segment, cycle the current Q until T. On success this fallback never occurs: the anchor phase uses at most $pn + H\mathcal{J}_T$ rounds and incurs regret at most $p\mathcal{A}_T\sqrt{n} + b_{\mathrm{anc}}n + H\mathcal{J}_T$, since $v_\varepsilon(\mathcal{X}) \ge f^\star$. Set $z := n^{-1}\sum_{s=1}^n q_s$. Convexity and Proposition 2.1 give

$$v_\varepsilon(K_z) - v_\varepsilon(\mathcal{X}) \le f(z) - f^\star \le \frac{1}{n}\sum_{s=1}^n (f(q_s) - f^\star) \le \frac{\mathcal{A}_T}{\sqrt{n}}.$$

Thus localization costs at most $\mathcal{A}_T T/\sqrt{n}$ in the remaining rounds. The cap $n \le T/(2p)$ already bounds the query-regret term by $O(\mathcal{A}_T\sqrt{pT})$; we balance $b_{\mathrm{anc}}n$ against $\mathcal{A}_T T/\sqrt{n}$ subject to this cap. When $T > 4H\mathcal{J}_T$, choose

$$n := \begin{cases} \lfloor T/(2p) \rfloor, & b_{\mathrm{anc}} = 0, \\ \left\lfloor \min\{T/(2p), (\mathcal{A}_T T/b_{\mathrm{anc}})^{2/3}\} \right\rfloor, & b_{\mathrm{anc}} > 0. \end{cases} \tag{5}$$

Then $pn + H\mathcal{J}_T < 3T/4$, leaving time for refinement. For shorter horizons, cycle any fixed feasible packing. The BCO failure event contributes at most T^{-1} to expected regret; Appendix G gives the full accounting. All operations stop at round T, and only complete refinement cycles update the cell statistics.

Local refinement. Condition on the completed anchor history and set $R := 2p\varepsilon$, $\beta := \min\{2LR, 1\}$. The fixed region K_z contains a feasible packing by Proposition 2.1; since $\mathrm{diam}(K_z) \le 2R$, its loss oscillation is at most β. The first entry may leave K_z and costs at

Algorithm 1 Anchor-and-refine policy

Require: Horizon T and the quantities defined in Section 3.2.
Stop all actions at round T; discard transfer observations.
Set $H \leftarrow 2p-1$.
If $T \le 4H\mathcal{J}_T$, cycle any $P_0 \in \mathcal{P}_p^\varepsilon(\mathcal{X})$ to T and stop.
Anchor phase
Set n by (5); start batched BCO with failure probability T^{-2}.
Set $j \leftarrow 0$. ▷ Number of query segments started
for $s = 1, \dots, n$ **do**
  Obtain the next virtual query q_s from BCO.
  if $s = 1$ or $q_s \neq q_{s-1}$ **then** ▷ A new query segment
    If $j = \mathcal{J}_T$, cycle the current Q to T and stop.
    Set $j \leftarrow j+1$ and enter the prescribed state $Q = P(q_s)$.
  end if
  Play one fresh cycle of Q; send only the observation at q_s to BCO.
end for
Refinement phase
Set $z \leftarrow n^{-1}\sum_{s=1}^n q_s$, $R \leftarrow 2p\varepsilon$, and $K_z \leftarrow \mathcal{X} \cap \overline{B}(z, R)$.
Choose the cell radius r by Lemma G.5.
If $L = 0$ or $r = 6R$, choose $P \in \mathcal{P}_p^\varepsilon(K_z)$, enter $B(P)$, record its arrival order Q, cycle Q to T, and stop.
Partition K_z into $C_1, \dots, C_M$; initialize $n_i = \widehat{f}_i = 0$.
while physical rounds remain **do**
  Form $\underline{f}_i$ and choose $P \in \mathcal{P}_p^\varepsilon(K_z)$ with $\underline{F}(P) \le \inf_{P' \in \mathcal{P}_p^\varepsilon(K_z)} \underline{F}(P') + T^{-2}$.
  Set $m_i \leftarrow m_i(P)$ and freeze the batch length $\kappa \leftarrow \max\{1, \min_{i:m_i>0} \lfloor n_i/m_i \rfloor\}$.
  Transfer to $B(P)$ and record its arrival order Q (first entry may leave K_z; later transfers stay inside K_z).
  If time remains, play up to κ cycles of Q and update $n_i, \widehat{f}_i$ from complete cycles.
end while

most $H_{\mathrm{in}} := H$ in rounds and regret. Later transfers stay in K_z, take at most $\bar{H} := 3p$ rounds, and incur excess loss at most β per round above $v_\varepsilon(K_z)$ by Lemma G.2. We bound the remaining horizon by T. For $r = 6R$ or $L = 0$, enter any feasible packing in K_z and repeat its arrival order. Otherwise choose $0 < r < 6R$, partition K_z into cells $C_1, \dots, C_M$ of diameter at most r, with $M \le (6R/r)^d$ (Lemma G.1), and set $\Lambda := \log(2(M+1)T^3)$. Cells pool observations; actions remain in the continuous set K_z. Let $\iota(x)$ index the cell containing x, and let n_i, $\widehat{f}_i$ be the observation count and empirical mean in C_i, using only complete target cycles from this phase. Define

$$\ell_i := \begin{cases} [\widehat{f}_i - Lr - \sigma\sqrt{2\Lambda/n_i}]_{[0,1]}, & n_i > 0, \\ 0, & n_i = 0, \end{cases}$$

where $[a]_{[0,1]} := \min\{1, \max\{0, a\}\}$. Set $N := \sum_i n_i$. For $N > 0$, let $\widehat{f}_{K_z} := N^{-1}\sum_i n_i \widehat{f}_i$ and $g := [\widehat{f}_{K_z} - \beta - \sigma\sqrt{2\Lambda/N}]_{[0,1]}$; otherwise set $g := 0$. Use $\underline{f}_i := \max\{\ell_i, g\}$. Both indices are lower bounds on the confidence event; g controls confidence widths in sparsely sampled cells through β and the pooled sample size N.

For $P = (u_1, \dots, u_p)$, define $\underline{F}(P) := p^{-1}\sum_j \underline{f}_{\iota(u_j)}$ and $m_i(P) := \sum_j \mathbf{1}\{u_j \in C_i\}$, both invariant under reordering. Each *refinement batch* selects P, freezes κ as in Algorithm 1, transfers to $B(P)$, and repeats its arrival order Q for up to κ cycles. Batches may reselect the same packing. Each completed batch makes one cell count positive for the first time or multiplies it by at least $3/2$, giving $O(M\log(eT))$ batches and transfers.

We assume access to an optimization oracle that minimizes the current LCB objective $\underline{F}$ over $\mathcal{P}_p^\varepsilon(K_z)$ up to additive error T^{-2}. This error contributes at most T^{-1} to total regret; see the proof of Lemma G.4 in Appendix G. Our guarantees bound regret over physical interaction rounds and do not bound the oracle's computation time. The geometric optimization handled by this oracle can be computationally difficult. Even for $p = 2$, the packing feasibility threshold is $\mathrm{diam}(\mathcal{X})$, whose

computation is NP-hard for polytopes specified by rational linear inequalities when the dimension is part of the input Gritzmann & Klee (1993). Certifying optimal equal-circle packings in a square has also required specialized global optimization Markót (2021). Beyond feasibility, Theorem C.1 in Appendix C shows that approximating the offline packing benchmark to inverse-polynomial additive accuracy is NP-hard when the dimension varies, even on the unit ball with $p = 2$ and a supplied minimizer of f.

Cell approximation, estimation, and sparse-cell exploration plus transfers contribute LrT, $\sigma\sqrt{\Lambda MT}$, and $\beta\bar{H}M\log(eT)$, respectively. Balancing the first against the other two gives the radius in Lemma G.5; the statistical term is $(LR)^{d/(d+2)}T^{(d+1)/(d+2)}$ up to noise and logarithmic factors. The tuned sparse-cell and transfer cost is absorbed in the final bound (6), using its statistical term or the cap $R_T^{\varepsilon} \le T$; see Proposition G.6 and Appendix G.

Combined regret bound. All auxiliary and transfer actions are ordinary physical rounds included in regret. Only designated target-cycle observations are returned to the virtual BCO routine, and the packing learner starts with fresh statistics. The query-segment cap and short-horizon branch make the physical-time budget deterministic; Appendix G gives the proof. The anchor guarantee, localization inequality, and localized packing bound now yield the main result.

Theorem 3.3 (General-period upper bound)**.** *Assume* (3)*. There exists an admissible policy such that, for every $T \ge p$,*

$$R_T^{\varepsilon} \le \min\left\{T,\ \widetilde{O}_{d,\sigma}\left(\sqrt{pT}\left[1+\left(L\varepsilon\sqrt{pT}\right)^{\frac{d}{d+2}}\right]\right)\right\}. \tag{6}$$

The notation suppresses additional polynomial factors in d, noise-dependent constants, and logarithmic factors.

The $\sqrt{pT}$ prefactor reflects the cost of obtaining a low-loss anchor through feasible cycles. The ε-dependent part inside the brackets covers its auxiliary actions and statistical estimation among localized packings. Sparse-cell exploration and transfers are included through the absorption argument in Appendix G. For fixed p, d, L, σ, the displayed bound is $\widetilde{O}\Big(\sqrt{T}[1+(\varepsilon\sqrt{T})^{d/(d+2)}]\Big)$, the rate stated in the introduction.

4 General Period: Statistical Lower Bound

We next ask how much of the general-period upper rate is intrinsic to learning the packing itself. We show that a nontrivial gap remains even on the unit Euclidean ball, where reconfiguration can be completed in only $O(p)$ rounds. Thus the difficulty is statistical rather than a consequence of expensive transitions.

Let $\mathcal{X} = B_2^d$ and let the observation noises be independent $N(0,1)$. Write $\mathcal{F}$ for the convex, 1-Lipschitz functions $f : \mathcal{X} \to [0,1]$. To state the bound uniformly in d and p, define

$$\chi_{d,p} := \max\{1, \frac{(4p)^{2/(d-1)}}{4}\}. \tag{7}$$

This factor ensures that the perturbation construction has at least $4p$ candidate directions in the regime where its blocking-dependent term is needed.

Theorem 4.1 (General-period lower bound)**.** *For every $d \ge 2$, $p \ge 2$, $0 < \varepsilon \le 1/[2(2p)^{1/d}]$, and integer $T \ge 1$,*

$$\begin{aligned}\inf_{\pi}\sup_{f\in\mathcal{F}} R_T^{\varepsilon}(\pi, f) &\ge \sqrt{T}\left[\frac{\min\{\sqrt{T}, d\}}{64} + \frac{(\varepsilon\sqrt{T})^{\frac{d-1}{d+3}}}{24576p^{1+1/d}\chi_{d,p}}\right] \\ &\ge \frac{\sqrt{T}}{24576p^{1+1/d}\chi_{d,p}}\left[1+(\varepsilon\sqrt{T})^{\frac{d-1}{d+3}}\right].\end{aligned} \tag{8}$$

Moreover, $1 \le \chi_{d,p} \le 4p^2$, and $\chi_{d,p} = 1$ whenever $d - 1 \ge \log_2(4p)$. The hard instances admit known bridges of at most $p-1$ intermediate actions before a prescribed p-round target cycle.

Here is the mechanism behind the blocking-dependent exponent. Start with a radial loss that is quadratic near the origin and linear outside; its optimal p-packing can be rotated without changing its loss. A downward perturbation of depth Δ around one of M separated directions makes the corresponding rotated packing better while preserving convexity. Its support has radius of order $\sqrt{\varepsilon\Delta}$ and diameter less than ε. For fixed d, p, the number of candidate directions and the information needed to distinguish them satisfy, up to constants,

$$M \gtrsim_{d,p} (\varepsilon/\Delta)^{(d-1)/2}, \qquad \Delta^2 T \lesssim M.$$

An admissible policy can visit any one support at most $\lceil T/p \rceil$ times. A Gaussian change-of-measure argument then gives regret of order $T\Delta/p$ when the improved direction remains unidentified. Balancing the displayed relations yields $\Delta \asymp_{d,p} \varepsilon^{(d-1)/(d+3)} T^{-2/(d+3)}$ and the second term of Theorem 4.1. The affine family supplies $\Omega(\min\{T, d\sqrt{T}\})$ separately; taking the larger of the two bounds gives their sum up to constants. Appendix H handles the short-horizon cases and the precise dependence on p. Using $\chi_{d,p} \le 4p^2$, we obtain the simpler consequence

$$\inf_{\pi} \sup_{f \in \mathcal{F}} R_T^{\varepsilon}(\pi, f) \ge \frac{\sqrt{T}}{98304 p^{3+1/d}} \left[1 + (\varepsilon\sqrt{T})^{\frac{d-1}{d+3}}\right]. \tag{9}$$

This is the form used in the introduction, while Theorem 4.1 retains the sharper dependence on p through $\chi_{d,p}$.

Remark 4.2. On B_2^d and for $T \ge p$, Theorems 3.3 and 4.1 apply under the same inball condition. For general p, their blocking-dependent terms leave gaps in the powers of ε and T, and their dependence on p also differs. Section 5 closes the ε and T gaps for $p = 2$, up to logarithmic factors.

5 Period Two: Matching Minimax Bounds

The general-period bounds above do not match. For $p = 2$, each feasible pair (x, y) has a separation direction $(x - y)/\|x - y\|$ on the unit sphere $S^{d-1} := \{v \in \mathbb{R}^d : \|v\| = 1\}$. Covering these directions lets us work with convex branches instead of discretizing the full pair space. This reduction improves the statistical rate and yields a matching minimax bound. Fix $0 < \theta \le 1/2$. A Euclidean θ-net $\mathcal{U}_\theta \subset S^{d-1}$ is a finite set such that every unit direction lies within distance θ of some $u \in \mathcal{U}_\theta$. With $C_{\mathrm{dir}}(d) := 2d(1 + 4\sqrt{d})^{d-1}$, we may choose such a net with $N := |\mathcal{U}_\theta| \le C_{\mathrm{dir}}(d)\theta^{-(d-1)}$.

For each $u \in \mathcal{U}_\theta$, define the branch $K_u := \{(x, y) \in \mathcal{X}^2 : \langle u, x - y\rangle \ge \varepsilon(1 + \theta^2)\}$. Each K_u is convex, and every pair in K_u is strictly feasible. The factor $1 + \theta^2$ provides enough slack to approximate an arbitrary feasible pair after discretizing its direction.

Lemma 5.1 (Quadratic directional approximation)**.** *Assume* $0 < \varepsilon < r_0$. *Then*

$$0 \le \min_{u \in \mathcal{U}_\theta} \min_{(x,y) \in K_u} \frac{f(x) + f(y)}{2} - v_\varepsilon(\mathcal{X}) \le \frac{2\varepsilon}{r_0}\theta^2.$$

Lemma 5.1 reduces learning to N convex branches with per-round approximation cost $O((\varepsilon/r_0)\theta^2)$. To query $P = (x, y) \in K_u$, transfer to $B(P) = \{x, y\}$ and play a fresh cycle in the arrival order Q. Although Q may lie outside K_u, the averaged feedback has mean $F_f(Q) = F_f(P)$ and $\sigma/\sqrt{2}$-sub-Gaussian noise. Algorithm 2 runs batched BCO (Lemma F.1) on each selected branch, using its latest completed run to form a lower-confidence index. Each query segment (a maximal block of identical queries within one run) shares one transfer; every new run starts a new segment. For the confidence bounds, set the total failure budget $\delta := (T + 1)^{-2}$, the failure budget per BCO run $\xi := \delta/[4N(T + 1)]$, and the logarithmic factor $\Lambda := \log(64N(T + 1)^4/\delta)$. Let A_{2d} be the regret constant in Lemma F.1 for a domain of affine dimension at most $2d$. For a run of $n \ge 1$ virtual queries, define the cumulative regret bound $B(n) := A_{2d}(1 + \sigma/\sqrt{2})\Lambda^4\sqrt{n}$ and the confidence radius for its average noise $r(n) := (\sigma/\sqrt{2})\sqrt{2\Lambda/n}$.

With empty history, the first transfer plays P_1 once to establish $Q = P_1$; these two rounds are charged as transfer rounds and their observations are ignored. A query segment of a virtual queries uses a complete cycles, hence $2a$ target-play rounds, plus one transfer. A completed m-query run uses m cycles, hence $2m$ target-play rounds, plus its query-segment transfers. Physical play stops at round T: incomplete cycles give no feedback, and incomplete runs do not update I_u or k_u.

Algorithm 2 Period-two convex-branch master

Require: Horizon T, branch family $\{K_u\}_{u\in\mathcal{U}_\theta}$, and confidence quantities defined in Section 5.

for all $u \in \mathcal{U}_\theta$ **do**
  Initialize the completed-run count $k_u \leftarrow 0$ and lower-confidence index $I_u \leftarrow -\infty$.
end for
while physical rounds remain **do**
  Choose $u \in \arg\min_{v\in\mathcal{U}_\theta} I_v$; set $m \leftarrow 2^{k_u}$ virtual pair queries for this run.
  Start a new batched BCO instance on K_u with planned horizon m and failure level ξ.
  for $j = 1, \dots, m$ **do**
    Obtain the virtual query $P_j = (x_j, y_j) \in K_u$ from BCO.
    if $j = 1$ or $P_j \neq P_{j-1}$ **then**
      Transfer admissibly to $B(P_j) = \{x_j, y_j\}$ and record its chronological arrival order Q; ignore transfer observations.
    end if
    Play one fresh cycle in order Q over two consecutive rounds; observe $Y_{j,1}, Y_{j,2}$.
    Return the average feedback $Z_j \leftarrow (Y_{j,1} + Y_{j,2})/2$ to BCO.
  end for
  if all m cycles were completed **then**
    Set the run mean $\bar{Z}_u \leftarrow m^{-1}\sum_{j=1}^{m} Z_j$ and update $I_u \leftarrow \bar{Z}_u - B(m)/m - r(m)$.
    Set $k_u \leftarrow k_u + 1$.
  end if
end while

The next theorem quantifies the regret of Algorithm 2. Its first term is the directional approximation error from Lemma 5.1, while the remaining terms are the statistical cost of learning over the branch family and the overhead of the batched implementation.

Theorem 5.2 (Period-two upper bound)**.** *Let $p = 2$ and $0 < \varepsilon < r_0$. There exists a constant $C_{d,\sigma} > 0$, depending only on d and σ, such that, for every $0 < \theta \le 1/2$, Algorithm 2 satisfies*

$$R_T^{\varepsilon} \le \frac{2\varepsilon}{r_0}\theta^2 T + C_{d,\sigma}\log^4\left(\frac{\mathrm{e}T}{\theta}\right)\left(\sqrt{T\theta^{-(d-1)}} + \theta^{-(d-1)}\right) + 5. \tag{10}$$

Choosing $\theta_T := \min\{1/2, ((\varepsilon/r_0)^2 T)^{-1/(d+3)}\}$ gives

$$R_T^{\varepsilon} = \widetilde{O}_{d,\sigma}\left(\sqrt{T} + \left(\frac{\varepsilon}{r_0}\right)^{\frac{d-1}{d+3}} T^{\frac{d+1}{d+3}}\right). \tag{11}$$

The rate follows by balancing the approximation term $(\varepsilon/r_0)\theta^2 T$ with the dominant statistical term $\sqrt{T\theta^{-(d-1)}}$. The improvement over the general-period bound comes from discretizing only the $(d-1)$-dimensional direction space.

On the unit Euclidean ball, the general lower bound has the same dependence on T and ε when $p = 2$. Together with Theorem 5.2, it gives the following minimax rate.

Corollary 5.3 (Minimax rate for period two)**.** *Let $\mathcal{X} = B_2^d := \{x \in \mathbb{R}^d : \|x\| \le 1\}$, $d \ge 2$, and $p = 2$. Assume independent standard Gaussian observation noise. Let $\mathcal{F}$ be the class of convex, 1-Lipschitz functions $f : \mathcal{X} \to [0, 1]$, and let π range over admissible policies. There are constants $0 < c_d \le C_d < \infty$, depending only on d, such that, for every $0 < \varepsilon \le 1/(2 \cdot 4^{1/d})$ and integer $T \ge 1$,*

$$c_d\sqrt{T}\left[1 + (\varepsilon\sqrt{T})^{\frac{d-1}{d+3}}\right] \le \inf_{\pi}\sup_{f\in\mathcal{F}} R_T^{\varepsilon}(\pi, f) \le C_d\log^4(\mathrm{e}T)\sqrt{T}\left[1 + (\varepsilon\sqrt{T})^{\frac{d-1}{d+3}}\right].$$

Proof. The unit ball has inradius $r_0 = 1$, so Theorem 5.2 with $\sigma = 1$ gives the upper bound. Indeed, its choice of θ_T satisfies $\theta_T \ge T^{-1/(d+3)}/2$, so the logarithmic factor in (10) is $O_d(\log(\mathrm{e}T))$. For the lower bound, apply the second inequality in Theorem 4.1 with $p = 2$. □

AI USE STATEMENT

Generative AI tools (OpenAI Codex) were used to assist with manuscript review. These tools were also used for literature lookup and LaTeX editing. The authors take full responsibility for the correctness, originality, and integrity of the final manuscript, including all AI-assisted content.

ETHICS STATEMENT

This paper presents theoretical results on stochastic bandit convex optimization with blocking constraints. It does not report experiments involving human participants or animals, or use personal or sensitive data.

REPRODUCIBILITY STATEMENT

The problem formulation, observation model, and assumptions are specified in Section 2 and in the individual theorem statements. Algorithms 1 and 2 describe the learning policies. Appendices D–H provide the geometric constructions, the batched convex-bandit subroutine, and the proofs of the regret bounds, including the accounting for auxiliary actions and transitions. The packing-selection oracle and its required accuracy are specified in Section 3.2.

REFERENCES

Alekh Agarwal, Dean P. Foster, Daniel Hsu, Sham M. Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

Jason M. Altschuler and Kunal Talwar. Online learning over a finite action set with limited switching. *Mathematics of Operations Research*, 46(1):179–203, 2021.

Alexia Atsidakou, Orestis Papadigenopoulos, Soumya Basu, Constantine Caramanis, and Sanjay Shakkottai. Combinatorial blocking bandits with stochastic delays. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 404–413, 2021.

Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. Blocking bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Soumya Basu, Orestis Papadigenopoulos, Constantine Caramanis, and Sanjay Shakkottai. Contextual blocking bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 271–279, 2021.

Matteo Berni, Paolo Erani, Nicola Francesco Lopomo, and Massimiliano Baleani. Optimization of in situ indentation protocol to map the mechanical properties of articular cartilage. *Materials*, 15(18): 6425, 2022. doi: 10.3390/ma15186425.

Nicholas Bishop, Hau Chan, Debmalya Mandal, and Long Tran-Thanh. Adversarial blocking bandits. In *Advances in Neural Information Processing Systems*, 2020.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011a.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pp. 144–158. Springer, 2011b.

Leonardo Cella and Nicoló Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 1168–1177, 2020.

Nicolò Cesa-Bianchi, Junya Honda, Yuko Kuroki, Atsushi Miyauchi, and Lukas Zierahn. Learning periodic strategies in blocking bandits is as hard as bandits with switching costs. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1001–1021, 2026.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 151–159, 2013.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2014.

Ayoub Foussoul, Vineet Goyal, Orestis Papadigenopoulos, and Assaf Zeevi. Last switch dependent bandits with monotone payoff functions. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10265–10284, 2023.

Peter Gritzmann and Victor Klee. Computational complexity of inner and outer j-radii of polytopes in finite-dimensional normed spaces. *Mathematical Programming*, 59(1–3):163–213, 1993. doi: 10.1007/BF01581243.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Bandit task assignment with unknown processing time. In *Advances in Neural Information Processing Systems*, 2023.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80(2–3):245–272, 2010.

Tomer Koren, Roi Livni, and Yishay Mansour. Multi-armed bandits with metric movement costs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Pierre Laforgue, Giulia Clerici, Nicolò Cesa-Bianchi, and Ran Gilad-Bachrach. A last switch dependent analysis of satiation and seasonality in bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 971–990, 2022.

Tor Lattimore and András György. Improved regret for zeroth-order stochastic convex bandits. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2938–2964, 2021.

Tor Lattimore and András György. A second-order method for stochastic bandit convex optimisation. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2067–2094, 2023.

Mihály Csaba Markót. Improved interval methods for solving circle packing problems in the unit square. *Journal of Global Optimization*, 81:773–803, 2021. doi: 10.1007/s10898-021-01086-z.

Gergely Neu and Michal Valko. Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, pp. 2780–2788, 2014.

Orestis Papadigenopoulos and Constantine Caramanis. Recurrent submodular welfare and matroid blocking semi-bandits. In *Advances in Neural Information Processing Systems*, pp. 23334–23346, 2021.

Jakub Roemer, Hassan Khawaja, Mojtaba Moatamedi, and Lukasz Pieczonka. Data processing scheme for laser spot thermography applied for nondestructive testing of composite laminates. *Journal of Nondestructive Evaluation*, 42:21, 2023. doi: 10.1007/s10921-023-00932-2.

Aadirupa Saha, Pierre Gaillard, and Michal Valko. Improved sleeping bandits with stochastic action sets and adversarial rewards. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8357–8366, 2020.

David Simchi-Levi and Yunzong Xu. Phase transitions in bandits with switching constraints. *Management Science*, 69(12):7182–7201, 2023.

A Additional Related Work

Availability-constrained bandits. In sleeping bandits (Kleinberg et al., 2010; Saha et al., 2020) and in the combinatorial setting of Neu & Valko (2014), the set of available actions changes over time. The environment decides this set, which is independent of the learner's past actions. In bandit task assignment (Ito et al., 2023), ongoing tasks restrict which new tasks can be started. Here availability depends on the learner, but the tasks form a finite set. In delay-dependent bandits (Cella & Cesa-Bianchi, 2020), the expected reward of an arm depends on the time since it was last played. Last-switch-dependent bandits (Laforgue et al., 2022; Foussoul et al., 2023) use the time since the arm last took part in a switch. In both cases, the history changes the rewards, and every arm can always be played. In our model, by contrast, the loss is fixed, and the history decides which actions are allowed.

Bandits with switching and movement costs. With a unit cost per switch, the minimax regret of adversarial bandits is of order $T^{2/3}$ up to logarithmic factors (Dekel et al., 2014). Other works limit the number or the total cost of switches and studies the resulting regret-switching tradeoffs Altschuler & Talwar (2021); Simchi-Levi & Xu (2023). Metric movement costs instead charge according to the distance between consecutive actions Koren et al. (2017). In these models, repeating the same action is free. Instead, the recent trajectory may make a desired action infeasible, so changing packings require intermediate admissible plays. Our transition bounds quantify the number of such plays and include their losses in the regret.

B Fixed-Action and Greedy Comparators

Best fixed action. The classical fixed-action comparator has value $f^\star = \min_{x\in\mathcal{X}} f(x)$ and expected pseudo-regret

$$R_T^{\mathrm{fixed}}(\pi, f) := \mathbb{E}_f^\pi\left[\sum_{t=1}^T f(x_t)\right] - Tf^\star = R_T^\varepsilon(\pi, f) + T\big(v_\varepsilon(\mathcal{X}) - f^\star\big). \tag{12}$$

The difference is a policy-independent blocking penalty. It is nonnegative and, by Proposition 2.1 and Lipschitzness, at most $2Lp\varepsilon T$, but can grow linearly with T even when f is known. For example, let $\mathcal{X} = [0, 1]$, $f(x) = x$, and $0 < \varepsilon \le 1/(p-1)$. Sorting any feasible p-packing shows that its ith point is at least $(i-1)\varepsilon$ for $i \ge 2$. The packing $(0, \varepsilon, \dots, (p-1)\varepsilon)$ therefore attains $v_\varepsilon(\mathcal{X}) = (p-1)\varepsilon/2$, whereas $f^\star = 0$. Every complete block of p admissible actions has loss at least $p(p-1)\varepsilon/2$, so every policy satisfies

$$R_T^{\mathrm{fixed}}(\pi, f) \ge \lfloor T/p \rfloor \, \frac{p(p-1)\varepsilon}{2} = \Omega(T)$$

for fixed p and ε.

Myopic greedy comparator. In stochastic finite-arm blocking, the oracle greedy scheduler of Basu et al. (2019) selects the currently available arm with the best known mean reward. Given its own admissible history $g_1, \dots, g_{t-1}$, its loss-minimizing analogue here would choose a point in

$$\arg\min_{x\in A_t(g_{1:t-1})} f(x), \quad A_t(g_{1:t-1}) := \{x \in \mathcal{X} : \|x - g_{t-s}\| \ge \varepsilon \text{ for } 1 \le s \le \min\{p-1, t-1\}\}.$$

For a specified feasible greedy trajectory $g_{1:T}$, one could define $G_T^{\mathrm{gr}} := \sum_{t=1}^T f(g_t)$ and $R_T^{\mathrm{gr}}(\pi, f) := \mathbb{E}_f^\pi[\sum_{t=1}^T f(x_t)] - G_T^{\mathrm{gr}}$. Then $R_T^{\mathrm{gr}} = R_T^\varepsilon - (G_T^{\mathrm{gr}} - Tv_\varepsilon(\mathcal{X}))$. Each nonempty A_t is compact, so its one-step minimum is attained. Nevertheless, a myopic choice can eliminate every next action: with $\mathcal{X} = [-1, 1]$, $p = 2$, $\varepsilon = 3/2$, and $f(x) = x^2$, the first greedy action is 0, after which A_2 is empty, although $(-1, 1)$ is a feasible packing. Thus the greedy prescription need not define a trajectory for the full horizon.

Even when an exact greedy trajectory is feasible, its long-run loss may exceed the periodic benchmark. Take the same $\mathcal{X}$ and f, but $0 < \varepsilon < 1$. Breaking the tie at every even round in favor of $+\varepsilon$ gives the greedy cycle $(0, \varepsilon)$, with average loss $\varepsilon^2/2$. By contrast,

$$v_\varepsilon(\mathcal{X}) = \frac{\varepsilon^2}{4},$$

as follows from $\big(u^2+v^2\big)/2 \geq (u-v)^2/4$ and the feasible pair $(-\varepsilon/2, \varepsilon/2)$. Hence this greedy cycle has a positive, linear-in-T loss gap relative to $v_\varepsilon(\mathcal{X})$. A greedy-regret guarantee would therefore depend on tie-breaking and continuation rules, and would not in general imply a regret guarantee against the optimal periodic packing.

C Computational Hardness of the Offline Benchmark

Theorem C.1 (Hardness of offline packing)**.** *Let B_2^d be the Euclidean unit ball. When d is part of the input, computing $v_\varepsilon(B_2^d)$ to additive accuracy $1/(10^4 d^2)$ is NP-hard, even for $p = 2$, rational $\varepsilon > 0$, and a convex piecewise-linear 1-Lipschitz loss $f : B_2^d \to [0,1]$ given as the maximum of $O(d)$ rational affine functions. The origin can be given as a global minimizer of f.*

Proof. We reduce from Max-Cut. Let $G = (V, E)$ have n vertices and $m \geq 1$ edges, and let $k \in \{1, \dots, m\}$. Write $C^\star$ for its maximum cut size and orient the edges arbitrarily. The signed incidence matrix $B \in \mathbb{R}^{m \times n}$ then has $(Bz)_e = z_i - z_j$ for an oriented edge $e = (i, j)$. Set $d = n+m$ and $a = 1/(100d)$, and write $x = (z, w) \in \mathbb{R}^n \times \mathbb{R}^m$. Define

$$f(z,w) = \frac{1}{2} \max\left\{0,\ \max_{i \in V}(|z_i| - a),\ \max_{e \in E} |w_e - (Bz)_e|\right\}. \tag{13}$$

This is the maximum of $O(d)$ rational affine functions. Their gradients, after the factor $1/2$, have norms at most $\sqrt{3}/2 < 1$, so f is convex, piecewise linear, and 1-Lipschitz. Since $f(0) = 0$ and $f \geq 0$, the origin is a global minimizer; Lipschitzness also gives $f(x) \leq \sqrt{3}/2 < 1$ on B_2^d.

The zero set is $Q = \{(z, Bz) : z \in [-a, a]^n\}$. It lies in the unit ball because $\|(z, Bz)\|^2 \leq a^2(n + 4m) \leq 4a^2 d < 1$. As Q is centrally symmetric, its diameter is twice its maximum norm. The convex quadratic $\|z\|^2 + \|Bz\|^2$ attains a maximum over the box at a vertex $z = as$, where $s \in \{-1, 1\}^n$. Each edge crossing the cut induced by s contributes $4a^2$ to $\|Bz\|^2$, while each other edge contributes zero. Consequently,

$$\operatorname{diam}(Q) = 2a\sqrt{n + 4C^\star}. \tag{14}$$

Choose a rational ε between $2a\sqrt{n + 4k - 3}$ and $2a\sqrt{n + 4k - 1}$. Such a choice is computable in polynomial time: the interval has width at least $a/\sqrt{d} > 1/(1000d^2)$, so one may take $\varepsilon = \lceil 20d\sqrt{n + 4k - 3}\rceil/(1000d^2)$; its numerator is an integer square root rounded upward. If $C^\star \geq k$, then (14) exceeds ε. A diameter-attaining pair in Q is feasible and has zero loss, giving $v_\varepsilon(B_2^d) = 0$.

Now suppose $C^\star \leq k-1$. By (14), $\operatorname{diam}(Q) \leq 2a\sqrt{n + 4k - 4}$, so rationalization and $n+4k \leq 4d$ give $\varepsilon - \operatorname{diam}(Q) \geq a/(2\sqrt{d})$. For any $x = (z, w) \in B_2^d$, put $r = 2f(x)$ and clip each z_i to $u_i \in [-a, a]$. The definition of f gives $|z_i - u_i| \leq r$ and $|w_e - (Bu)_e| \leq 3r$ for every edge. Thus $(u, Bu) \in Q$ and

$$\operatorname{dist}(x, Q) \leq 2\sqrt{n + 9m}\, f(x). \tag{15}$$

For every feasible pair x, y, the triangle inequality therefore yields $\varepsilon \leq \|x - y\| \leq \operatorname{diam}(Q) + 2\sqrt{n + 9m}\,[f(x) + f(y)]$. Taking the minimum and using $n + 9m \leq 9d$, we obtain

$$v_\varepsilon(B_2^d) \geq \frac{\varepsilon - \operatorname{diam}(Q)}{4\sqrt{n + 9m}} \geq \frac{a}{24d} = \frac{1}{2400d^2}. \tag{16}$$

An approximation with error at most $1/(10^4 d^2)$ therefore distinguishes $C^\star \geq k$ from $C^\star \leq k - 1$ by comparison with $1/(5000d^2)$. The reduction has polynomial size, proving the claim. □

D Additional Transition Geometry

We prove Lemma 3.2. Put $N = \lfloor D/\varepsilon \rfloor + 1 = p + s$. When $D \geq p\varepsilon$, we have $s \geq 1$. A diameter line segment of $\mathcal{X}$ contains an N-point parking set C with consecutive points spaced exactly ε apart, because $(N-1)\varepsilon \leq D$. Thus C is ε-separated, and every p-element subset is a feasible target set. Every point of $\mathcal{X}$ conflicts with at most two points of C: if it were at distance less than ε from three,

the two extreme parking points would be at distance less than 2ε, contradicting their separation by at least 2ε.

Let S be the current chronological state, put $A = B(S)$, and define $R = A \setminus C$, $U = C \setminus A$, and $q = |R|$. Then $|U| = q + s$. In the bipartite conflict graph between R and U, two points conflict when their distance is less than ε; every vertex of R has degree at most two. Let $U_0, U_1, U_{\geq 2}$ be the unused parking points of degree zero, one, and at least two. Edge counting yields

$$|U_1| + 2|U_{\geq 2}| \leq 2q, \qquad 2|U_0| + |U_1| \geq 2s.$$

If R_1 is the set of neighbors of U_1, then $|R_1| \geq |U_1|/2$, since each point of R has degree at most two. Hence $|U_0| + |R_1| \geq s$.

We may therefore choose $t = \min\{q, s\}$ disjoint pairs (r_i, u_i) with $r_i \in R$ and $u_i \in U$ such that u_i conflicts with no point of $A \setminus \{r_i\}$. Use a degree-one neighbor for each selected point of R_1, and if necessary pair distinct degree-zero vertices with distinct remaining points of R. The inserted u_i are mutually compatible because they belong to C.

During one block of p physical rounds, process the chronological state once. When r_i expires, play u_i; otherwise replay the expired point. Every action is admissible: a new u_i is compatible with all original points except possibly the expired r_i, and with every other inserted parking point. After the block, at least s points outside C have disappeared unless fewer than s remained. Thus any chronological state reaches a p-element subset of C within $\lceil p/s \rceil$ such blocks. These replacement blocks need not repeat a fixed ordered packing.

Each such replacement can be reversed at the level of point sets in one p-round block, starting from any ordering of the resulting set. Indeed, the old r_i is compatible with every retained original point, every u_j for $j \neq i$, and all previously restored old points. Restore r_i when u_i expires. Apply the construction separately to the source state and to any ordering of the target set. Their resulting p-subsets of C can be connected in one additional block by pairing and replacing their distinct elements. Reversing the target replacements reaches the target set in some chronological order, which can then be repeated; it need not recover the target ordering used to construct the replacements. Hence

$$H(\mathcal{X}; p, \varepsilon) \leq p\left(2\left\lceil \frac{p}{s} \right\rceil + 1\right).$$

Finally let $L = D/\varepsilon - p + 1$. Since $s = \lfloor D/\varepsilon - p \rfloor + 1$, we have $s \leq L < s + 1 \leq 2s$. Therefore

$$p\left(2\left\lceil \frac{p}{s} \right\rceil + 1\right) \leq 3p + \frac{2p^2}{s} \leq 3p + \frac{4p^2}{L},$$

which proves (4).

D.1 Matching lower bound for the geometric transition time

For completeness, we record the matching reachability statement corresponding to the diameter bridge. The quantity H measures the worst-case time from an ordered chronological state to an unordered target set, rather than to a prescribed ordered state. This result concerns only admissible transitions and is not used in the statistical lower bound.

Theorem D.1 (Worst-case tightness of the diameter dependence)**.** *For every $p \geq 2$, $\varepsilon > 0$, and $D \geq p\varepsilon$,*

$$\frac{1}{64}\left(p + \frac{p^2}{D/\varepsilon - p + 1}\right) \leq \sup_{\operatorname{diam}(K)=D} H(K; p, \varepsilon) \leq 3p + \frac{4p^2}{D/\varepsilon - p + 1}. \tag{17}$$

The supremum ranges over all ambient dimensions and all compact convex K of diameter D. The lower bound already holds on the interval $[0, D]$, including at $D = p\varepsilon$.

We now prove the lower half of Theorem D.1. This subsection uses only the blocking geometry; no loss function or observations are involved. Measure spatial coordinates in units of ε, so the interval becomes $[0, D/\varepsilon]$ and the exclusion radius is 1.

List the points of a feasible packing set in spatial order as $z_1 < \cdots < z_p$ in the interval $[0, D/\varepsilon]$; this ordering is distinct from their chronological order. Define the reduced coordinates

$$y_i := z_i - (i - 1), \qquad i = 1, \ldots, p, \tag{18}$$

and put

$$L := D/\varepsilon - p + 1.$$

Then feasibility is equivalent to

$$0 \le y_1 \le y_2 \le \cdots \le y_p \le L.$$

A replacement is said to *change spatial rank* if the newly inserted action has a different rank in the sorted packing from the expired action that it replaces.

Lemma D.2 (Propagation envelope before the first rank change)**.** *Assume $p \ge 3$ and set $h := L/(p+1) > 0$, where $L = D/\varepsilon - p + 1$. Let*

$$\sigma = \big(1, 3, 5, \ldots, 2\lceil p/2 \rceil - 1,\ 2, 4, \ldots, 2\lfloor p/2 \rfloor\big)$$

be the chronological rank order of the initial packing and put

$$q := \left\lfloor \frac{p-1}{2} \right\rfloor.$$

Suppose the initial reduced coordinates are $y_i(0) = ih$ for $i = 1, \ldots, p$, and suppose no spatial-rank-changing replacement occurs through round t. Define

$$K(t) := 1 + \left\lfloor \frac{t-1}{q} \right\rfloor \quad (t \ge 1), \qquad K(0) := 0.$$

Then

$$h \max\{0, i - K(t)\} \le y_i(t) \le h \min\{p+1, i + K(t)\}, \qquad i = 1, \ldots, p. \tag{19}$$

Proof. Until the first rank change, chronological ranks rotate with the same cyclic order. Rank i is updated once every p rounds at its fixed phase in σ. The phase positions are

$$\mathrm{pos}(2j-1) = j, \qquad \mathrm{pos}(2j) = \lceil p/2 \rceil + j.$$

For every adjacent spatial pair $i, i+1$, both positive cyclic distances between their phases are at least $q = \lfloor (p-1)/2 \rfloor$. Indeed, for an odd–even pair the two distances are $\lceil p/2 \rceil$ and $\lfloor p/2 \rfloor$, while for an even–odd pair they are $\lceil p/2 \rceil - 1$ and $\lfloor p/2 \rfloor + 1$. Consequently, two strictly time-ordered updates of neighboring reduced coordinates are separated by at least q physical rounds.

For the lower envelope, trace the most recent update of coordinate i backward in time. If rank $i > 1$ is updated at time u, admissibility gives

$$y_i(u) \ge y_{i-1}(u-1).$$

Continue with the most recent update of coordinate $i-1$ strictly before u, and iterate toward the left. Stop when reaching either an initial coordinate or the fixed boundary value $y_0 = 0$. A chain containing r positive-time updates requires at least $1 + (r-1)q$ rounds, hence $r \le K(t)$. Since the initial coordinates are jh, this proves the lower inequality in (19). The upper inequality follows symmetrically by tracing updates to the right using $y_i(u) \le y_{i+1}(u-1)$ and the fixed boundary $y_{p+1} = L = (p+1)h$. At $t = 0$ the claim is immediate. □

Lemma D.3 (A rank change requires an insertion gap)**.** *For a feasible packing of p points in $[0, D/\varepsilon]$ with reduced coordinates $0 \le y_1 \le \cdots \le y_p \le L$, an additional point compatible with all p packing points exists if and only if at least one of*

$$y_1 \ge 1, \qquad L - y_p \ge 1, \qquad y_{i+1} - y_i \ge 1 \text{ for some } i < p \tag{20}$$

holds. In particular, if none of these inequalities holds, the next admissible replacement cannot change spatial rank.

Proof. An additional compatible point can lie to the left of the packing, to the right, or between two adjacent packing points. These possibilities require, respectively, a physical endpoint gap at least 1, another endpoint gap at least 1, or an interior physical gap at least 2. Since

$$z_{i+1} - z_i = 1 + y_{i+1} - y_i,$$

these conditions are exactly (20).

Now suppose a replacement changed spatial rank. Then a retained packing point w lies strictly between the expired point u and the inserted point v on the line. Both the old and the new packing are feasible, hence $|u - w| \ge 1$ and $|v - w| \ge 1$, so $|u - v| \ge 2$. Thus v is compatible not only with every retained point but also with u, and therefore is an additional point compatible with the entire old packing. If no insertion gap exists, this is impossible. □

Proposition D.4 (Parameterized transition lower bound)**.** *For every $p \geq 2$, $\varepsilon > 0$, and $D \geq p\varepsilon$, there is a reachable chronological state S_0 of length p and a feasible unordered target set B on $[0, D]$ such that*

$$H(S_0, B) \geq \frac{1}{64}\left(p + \frac{p^2}{D/\varepsilon - p + 1}\right). \tag{21}$$

The constructed state–target pairs have finite transition time.

Proof. Work in the normalized coordinates above. Put

$$L := D/\varepsilon - p + 1 \geq 1, \qquad h := \frac{L}{p+1}, \qquad a_i := i - 1 + ih, \quad i = 1, \ldots, p.$$

The points a_i lie in $[0, D/\varepsilon]$, because $a_1 = h$ and $a_p = D/\varepsilon - h$, and adjacent gaps equal $1 + h > 1$. For $p \geq 3$, give them the chronological rank order σ from Lemma D.2; the choice $h = L/(p+1)$ and the initial reduced coordinates $y_i(0) = ih$ satisfy that lemma's boundary and initialization assumptions. For $p = 2$, use (a_1, a_2). The initial history is reachable by playing these p points once in the specified order.

Choose

$$0 < \lambda < \frac{h}{2p}$$

outside the finite set of values for which one of

$$b_i := i - 1 + i\lambda, \qquad i = 1, \ldots, p,$$

coincides with an initial point. The target $B = \{b_1, \ldots, b_p\}$ is a feasible packing set, lies in $[0, D/\varepsilon]$, and is disjoint from the initial packing set. Its largest reduced coordinate is

$$y_p^B = p\lambda < h/2.$$

Since B is disjoint from the initial window, reaching B requires at least p new physical rounds:

$$H(S_0, B) \geq p. \tag{22}$$

Let

$$q := \left\lfloor \frac{p-1}{2} \right\rfloor, \qquad r := \left\lfloor \frac{p+1}{4L} \right\rfloor, \qquad T_0 := qr.$$

If $q = 0$ or $r = 0$, the additional assertion $H(S_0, B) > T_0$ is trivial. Otherwise $p \geq 3$, $h \leq 1/4$, and for every $1 \leq t \leq T_0$ we have $K(t) \leq r$. Under the hypothesis that no rank change has yet occurred, Lemma D.2 yields

$$\begin{aligned} y_1(t) &\leq (r+1)h, \\ L - y_p(t) &\leq (r+1)h, \\ y_{i+1}(t) - y_i(t) &\leq (2r+1)h. \end{aligned} \tag{23}$$

The same bounds hold at $t = 0$. Since $r \leq (p+1)/(4L)$ and $h = L/(p+1)$,

$$(2r+1)h \leq \frac{1}{2} + h \leq \frac{3}{4} < 1, \qquad (r+1)h \leq \frac{1}{2} < 1.$$

Hence a first rank change cannot occur during rounds $1, \ldots, T_0$: if it first occurred at u, then (23) at time $u - 1$ would rule out every insertion gap in Lemma D.3, contradicting the rank change at round u.

Therefore Lemma D.2 is valid throughout this interval. In particular,

$$y_p(t) \geq (p - r)h > h/2, \qquad t \leq T_0,$$

where the final inequality follows from $r \leq (p+1)/4$ and $p \geq 3$. Since the target satisfies $y_p^B < h/2$, it cannot have been reached by time T_0. Thus

$$H(S_0, B) \geq \max\left\{p, 1 + \left\lfloor \frac{p-1}{2} \right\rfloor \left\lfloor \frac{p+1}{4L} \right\rfloor\right\}. \tag{24}$$

It remains to simplify constants. If $L > (p+1)/8$, then $p^2/L < 8p$, so (22) is stronger than (21). If $L \le (p+1)/8$, then $p \ge 7$ and

$$q \ge \frac{p}{4}, \qquad r \ge \frac{p+1}{8L}, \qquad T_0 \ge \frac{p^2}{32L}.$$

Consequently

$$H(S_0, B) \ge \max\left\{p, \frac{p^2}{32L}\right\} \ge \frac{1}{33}\left(p + \frac{p^2}{L}\right),$$

which implies the weaker universal constant $1/64$ in (21).

Finally, the state–target pair has finite transition time. Indeed, sort an arbitrary initial packing and target packing as $a_1 < \cdots < a_p$ and $b_1 < \cdots < b_p$. First replace every coordinate with $b_i < a_i$ in increasing i, and then every remaining coordinate with $b_i > a_i$ in decreasing i. Each replacement preserves feasibility by comparison with its two neighbors. Wait for the designated rank by replaying expired points; the first replacement takes at most p rounds and each later one at most $p-1$ additional rounds. Hence every interval state–target pair can be connected in at most $p(p-1)+1$ rounds. □

Proof of Theorem D.1. The upper bound in (17) is Lemma 3.2. For the lower bound, apply Proposition D.4 to the interval $[0, D]$, which is admissible in the supremum in (17). All lower bounds are pathwise, so randomization cannot shorten them. □

E Local Buffers and the Inball Upper Bound

We record the local geometric construction used in Lemma 3.1 and the general-period upper bound. Put

$$b = 2(2p)^{1/d}\varepsilon, \qquad \rho_{\mathrm{buf}} = b\left(1 + \frac{D}{r_0}\right).$$

Under (3), $b \le r_0$.

Lemma E.1 (Local admissible buffers)**.** *Assume* (3)*. Fix any center $a \in \mathcal{X}$, any current admissible history, and any nonempty set U of at most p future target points. There exist $p-1$ admissible buffer actions such that every buffer lies within distance ρ_{buf} of a, every buffer is at least ε from every target in U, and after the buffers any target in U can be played. If $U = B(P)$ for some $P \in \mathcal{P}_p^{\varepsilon}(\mathcal{X})$, the prescribed ordering P can then be repeated indefinitely.*

Proof. Set $\theta = b/r_0 \le 1$ and define

$$D_a := \overline{B}((1-\theta)a + \theta c, b).$$

For every $w \in D_a$, write $w = (1-\theta)a + \theta v$ with $v \in \overline{B}(c, r_0)$. Convexity gives $D_a \subseteq \mathcal{X}$, and

$$\|w - a\| \le b + \theta\,\|c - a\| \le b\left(1 + \frac{D}{r_0}\right) = \rho_{\mathrm{buf}}.$$

At one buffer step, exclude the closed ε-balls centered at the last at most $p-1$ actual actions and at all targets in U. There are at most $2p-1$ excluded balls. If v_d is the volume of the unit ball, their total volume inside D_a is at most $(2p-1)v_d\varepsilon^d$, whereas

$$\mathrm{vol}(D_a) = v_d b^d = 2^d(2p)v_d\varepsilon^d > (2p-1)v_d\varepsilon^d.$$

Hence an admissible buffer outside all excluded closed balls exists. Repeat this argument $p-1$ times. The pre-buffer history has then expired, all remaining buffer actions avoid every target, and feasibility among the targets makes the first and all later cycles admissible. □

Proof of Lemma 3.1. Lemma E.1 gives $p-1$ buffer rounds that avoid all p target points. Playing the p targets next in their prescribed order reaches that chronological state, so $H(\mathcal{X}; p, \varepsilon) \le 2p-1$. These $2p-1$ rounds include the target plays used to establish the state; any subsequent fresh cycle uses another p rounds.

For the lower bound, (3) implies

$$r_0^d \geq 2^{d+1} p \varepsilon^d.$$

A greedy volume argument therefore selects $2p$ points in $\overline{B}(c, r_0)$ with pairwise distances strictly greater than ε: before the last selection, fewer than $2p$ closed ε-balls have total volume strictly below that of the inball. Use the first p points as a reachable initial state and the last p as a disjoint target set. After fewer than p new rounds, at least one initial point remains among the last p actions, so their set cannot equal the target set. Hence $H(\mathcal{X}; p, \varepsilon) \geq p$. □

F Proofs for Period Two

A *query segment* is a maximal consecutive interval of identical virtual queries within a single run of the BCO procedure. Segment lengths count virtual queries, not physical rounds; a new run starts a new query segment even if its first query equals the preceding run's last query.

Lemma F.1 (Stochastic convex-bandit primitive with few query segments)**.** *Let K be a nonempty compact convex set of affine dimension at most k, and let $g : K \to [0, 1]$ be continuous and convex. Under conditionally centered s-sub-Gaussian value noise, there is a value-feedback procedure such that, for every planned horizon $n \geq 1$ and failure probability $\xi \in (0, 1)$, with probability at least $1 - \xi$,*

$$\sum_{j=1}^{n} (g(z_j) - \min_K g) \leq A_k (1 + s) \sqrt{n} \, [\log(en/\xi)]^4, \tag{25}$$

$$\#\{\textit{query segments}\} \leq B_k [\log(en/\xi)]^3. \tag{26}$$

The constants A_k, B_k can be chosen polynomial in k.

Proof. We use the stochastic convex-bandit construction and analysis of Agarwal et al. Agarwal et al. (2013). Their Algorithm 2 already evaluates every confidence-test point by a consecutive batch of repeated observations, and the proof of their Theorem 2 gives $\widetilde{O}_k(\sqrt{n})$ regret under their normalization. We adapt that proof to bounded convex losses and conditional sub-Gaussian noise, and we make explicit the number of query segments, which is not stated separately in Agarwal et al. (2013).

First identify the affine hull of K with $\mathbb{R}^r$, where $r \leq k$. If $r = 0$, K is a singleton and the claim is immediate. For $r = 1$, use Algorithm 1 of Agarwal et al. (2013); its interval-shrinkage proof gives the same or a smaller number of query segments with the confidence and range modifications below. For $r \geq 2$, use their Algorithm 2. Apply an affine rounding of K. This changes neither convexity nor the number of query segments and only introduces constants depending on r. Normalize the observations by

$$\widetilde{Y} = \frac{Y}{1 + s}, \qquad \widetilde{g} = \frac{g}{1 + s}.$$

The normalized noise is conditionally 1-sub-Gaussian, and $0 \leq \widetilde{g} \leq 1$. Put

$$\ell := \log \frac{8en}{\xi}.$$

Run the appropriate algorithm of Agarwal et al. (2013) with its confidence-test sample sizes enlarged, if necessary, from a constant multiple of $\log n/\gamma^2$ to a constant multiple of ℓ/γ^2. If a test point z is chosen from the past and queried m consecutive times, the conditional sub-Gaussian assumption gives

$$\Pr\left(\left| \frac{1}{m} \sum_{q=1}^{m} (\widetilde{Y}_q - \widetilde{g}(z)) \right| > \gamma \;\middle|\; \text{history before the batch} \right) \leq 2e^{-m\gamma^2/2}.$$

There can be at most n nonempty query batches before the horizon ends. Hence, after increasing the numerical constant in the batch size, a union bound gives a simultaneous confidence event $\mathcal{E}$ of probability at least $1 - \xi$ on which every confidence interval used by the algorithm is valid. The geometric arguments in Agarwal et al. (2013) are deterministic conditional on this event, so their proof applies unchanged with ℓ in place of $\log n$, apart from the two boundedness points discussed below.

We next count query segments for $r \geq 2$; Algorithm 1 has the analogous interval tests with fewer points. In round i of an epoch, Algorithm 2 uses the outer confidence level

$$\gamma_i = 2^{-i}.$$

By Lemma 8 of Agarwal et al. (2013), the number of successive pyramids constructed in one round is bounded by a constant C_r depending only on r. For each such pyramid the inner confidence level starts at $1/2$ and is halved until either the epoch terminates or it drops below γ_i. Thus a pyramid uses at most i confidence levels. At each confidence level at most $r+2$ points are tested, and the initial simplex contributes at most $r+1$ additional test points. Executing all repetitions of one test point consecutively therefore yields at most

$$C'_r(1+i) \tag{27}$$

query segments in round i.

To start outer round $i \geq 2$, the previous round $i-1$ must have completed; it has spent at least $c\ell 4^{i-1}$ observations on a confidence test. Thus even an interrupted final round has index at most

$$I_n \leq C + \frac{1}{2}\log_2 \frac{n}{\ell} \leq C' \log \frac{en}{\xi},$$

where the case $n \leq c\ell$ is absorbed into the constants by the trivial bound $R_n \leq n$ and one query segment.

It remains to bound the number of epochs without adding a Lipschitz assumption. Lemmas 12 and 13 of Agarwal et al. (2013) imply, on $\mathcal{E}$, that each completed epoch contracts the enclosing-ellipsoid volume by a factor $\rho_r < 1$ depending only on r, and every point discarded at an epoch ending with confidence level γ has suboptimality at least γ. Let $z^\star$ minimize $\widetilde{g}$ over K, and let $\gamma_{\min}$ denote the smallest confidence level that can be reached before exhausting n observations. Up to universal constants,

$$\gamma_{\min} \geq \sqrt{\frac{\ell}{n}}.$$

For $\lambda = \gamma_{\min}/2$, consider the homothetic copy

$$K_\lambda = (1-\lambda)z^\star + \lambda K.$$

For $z = (1-\lambda)z^\star + \lambda x \in K_\lambda$, convexity and $0 \leq \widetilde{g} \leq 1$ give

$$\widetilde{g}(z) - \widetilde{g}(z^\star) \leq \lambda\big(\widetilde{g}(x) - \widetilde{g}(z^\star)\big) \leq \lambda < \gamma_{\min}.$$

Hence no point of K_λ is ever discarded. Its r-dimensional volume equals $\lambda^r \operatorname{vol}_r(K)$. The initial affine rounding makes the ratio between the initial enclosing-ellipsoid volume and $\operatorname{vol}_r(K)$ bounded by a constant depending only on r. Therefore, if E epochs are completed,

$$c_r \rho_r^E \operatorname{vol}_r(K) \geq \operatorname{vol}_r(K_\lambda) = \lambda^r \operatorname{vol}_r(K),$$

which implies

$$E \leq C''_r \log \frac{en}{\xi}. \tag{28}$$

This is the bounded-range analogue of Lemma 14 in Agarwal et al. (2013); it replaces the Euclidean near-optimal ball used there by the homothetic set K_λ. The dimension dependence remains polynomial: Lemma 8 of that paper gives $C_r = O(r^2 \log(er))$, Lemma 12 gives $-\log \rho_r = \Omega((r+1)^{-3/2})$, and affine rounding gives $\log c_r = O(r \log(er))$.

Count the current unfinished epoch as one additional epoch. Combining (27) and (28),

$$\#\{\text{query segments}\} \leq C''_r \log \frac{en}{\xi} \sum_{i=1}^{I_n} (1+i) \leq B_k \left[\log \frac{en}{\xi}\right]^3,$$

after enlarging B_k and using $r \leq k$. This proves (26).

Finally, the regret proof of Theorem 2 in Agarwal et al. (2013), conditional on the simultaneous confidence event, gives $O_r(\sqrt{n}\,\ell^2)$ regret for the normalized problem, with a coefficient polynomial in r. In that proof, the first-round use of Lipschitzness is replaced here by the direct range bound $0 \leq \widetilde{g} \leq 1$, while the Lipschitz-based epoch-count argument is replaced by the homothetic argument

above. Enlarging the logarithmic envelope from ℓ^2 to $[\log(en/\xi)]^4$ and rescaling by $1+s$ gives (25). The regret accounting also includes the final interrupted confidence test: at its scale γ, its per-query gap is bounded by $C_r\gamma$ by the preceding valid test (or by the range bound at the initial scale), and its realized length is at most $\min\{n, C\ell/\gamma^2\}$. Its contribution is therefore at most $C_r\sqrt{n\ell}$, within the displayed envelope. Increasing A_k, B_k handles the finitely many small-horizon cases and completes the proof. □

The bound on the number of query segments is essential because a transfer is paid once per query segment, not once per observation.

F.1 A FINITE FAMILY OF CONVEX PACKING BRANCHES

Let $K_1, \dots, K_N$ be nonempty compact convex subsets of $\mathcal{X}^p$ such that

$$K_i \subseteq \mathcal{P}_p^{\varepsilon}(\mathcal{X}), \qquad v_i := \min_{P\in K_i} F_f(P), \qquad v_{\mathcal{K}} := \min_i v_i \le v_{\varepsilon}(\mathcal{X}) + a. \tag{29}$$

Suppose a known transition procedure reaches any specified feasible target set B from any chronological state within at most H rounds. At arrival it returns the chronological tuple Q of the last p actions, so $B(Q) = B$ and repeating Q is admissible. From an empty history, playing any ordering of B once initializes such a state in p rounds.

A virtual query $P = (u_1, \dots, u_p) \in K_i$ is implemented by transferring to $B(P)$ and then playing one fresh complete cycle in the arrival order Q. The virtual query remains P: the permutation Q need not belong to K_i, but $F_f(Q) = F_f(P)$. Transfer observations, including target plays used to establish the arrival state, are discarded. The fresh cycle's averaged observation is

$$Z = \frac{1}{p}\sum_{j=1}^{p} Y_{t+j} = F_f(P) + \zeta, \qquad \mathbb{E}[e^{\lambda\zeta} \mid \text{before the cycle}] \le e^{\lambda^2\sigma^2/(2p)}. \tag{30}$$

For a query segment of length m, keep the arrival order Q fixed throughout. Its physical cost is at most $H + pm$, with one transition and m fresh cycles. If the segment starts from an empty history, its initialization costs p rounds instead of the transfer; this is charged separately below.

Lemma F.2 (Finite convex-family master)**.** *Under* (29) *and Lemma F.1, let*

$$\Lambda := \log\frac{64N(T+1)^4}{\delta}.$$

There is a constant $C_{d,p,\sigma} > 0$, depending only on d, p, σ and the primitive constants A_{dp}, B_{dp}, such that an admissible algorithm satisfies

$$R_T^{\varepsilon} \le aT + C_{d,p,\sigma}\Lambda^4\big(\sqrt{pNT} + HN\big) + 2p + \delta T. \tag{31}$$

Proof. For every branch i, keep only its most recently completed fresh BCO run. If this run has length n_i and mean observation $\bar{Z}_i$, define

$$B(n) = A_{dp}\left(1 + \frac{\sigma}{\sqrt{p}}\right)\Lambda^4\sqrt{n}, \qquad r(n) = \frac{\sigma}{\sqrt{p}}\sqrt{\frac{2\Lambda}{n}}, \qquad L_i = \bar{Z}_i - \frac{B(n_i)}{n_i} - r(n_i).$$

An unvisited branch has $L_i = -\infty$. Select a branch with minimum L_i and use fresh run lengths $1, 2, 4, \dots$ on each branch. Give each run failure probability $\xi := \delta/[4N(T+1)]$; its logarithmic envelope is at most Λ. Within every inner run, execute query segments in physical time without interleaving other queries.

For a completed length-n run on branch i, the BCO guarantee and (30) imply, on the allocated high-probability event,

$$v_i - \frac{B(n_i)}{n_i} - 2r(n_i) \le L_i \le v_i.$$

A union bound over all branches and the $O(\log T)$ planned run lengths gives probability at least $1-\delta$. If branch i is selected after a previous run of length n, comparison with an optimal branch yields

$$v_i - v_{\mathcal{K}} \le \frac{B(n)}{n} + 2r(n).$$

The next run has planned length $2n$, so its virtual regret is at most $C_{d,p,\sigma}\Lambda^4\sqrt{n}$. The same bound covers the final run if the physical horizon interrupts it after $q \le 2n$ complete query cycles: extend that run hypothetically to all $2n$ virtual queries with valid additional noise. On the primitive's success event, the intrinsic regret of its actual prefix is at most $B(2n)$, since each term $F_f(P) - v_i$ is nonnegative; its branch-selection gap contributes at most $q(B(n)/n + 2r(n))$. An interrupted first run has planned length one and costs at most one query cycle. At most one run is interrupted, and its planned length is at most one plus the number of previously completed virtual queries on that branch, so it changes the geometric summation by only a constant factor. Summing the geometric run lengths over branches and using Cauchy–Schwarz gives a total statistical contribution at most

$$C_{d,p,\sigma}\Lambda^4\sqrt{pNT}.$$

Each branch has only $O(\log T)$ fresh runs, and Lemma F.1 gives at most $B_{dp}\Lambda^3$ query segments per run, including the prefix of the final interrupted run by the same hypothetical completion. Since $\log T \lesssim \Lambda$, all bridge rounds contribute at most

$$C_{d,p,\sigma}\Lambda^4 HN.$$

Initialization and a final incomplete cycle cost at most $2p$. Finally, replacing $v_{\mathcal{K}}$ by $v_\varepsilon(\mathcal{X})$ costs at most aT, and the failure event contributes at most δT. This proves (31). □

F.2 Proof of the directional approximation and Theorem 5.2

For $d = 1$, the two points of S^0 form an exact net and satisfy the stated bound. Suppose $d \ge 2$. Cover S^{d-1} by the $2d$ coordinate charts $\{u : \pm u_j \ge 1/\sqrt{d}\}$. Each chart is the image of

$$A := \{w \in \mathbb{R}^{d-1} : \|w\| \le \sqrt{1 - 1/d}\}$$

under the map $\phi_{j,\pm}$ that inserts $\pm\sqrt{1 - \|w\|^2}$ as coordinate j. For $w \in A$ and $v \in \mathbb{R}^{d-1}$,

$$\|D\phi_{j,\pm}(w)v\|^2 = \|v\|^2 + \frac{\langle w, v\rangle^2}{1 - \|w\|^2} \le d\,\|v\|^2 .$$

Since A is convex, each chart map is $\sqrt{d}$-Lipschitz. Put $\delta = \theta/(2\sqrt{d})$ and take a maximal finite Euclidean δ-separated set $W \subseteq A$. Maximality makes W a Euclidean δ-net of A. The open balls of radius $\delta/2$ centered at its points are disjoint and lie in the radius-$(1 + \delta/2)$ ball in $\mathbb{R}^{d-1}$, so volume comparison gives

$$|W| \le (1 + 2/\delta)^{d-1} = (1 + 4\sqrt{d}/\theta)^{d-1}.$$

The same bound guarantees that a greedy construction of such a maximal set terminates. The images $\phi_{j,\pm}(W)$ cover their respective charts within Euclidean distance $\sqrt{d}\,\delta \le \theta$. Their union is therefore a θ-net of the sphere, and, since $\theta \le 1$,

$$|\mathcal{U}_\theta| \le 2d(1 + 4\sqrt{d}/\theta)^{d-1} \le 2d(1 + 4\sqrt{d})^{d-1}\theta^{-(d-1)} = C_{\mathrm{dir}}(d)\theta^{-(d-1)}.$$

Proof of Lemma 5.1. Nonemptiness follows because $Q_u = (c + r_0 u, c - r_0 u) \in \mathcal{X}^2$ has projected separation $2r_0 > \varepsilon(1+\theta^2)$. Take any feasible pair $P = (x, y)$ and let $v = (x-y)/\|x - y\|$. Choose $u \in \mathcal{U}_\theta$ with $\|u - v\| \le \theta$. Since u, v are unit vectors,

$$\langle u, v\rangle = 1 - \frac{1}{2}\|u - v\|^2 \ge 1 - \theta^2/2.$$

Writing $A = \langle u, x - y\rangle$ gives $A \ge \varepsilon(1 - \theta^2/2)$. If $A \ge \varepsilon(1 + \theta^2)$, then $P \in K_u$. Otherwise define

$$\alpha = \frac{\varepsilon(1 + \theta^2) - A}{2r_0 - A}, \qquad P' = (1 - \alpha)P + \alpha Q_u.$$

Then $0 < \alpha < 1$, $P' \in \mathcal{X}^2$, and its projected separation is exactly $\varepsilon(1 + \theta^2)$. Moreover

$$\varepsilon(1 + \theta^2) - A \le \frac{3}{2}\varepsilon\theta^2, \qquad 2r_0 - A > 2r_0 - \varepsilon(1 + \theta^2) \ge \frac{3}{4}r_0,$$

so $\alpha \le 2(\varepsilon/r_0)\theta^2$. By convexity and $0 \le f \le 1$,

$$F_f(P') \le (1 - \alpha)F_f(P) + \alpha F_f(Q_u) \le F_f(P) + \alpha.$$

Taking the minimum over feasible pairs proves the upper inequality. The lower inequality holds because every branch consists only of feasible packings. □

Proof of Theorem 5.2. Apply Lemma F.2 with $p = 2$, $a = (2\varepsilon/r_0)\theta^2$, $N \le C_{\mathrm{dir}}(d)\theta^{-(d-1)}$, $H \le 10$, and $\delta = (T+1)^{-2}$. Since the resulting logarithmic factor is $O_d(\log(eT/\theta))$, this gives (10) after enlarging $C_{d,\sigma}$.

Let $q = d-1$. When $\theta_T = ((\varepsilon/r_0)^2 T)^{-1/(d+3)}$,

$$\frac{\varepsilon}{r_0}\,\theta_T^2 T = \left(\frac{\varepsilon}{r_0}\right)^{q/(q+4)} T^{(q+2)/(q+4)}, \qquad \sqrt{T\theta_T^{-q}} = \left(\frac{\varepsilon}{r_0}\right)^{q/(q+4)} T^{(q+2)/(q+4)}.$$

The remaining term $\theta_T^{-q} = ((\varepsilon/r_0)^2 T)^{q/(q+4)}$ is lower order. If $\theta_T = 1/2$, then $(\varepsilon/r_0)^2 T \le 2^{d+3}$ and every term in (10) is $\widetilde{O}_{d,\sigma}(\sqrt{T})$. This proves (11). □

G Proofs of the General-Period Upper Bounds

Lemma G.1 (Dimension-explicit statistical partition)**.** *Let $K \subseteq B(z,R)$ be nonempty and compact. For every $0 < r \le 6R$, there is a partition into M cells of diameter at most r with*

$$M \le (6R/r)^d. \tag{32}$$

The cells are used only for statistics; packing points are not required to be cell representatives.

Proof. If $r \ge 2R$, use one cell. Otherwise take a maximal $r/2$-separated finite subset of K and its Voronoi partition, breaking ties deterministically. Maximality gives covering radius at most $r/2$, hence cell diameter at most r. The open balls of radius $r/4$ around the centers are disjoint and contained in $B(z, R + r/4)$. Comparing volumes gives

$$M \le (1 + 4R/r)^d \le (6R/r)^d,$$

where the last inequality uses $r < 2R$. Greedy construction terminates because the same volume argument bounds the number of separated centers. □

Lemma G.2 (Transitions confined to the localized region)**.** *Let $H \ge p$ be a known uniform budget for reaching any feasible target set from any chronological state in $\mathcal{X}$. For any $z \in \mathcal{X}$, set*

$$K_z = \mathcal{X} \cap \overline{B}(z, 2p\varepsilon), \qquad \beta := \min\{1, 4Lp\varepsilon\}.$$

There is a known budget $\bar{H} \le 3H$ such that, from any chronological state $S \in \mathcal{P}_p^\varepsilon(K_z)$, every target set $B(P)$ with $P \in \mathcal{P}_p^\varepsilon(K_z)$ can be reached within $\bar{H}$ rounds using only actions in K_z. The arrival order Q may differ from P; repeating Q is admissible. An arbitrary initial chronological state in $\mathcal{P}_p^\varepsilon(\mathcal{X})$ can reach the same target set within H rounds. From an empty history, playing P once also establishes a target state within $p \le H$ rounds. Consequently, except for this first entry, every transition action has excess loss at most β above $v_\varepsilon(K_z)$.

Proof. Write $D = \mathrm{diam}(\mathcal{X})$. If $D \le 2p\varepsilon$, then $K_z = \mathcal{X}$, and the given transition procedure works with $\bar{H} = H$.

Suppose $D > 2p\varepsilon$, and let $a, b \in \mathcal{X}$ be diameter endpoints. With $\lambda = 2p\varepsilon/D$, define

$$a' = (1-\lambda)z + \lambda a, \qquad b' = (1-\lambda)z + \lambda b.$$

Convexity puts the whole line segment $[a', b']$ in $\mathcal{X}$, while $\|a' - z\|, \|b' - z\| \le \lambda D = 2p\varepsilon$. The segment therefore lies in K_z and has length exactly $2p\varepsilon$. It contains $2p+1$ equally spaced parking points with separation ε. Applying the parking construction in the proof of Lemma 3.2 with $N = 2p+1$ and $s = p+1$ connects any source chronological state to any feasible target set in K_z in at most

$$p\left(2\left\lceil \frac{p}{p+1} \right\rceil + 1\right) = 3p \le 3H$$

rounds, with the target order determined by the construction. Every action in that construction is a source point, a target point, or a parking point, so the entire transition stays in K_z. Take $\bar{H} = 3p$ in this case. The initial entry uses the originally assumed budget H; its intermediate actions need not lie in K_z.

Finally, $\mathrm{diam}(K_z) \le 4p\varepsilon$, so the oscillation of f on K_z is at most β. Since $v_\varepsilon(K_z) \ge \min_{K_z} f$, every $x \in K_z$ satisfies $f(x) - v_\varepsilon(K_z) \le \beta$. □

G.1 THE PACKING LOOP

Condition on the history when a phase starts and fix a region K and its partition. Assume that a known $\beta \in [0,1]$ satisfies

$$\sup_{x\in K} f(x) - \inf_{x\in K} f(x) \le \beta.$$

Initialize fresh target-only statistics $n_i, \widehat{f}_i$ and a global target count and mean

$$N = \sum_i n_i, \qquad \widehat{f}_K = \frac{1}{N}\sum_i n_i \widehat{f}_i \quad (N>0).$$

Transition and pre-phase observations are not inserted. At every batch start form the ordinary cell index

$$\ell_i = \begin{cases} \left[\widehat{f}_i - Lr - \sigma\sqrt{2\Lambda/n_i}\right]_{[0,1]}, & n_i > 0, \\ 0, & n_i = 0, \end{cases}$$

and the shared lower index

$$g = \begin{cases} \left[\widehat{f}_K - \beta - \sigma\sqrt{2\Lambda/N}\right]_{[0,1]}, & N > 0, \\ 0, & N = 0. \end{cases}$$

Use $\underline{f}_i = \max\{\ell_i, g\}$ and $\underline{f}(x) = \underline{f}_i$ on cell C_i. The shared index prevents an unvisited cell from receiving an arbitrarily low absolute baseline when all losses in K are close together. It uses the same target observations as the cell estimates.

At a batch start select an ordered packing $P = (u_1, \dots, u_p) \in \mathcal{P}_p^{\varepsilon}(K)$ whose average index $\underline{F}(P) := p^{-1}\sum_{j=1}^{p} \underline{f}(u_j)$ is within T^{-2} of the infimum. Set $m_i = \sum_{j=1}^{p} \mathbf{1}\{u_j \in C_i\}$ and freeze

$$\kappa = \max\left\{1, \min_{i:m_i>0} \lfloor n_i/m_i \rfloor\right\}. \tag{33}$$

Use one admissible transition to $B(P)$, record the arrival order Q, and then play κ fresh complete cycles in that fixed order, unless time ends. Since Q is a permutation of P, the average index and all cell multiplicities m_i are unchanged. One *refinement batch* consists of this selection, transition, and the subsequent cycles; it is complete when all κ cycles finish. A new batch recomputes the indices and κ, even if it selects the same packing as the previous batch. Only observations from complete fresh cycles update the statistics.

Lemma G.3 (Uniform confidence event)**.** *If $\Lambda \ge \log(2(M+1)T^3)$, then with conditional probability at least $1 - T^{-2}$, all the following statements hold simultaneously at every reached batch start. For every $x \in C_i$,*

$$\underline{f}(x) \le f(x), \qquad f(x) - \underline{f}(x) \le 2Lr + 2\sigma\sqrt{2\Lambda/n_i} \quad (n_i > 0). \tag{34}$$

Moreover, for every $x \in K$ and every reached $N > 0$,

$$f(x) - \underline{f}(x) \le 2\beta + 2\sigma\sqrt{2\Lambda/N}. \tag{35}$$

All widths are also at most one.

Proof. For cell i, expose the first m target observations in that cell. The exponential supermartingale for the predictable indicator of that cell, stopped at its mth visit and at the horizon, gives

$$\mathbb{P}\left(\left|\sum_{s:\, x_s\in C_i}^{\text{first } m} \eta_s\right| > \sigma\sqrt{2\Lambda m}, \text{ the } m\text{th visit occurs}\right) \le 2e^{-\Lambda}.$$

The same argument for the indicator of all target-cycle observations controls their first m noise terms. A union bound over at most $(M+1)T$ count pairs gives probability at least $1 - T^{-2}$.

On this event, the empirical mean in a cell differs from the average true loss of its observations by at most $\sigma\sqrt{2\Lambda/n_i}$. The average true loss differs from $f(x)$ by at most Lr for $x \in C_i$. This proves optimism and the cell width for ℓ_i.

For the global estimate, let μ_N be the average true loss of the first N target observations. All these actions lie in K, so $|\mu_N - f(x)| \le \beta$ for every $x \in K$. Consequently

$$\widehat{f}_K - \beta - \sigma\sqrt{2\Lambda/N} \le f(x), \qquad f(x) - \Big(\widehat{f}_K - \beta - \sigma\sqrt{2\Lambda/N}\Big) \le 2\beta + 2\sigma\sqrt{2\Lambda/N}.$$

Clipping to $[0,1]$ preserves these conclusions. Taking the maximum of the two optimistic indices preserves optimism and both width bounds. Conditional on the pre-phase history, the region and its partition are fixed; hence a random anchor causes no selection bias. □

Lemma G.4 (Physical-time packing-loop bound)**.** *Suppose the first entry into a selected target set $B(P)$ uses at most H_{in} rounds, while every later transition uses at most $\bar{H} \ge p$ rounds and stays inside K. The first entry may leave K. With total remaining horizon at most T, the refinement loop of Algorithm 1 has expected excess above $v_\varepsilon(K)$ at most*

$$4LrT + C\sigma\sqrt{\Lambda MT} + C\beta\big(p + \bar{H}\log(eT)\big)M + H_{\text{in}} + \beta\bar{H} + \beta p + 2, \tag{36}$$

where $C > 0$ is universal and one may take $\Lambda = \log(2(M+1)T^3)$. The statement is uniform over initial admissible histories for which the assumed first-entry procedure is available.

Proof. Charge each completed batch to a selected cell minimizing $\lfloor n_i/m_i \rfloor$. If this minimum is zero, the charged count either becomes positive for the first time or more than doubles. If it is $k \ge 1$, then

$$\kappa m_i = k m_i > \frac{k}{k+1} n_i \ge n_i/2.$$

Thus each cell is charged at most $C\log(eT)$ completed batches, and there are at most $CM\log(eT)+1$ attempted batches. The first transition costs at most H_{in} excess loss. All later transition actions lie in K and have excess at most β, giving the transition terms in (36).

We now control complete target cycles on the confidence event. The first batch has $N = 0$ and $\kappa = 1$; its target-cycle regret is at most $p\beta$, which we charge directly. For every later batch, $N \ge p$. Let a_i be the number of its observations in cell i from complete cycles, and let $a = \sum_i a_i$. The batch rule ensures

$$a \le p\kappa \le N. \tag{37}$$

Indeed, if $\kappa = 1$ this follows from $N \ge p$; otherwise $\kappa m_i \le n_i$ for every selected cell, and summing proves the claim.

Call a cell cold in a batch if its initial count satisfies $n_i < p$. For any fixed cell, all observations from such batches total at most $2p$: before the last cold batch its count is below p, and that batch adds at most p observations. The latter fact follows from $a_i \le m_i \le p$ when $\kappa = 1$, and from $a_i \le n_i < p$ when $\kappa > 1$. Using (35), the constant part of the cold-cell width charge is therefore at most $4\beta pM$.

The noise part of the global bound can be charged to all observations in a batch. By (37),

$$\frac{a}{\sqrt{N}} \le (1+\sqrt{2})\big(\sqrt{N+a} - \sqrt{N}\big).$$

Telescoping over batches bounds the total global noise charge by $C\sigma\sqrt{\Lambda T}$.

For a warm cell, whose initial count is $n_i \ge p$, the batch rule gives $a_i \le n_i$, including when $\kappa = 1$. Hence

$$\frac{a_i}{\sqrt{n_i}} \le (1+\sqrt{2})\big(\sqrt{n_i + a_i} - \sqrt{n_i}\big).$$

Apply (34), telescope over all warm-cell increments, and use $\sum_i \sqrt{N_i} \le \sqrt{MT}$. Their total width charge is at most $2LrT + C\sigma\sqrt{\Lambda MT}$. Omitted cold-cell potential increments are nonnegative, so they cannot invalidate this telescoping upper bound.

Optimism and the oracle tolerance charge every complete cycle after the first to its pointwise widths plus p/T^2. The accumulated tolerance is at most $1/T$. At most $p-1$ final target observations lie in an incomplete cycle, each with excess at most β. On confidence failure the entire phase excess is at most T, adding at most T^{-1} in expectation. Combining these bounds, absorbing the first cycle into $C\beta pM$, and enlarging $2LrT$ to $4LrT$ proves (36). □

Combining Lemma G.1 with the physical-time bound gives the following tuning.

Lemma G.5 (Optimized cell radius)**.** *For a phase restricted to $K \subseteq B(z, R)$, under the transition and oscillation assumptions of Lemma G.4, the refinement loop of Algorithm 1 can be tuned so that its expected excess above $v_\varepsilon(K)$ is*

$$\min\left\{T,\ \widetilde{O}_{d,\sigma}\left(\sqrt{T} + (LR)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} + (LR)^{\frac{d}{d+1}} (\beta\bar{H})^{\frac{1}{d+1}} T^{\frac{d}{d+1}} + H_{\mathrm{in}} + \beta\bar{H}\right)\right\}. \tag{38}$$

Here and throughout this section, the dependence on d suppressed by $\widetilde{O}_{d,\sigma}$ is at most polynomial. In particular, for $K = K_z$, Lemma G.2 allows $H_{\mathrm{in}} = H$, $\bar{H} \le 3H$, and $\beta = \min\{1, 4Lp\varepsilon\}$.

Proof. If $L = 0$, take $r = 6R$ and use the fixed-packing branch of Algorithm 1; the claim is immediate because f is constant. Otherwise write $r = 6Rx$ with $0 < x \le 1$, so Lemma G.1 gives $M \le x^{-d}$. Since $\bar{H} \ge p$, substituting into (36) yields, up to logarithmic factors, a factor polynomial in d, and constants depending only on σ,

$$LRT\,x + \sqrt{T}\,x^{-d/2} + \beta\bar{H}x^{-d} + H_{\mathrm{in}} + \beta\bar{H}.$$

Choose x as the larger of

$$x_{\mathrm{stat}} = \left((LR)^{-2}T^{-1}\right)^{1/(d+2)}, \qquad x_{\mathrm{tr}} = \left(\frac{\beta\bar{H}}{LRT}\right)^{1/(d+1)},$$

clipped at one. When neither value exceeds one, the increasing discretization term is bounded by the sum of its values at these two balancing points, and each decreasing term is bounded by its value at its own balancing point. This gives the displayed rates.

To control the logarithmic factors uniformly in d, put $S := (LR)^{d/(d+2)}T^{(d+1)/(d+2)}$. If $S \ge T$, the trivial cap by T suffices. Otherwise $(LR)^d < T$, and, when clipping is inactive, $x \ge x_{\mathrm{stat}}$ implies

$$M \le x_{\mathrm{stat}}^{-d} = \left((LR)^2T\right)^{d/(d+2)} \le T.$$

Thus we may take $\Lambda = \log(2(M+1)T^3) = O(\log(eT))$ with an absolute implicit constant. The substitution $r = 6Rx$ also shows that the partition bound introduces no factor 6^d.

For completeness, if the clipping is active, take $r = 6R$, enter an arbitrary fixed feasible packing set in K, and repeat its arrival order as in the first branch of Algorithm 1. This has regret at most

$$H_{\mathrm{in}} + \min\{\beta, 2LR\}T.$$

If $x_{\mathrm{stat}} > 1$, then $LRT < \sqrt{T}$, so this is at most $H_{\mathrm{in}} + 2\sqrt{T}$. If $x_{\mathrm{tr}} > 1$, then $LRT < \beta\bar{H}$, so it is at most $H_{\mathrm{in}} + 2\beta\bar{H}$. Either case is covered by (38). Finally $f \in [0,1]$ always gives the cap T. □

Proposition G.6 (Localized packing learner)**.** *Assume* (3)*, fix $z \in \mathcal{X}$, and let $K_z = \mathcal{X} \cap \overline{B}(z, 2p\varepsilon)$. Starting from any admissible history, the refinement loop of Algorithm 1 can be implemented admissibly so that its expected excess above $v_\varepsilon(K_z)$ is at most*

$$\widetilde{O}_{d,\sigma}\Bigg(\sqrt{T} + (Lp\varepsilon)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} + (Lp\varepsilon)^{\frac{d}{d+1}} \left(p\min\{Lp\varepsilon, 1\}\right)^{\frac{1}{d+1}} T^{\frac{d}{d+1}} + p\Bigg). \tag{39}$$

Proof. Proposition 2.1 ensures $\mathcal{P}_p^\varepsilon(K_z) \neq \varnothing$. Set $H := 2p-1$, as allowed by Lemma 3.1. Lemma E.1 gives the first entry from any admissible history within H rounds, including a history with fewer than p actions. By Lemma G.2, later transfers stay in K_z; one may use the uniform budget $\bar{H} := 3p \le 3H$ and $\beta = \min\{4Lp\varepsilon, 1\}$. Apply Lemma G.5 with $R = 2p\varepsilon$ and $H_{\mathrm{in}} = H$. The displayed bound follows from $H = O(p)$ and $\beta\bar{H} = O(p\min\{Lp\varepsilon, 1\})$. □

Lemma G.7 (Packing compression)**.** *For every $z \in \mathcal{X}$, $K_z := \mathcal{X} \cap B(z, 2p\varepsilon)$ obeys*

$$0 \le v_\varepsilon(K_z) - v_\varepsilon(\mathcal{X}) \le f(z) - f^\star. \tag{40}$$

Proof. For compression, retain all points of a given feasible packing already in K_z. For a point outside K_z, its line segment toward z has a length-$2p\varepsilon$ portion in $\mathcal{X}$. Excluding the closed radius-ε intervals around at most $p-1$ previously assigned points removes total length at most $2(p-1)\varepsilon < 2p\varepsilon$. One may therefore assign a point on the segment outside these intervals. Continue through all remaining points. A moved point $y_i = (1-\lambda_i)z + \lambda_i x_i$ satisfies

$$f(y_i) \le f(x_i) + (1-\lambda_i)(f(z) - f(x_i)) \le f(x_i) + f(z) - f^\star.$$

Taking the minimum over the original packings proves (40).

□

Lemma G.8 (Local completion of an embeddable query)**.** *If $q \in \mathcal{X}$ belongs to a feasible p-packing, it belongs to one contained in $\mathcal{X} \cap \overline{B}(q, 2p\varepsilon)$. Every other point u of this completion satisfies $f(u) - f(q) \le \min\{2Lp\varepsilon, 1\}$.*

Proof. Start from any feasible packing containing q and retain q. For each point outside the stated ball, consider the length-$2p\varepsilon$ portion, adjacent to q, of its line segment toward q. This segment belongs to $\mathcal{X}$ by convexity. The other $p-1$ current points have closed radius-ε intervals of total length at most $2(p-1)\varepsilon < 2p\varepsilon$ on it. Replace the point by any location outside these intervals. Each replacement preserves feasibility while putting one more point in the ball. After at most $p-1$ replacements the construction is complete. Lipschitzness and the range $[0,1]$ prove the loss bound. □

Proof of Theorem 3.3. Set $H := 2p-1$, as allowed by Lemma 3.1, and put $s := Lp\varepsilon$, $a := \min\{s,1\}$, $a_\varepsilon := \min\{2s,1\}$, $b_{\mathrm{anc}} := (p-1)a_\varepsilon$, and $\beta := \min\{4s,1\}$. Apply Lemma F.1 with failure probability $\xi := T^{-2}$. Because every planned virtual horizon is at most T, the deterministic envelopes

$$\mathcal{A}_T := \max\{1, A_d(1+\sigma)[\log(eT/\xi)]^4\}, \qquad \mathcal{J}_T := \max\{1, \lceil B_d[\log(eT/\xi)]^3\rceil\}$$

bound, on the primitive's success event, its cumulative excess by $\mathcal{A}_T\sqrt{n}$ and its number of query segments by $\mathcal{J}_T$. If $T \le 4H\mathcal{J}_T$, play any fixed feasible cycle: its regret is at most $T \le 4H\mathcal{J}_T$, within the claimed $\widetilde{O}_{d,\sigma}(p)$ additive term. Otherwise choose

$$n = \left\lfloor \min\left\{ \frac{T}{2p}, \left(\frac{\mathcal{A}_T T}{b_{\mathrm{anc}}}\right)^{2/3} \right\} \right\rfloor,$$

where the second quantity is omitted when $b_{\mathrm{anc}} = 0$. Since $T > 4H\mathcal{J}_T \ge 4p$ and $b_{\mathrm{anc}} \le p-1$, we have $n \ge 1$.

For each query segment at q, Lemma E.1 makes q embeddable and Lemma G.8 constructs the local completion $P(q)$ described in Section 3.1. Transfer to $B(P(q))$ once, record the arrival order Q, and supply each requested observation at q using a new complete cycle of Q. The inball construction permits $Q = P(q)$; the loss and feedback arguments only require that q occurs once per cycle. Initializing from an empty history needs at most one cycle, which is covered by $H \ge p$. The observation sent to the BCO routine is always fresh: observations collected during a transition or earlier cycle are never reused. Conditioning on the physical history before each designated observation verifies the conditional mean-zero and sub-Gaussian assumptions of the primitive.

Stop the BCO procedure if a $(\mathcal{J}_T+1)$st query segment is requested, and use the current feasible cycle thereafter. On the primitive's success event $\mathcal{E}_{\mathrm{BCO}}$ this stop never occurs; its failure probability is at most ξ. In all cases Phase I before such a stop fits within the deterministic budget

$$pn + H\mathcal{J}_T \le T/2 + H\mathcal{J}_T < 3T/4.$$

On $\mathcal{E}_{\mathrm{BCO}}$ all n designated queries finish. Write $v = v_\varepsilon(\mathcal{X})$, and set $z = n^{-1}\sum_{s=1}^n q_s$. Convexity gives

$$f(z) - f^\star \le \frac{1}{n}\sum_{s=1}^n (f(q_s) - f^\star) \le \frac{\mathcal{A}_T}{\sqrt{n}}.$$

Each fresh cycle containing q_s has total loss at most $pf(q_s) + b_{\mathrm{anc}}$. Since $v \ge f^\star$, target-cycle regret in the anchor phase is at most

$$p\sum_{s=1}^n (f(q_s) - f^\star) + b_{\mathrm{anc}} n \le p\mathcal{A}_T\sqrt{n} + b_{\mathrm{anc}} n.$$

All anchor transitions together cost at most $H\mathcal{J}_T$ in regret.

Now use $K_z = \mathcal{X} \cap \overline{B}(z, 2p\varepsilon)$. Lemma G.7 bounds the comparator increase by $\mathcal{A}_T/\sqrt{n}$ on $\mathcal{E}_{\mathrm{BCO}}$. Lemma G.2 gives a first-entry budget H, internal budget $\bar{H} := 3p \le 3H$, and oscillation β. Conditional on every completed anchor history, the fresh-data bound of Lemma G.5 is valid. With $R = 2p\varepsilon$, its contribution is at most

$$\widetilde{O}_{d,\sigma}\Big(\sqrt{T} + s^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} + s^{\frac{d}{d+1}} (\beta H)^{\frac{1}{d+1}} T^{\frac{d}{d+1}} + H\Big)$$

after absorbing numerical constants. The restriction of all remaining rounds to K_z adds at most $\mathcal{A}_T T/\sqrt{n}$. The BCO failure event contributes at most ξT, because every physical-round excess is at most one. Hence

$$\begin{aligned} R_T^{\varepsilon} \le{}& H\mathcal{J}_T + p\mathcal{A}_T\sqrt{n} + b_{\mathrm{anc}} n + \frac{\mathcal{A}_T T}{\sqrt{n}} + \xi T \\ &+ \widetilde{O}_{d,\sigma}\Big(\sqrt{T} + s^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} + s^{\frac{d}{d+1}} (\beta H)^{\frac{1}{d+1}} T^{\frac{d}{d+1}} + H\Big). \end{aligned}$$

Integer rounding and the two cases in the choice of n give

$$p\mathcal{A}_T\sqrt{n} + b_{\mathrm{anc}} n + \frac{\mathcal{A}_T T}{\sqrt{n}} \le C\left(\mathcal{A}_T\sqrt{pT} + \mathcal{A}_T^{2/3} b_{\mathrm{anc}}^{1/3} T^{2/3}\right).$$

To absorb the auxiliary-action term, put $C_0 = \mathcal{A}_T\sqrt{p}$, $u = a\sqrt{T}$, and $e = d/(d+2) \ge 1/3$. Since $a_\varepsilon \le 2a$ and hence $b_{\mathrm{anc}} \le 2pa$,

$$\mathcal{A}_T^{2/3} b_{\mathrm{anc}}^{1/3} T^{2/3} \le 2^{1/3}\sqrt{T}\, C_0^{2/3} u^{1/3} \le 2^{1/3}\left(C_0\sqrt{T} + C_0^{1-e}\sqrt{T}\, u^e\right).$$

For $u \le C_0$ the first term bounds the left side; for $u \ge C_0$ use $e \ge 1/3$ and the second term. Since

$$C_0^{1-e}\sqrt{T}\, u^e = (\mathcal{A}_T^2 p)^{1/(d+2)} a^{d/(d+2)} T^{(d+1)/(d+2)} \le (\mathcal{A}_T^2 p)^{1/(d+2)} s^{d/(d+2)} T^{(d+1)/(d+2)},$$

while $\beta H \le 4(2p-1)a$. Substituting these inequalities, $\mathcal{J}_T = \widetilde{O}_d(1)$, and $\mathcal{A}_T = \widetilde{O}_{d,\sigma}(1)$ gives a bound with terms $\sqrt{pT}$, A_2, B, and p, where

$$A_2 := p^{\frac{1}{d+2}} s^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}, \qquad B := s^{\frac{d}{d+1}} \big(p \min\{s, 1\}\big)^{\frac{1}{d+1}} T^{\frac{d}{d+1}}.$$

The additive p is at most $\sqrt{pT}$. If $0 < s \le 1$, then $B/A_2 = (p/T)^{1/((d+1)(d+2))} s^{2/(d+2)} \le 1$; for $s = 0$ both terms vanish. If $s > 1$, then $B/A_2 = (ps^d/T)^{1/((d+1)(d+2))}$. Whenever this ratio exceeds one, $A_2/T = (ps^d/T)^{1/(d+2)} > 1$, so the trivial cap by T absorbs B. This cap holds pathwise because $f \in [0,1]$. Finally, $\sqrt{pT} + A_2 = \sqrt{pT}[1 + (L\varepsilon\sqrt{pT})^{d/(d+2)}]$. Lemma F.1 bounds A_d, B_d by fixed polynomials in d, and the tuning and absorption steps above introduce at most further polynomial factors in d. These can all be included in a single fixed polynomial $P(d)$, proving (6). □

Diameter-based implementation. Suppose $D \ge p\varepsilon$ and set

$$H_D := \left\lceil 3p + \frac{4p^2}{D/\varepsilon - p + 1} \right\rceil.$$

The diameter-parking set in the proof of Lemma 3.2 has at least $p+1$ points. Any $q \in \mathcal{X}$ conflicts with at most two of them, so at least $p-1$ remain compatible with q. Thus every query is embeddable, and Lemma G.8 compresses its packing locally. Lemma 3.2 supplies a uniform transition budget H_D for target sets. After each anchor transfer to $B(P(q))$, keep its arrival order Q fixed for the query segment and return only the fresh observation at q in each cycle; q need not be the first physical action. Refinement likewise cycles the arrival order of each selected target set. Initialization from an empty history uses $p \le H_D$ rounds; every subsequent transfer starts from a chronological state of length p. Applying the same capped two-phase construction as above, with Lemma G.2 for transfers inside K_z, gives

$$\begin{aligned} R_T^{\varepsilon} \le \min\Bigg\{ T,\ \widetilde{O}_{d,\sigma}\Bigg(& \sqrt{pT} + p^{\frac{1}{d+2}} (Lp\varepsilon)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} \\ & + (Lp\varepsilon)^{\frac{d}{d+1}} \big(H_D \min\{Lp\varepsilon, 1\}\big)^{\frac{1}{d+1}} T^{\frac{d}{d+1}} + H_D \Bigg) \Bigg\}. \end{aligned}$$

H Proof of the General-Period Lower Bound

Let

$$\mathfrak{R}_T := \inf_{\pi} \sup_{f \in \mathcal{F}} R_T^{\varepsilon}(\pi, f), \qquad r_p := p^{1/d}.$$

H.1 Volume relocation and cheap transitions

Lemma H.1 (Template radius)**.** *The minimum*

$$S = \min_{\|q_i - q_j\| \geq 1} \sum_{i=1}^{p} \|q_i\|^2$$

is attained, and every minimizer satisfies

$$\max_i \|q_i\| \leq r_p, \qquad \frac{p-1}{2} \leq S \leq p r_p^2. \tag{41}$$

Every minimizer has centroid $p^{-1}\sum_i q_i = 0$. Moreover, a maximally distant point of a minimizing tuple has radius $r_\star \geq 1/2$.

Proof. Closedness and coercivity give existence. Subtracting the centroid from each point of a feasible tuple preserves all pairwise distances and decreases its squared-norm sum by p times the squared norm of the centroid. Thus every minimizer has centroid zero. Given any $p-1$ points, their closed unit balls have total volume at most $(p-1)v_d$, strictly smaller than $\mathrm{vol}(B(0, r_p)) = p v_d$. There is therefore a point in this latter ball strictly more than one unit away from all of them. If a minimizing tuple had a point of norm greater than r_p, replacing it by this point would strictly reduce the objective. This gives the upper bound on S. Finally,

$$\sum_{i<j} \|q_i - q_j\|^2 = p \sum_i \|q_i\|^2 - \left\| \sum_i q_i \right\|^2 \leq p \sum_i \|q_i\|^2$$

gives $S \geq (p-1)/2$, so $r_\star^2 \geq S/p \geq (p-1)/(2p) \geq 1/4$. □

Lemma H.2 (Cheap bridges and affine comparators)**.** *If $\varepsilon \leq 1/[2(2p)^{1/d}]$, any admissible history can be connected to any prescribed feasible p-cycle using at most $p-1$ intermediate buffer actions before that cycle. For every unit vector u, there is a feasible p-packing with average u-projection $-1/2$.*

Proof. The unit ball has inradius $r_0 = 1$, so the assumed radius is exactly (3). Lemma E.1 supplies the $p-1$ intermediate actions for any target cycle.

Choose a minimizer $Q^\star = (q_1^\star, \ldots, q_p^\star)$ from Lemma H.1 and set $x_i = -u/2 + \varepsilon q_i^\star$. The points are pairwise ε-separated, and

$$\|x_i\| \leq \frac{1}{2} + r_p \varepsilon \leq \frac{1}{2} + \frac{1}{2 \cdot 2^{1/d}} < 1.$$

They therefore lie in $\mathcal{X} = B_2^d$. Since $Q^\star$ has centroid zero, $p^{-1}\sum_i \langle u, x_i \rangle = -1/2$. □

H.2 Affine alternatives

Lemma H.3 (Affine dimension lower bound)**.** *In the radius range of Theorem 4.1,*

$$\mathfrak{R}_T \geq \frac{1}{32} \min\{T, d\sqrt{T}\}. \tag{42}$$

Also $\mathfrak{R}_T \geq \sqrt{T}/16$.

Proof. Fix an admissible policy. Let V be uniform on $\{-1, 1\}^d$ and set

$$\delta = \min\{1/(4\sqrt{d}), \sqrt{d}/(8\sqrt{T})\}, \qquad f_V(x) = \tfrac{1}{2} + \delta \langle V, x \rangle.$$

These losses belong to $\mathcal{F}$. By Lemma H.2, $v_\varepsilon(f_V) \le 1/2 - \delta\sqrt{d}/2$. Let $\mathcal{H}_{t-1}$ include the learner's independent randomization and put $m_t = \mathbb{E}[V \mid \mathcal{H}_{t-1}]$. We claim

$$\mathbb{E}\,\|m_t\|^2 \le 2\delta^2(t-1). \tag{43}$$

For a bit j, Pinsker's inequality applied to its posterior versus its uniform prior gives

$$\mathbb{E}[m_{t,j}^2] \le 2I(V_j; \mathcal{H}_{t-1}) \le 2I(V_j; \mathcal{H}_{t-1} \mid V_{-j}).$$

For fixed V_{-j}, the latter mutual information is the Jensen–Shannon divergence of the two interaction laws. Convexity of $-\log$ bounds it by one quarter of the sum of the two directed KL divergences. The adaptive Gaussian likelihood identity gives these KL divergences as $2\delta^2\mathbb{E}_\pm \sum_{s<t} x_{s,j}^2$. Averaging over V_{-j} yields

$$I(V_j; \mathcal{H}_{t-1} \mid V_{-j}) \le \delta^2 \mathbb{E} \sum_{s<t} x_{s,j}^2.$$

Summing in j and using $\|x_s\| \le 1$ proves (43). Hence

$$\mathbb{E}\langle V, x_t\rangle = \mathbb{E}\langle m_t, x_t\rangle \ge -\sqrt{2}\delta\sqrt{t-1}.$$

The average regret is at least

$$\frac{\delta\sqrt{d}}{2}T - \sqrt{2}\delta^2 T\sqrt{T} \ge \frac{\delta\sqrt{d}}{4}T = \min\{T/16, d\sqrt{T}/32\}.$$

This proves (42).

For the second claim take $f_\pm(x) = 1/2 \pm x_1/(4\sqrt{T})$ and the reference $f_0 = 1/2$. Both comparator values are at most $1/2 - 1/(8\sqrt{T})$. For $Z = \sum_t x_{t,1} \in [-T, T]$, the reference-to-alternative KL is at most $1/32$, so $|\mathbb{E}_\pm Z - \mathbb{E}_0 Z| \le T/4$. Averaging the two regrets gives at least $\sqrt{T}/16$. □

H.3 Radial baseline and localized bumps

Fix a minimizer $Q^\star$ in Lemma H.1, with its first point of radius $r_\star$. Put $B = 2r_p\varepsilon \le 2^{-1/d} < 1$ and define

$$h_B(r) = \begin{cases} r^2/(4B), & r \le B, \\ r/2 - B/4, & r > B, \end{cases} \qquad f_0(x) = \tfrac{1}{4} + h_B(\|x\|), \qquad b_\varepsilon = \tfrac{1}{4} + \frac{\varepsilon S}{8r_p p}. \tag{44}$$

The profile is convex, nondecreasing, and $1/2$-Lipschitz. Since $B < 1$, its values on the unit ball lie in $[0, 1]$, so $f_0 \in \mathcal{F}$.

Lemma H.4 (Exact benchmark)**.** *The periodic benchmark of f_0 is b_ε, and every admissible sequence satisfies*

$$\sum_{t=1}^{T} f_0(x_t) \ge Tb_\varepsilon - C\varepsilon, \qquad C = \frac{S}{8r_p} \le \frac{pr_p}{8}, \qquad C\varepsilon \le p/16. \tag{45}$$

If $p \mid T$, no remainder is needed.

Proof. Given any feasible packing, replace any point outside $B(0, B)$ by a point in $B(0, r_p\varepsilon)$ outside the closed ε-balls of the other $p-1$ points. Volume guarantees such a point, as in Lemma H.1; it lies in $\mathcal{X}$ and strictly decreases f_0. Repeat until all points are in the quadratic region. Their squared norms sum to at least $\varepsilon^2 S$, so the original packing has average at least b_ε.

The template $(\varepsilon q_i^\star)_i$ lies within radius $r_p\varepsilon = B/2$, is feasible, and has average loss exactly b_ε. Thus it attains the benchmark minimum. Partition an admissible sequence into complete blocks of p actions and a remainder of fewer than p actions. Each block costs at least pb_ε and each remainder action at least $1/4$, which yields (45). □

Lemma H.5 (Separated sphere centers)**.** *If $s \le R/4$, the sphere of radius R in $\mathbb{R}^d$ contains a finite set of points at pairwise distances greater than $2s$, with*

$$M \ge (R/(2s))^{d-1}. \tag{46}$$

Proof. Take a maximal such set. Its chordal caps of radius $2s$ cover the sphere. Put $m = d - 1$ and let v_m be the volume of the unit m-ball. Rotate one cap center to the north pole. A point in the cap has height at least $R - 2s^2/R \geq 7R/8$, and its projection to the equatorial plane has norm at most $2s$. The graph-area Jacobian is at most $8/7$, so the cap surface area is at most $(8/7)v_m(2s)^m$. Each hemisphere projects onto the radius-R m-ball with area Jacobian at least one, so the sphere's total area is at least $2v_m R^m$. Covering implies $M \geq (7/4)(R/(2s))^m$, hence (46). □

Choose

$$0 < \Delta \leq \frac{\varepsilon}{128 r_p}, \qquad s = \sqrt{2 r_p \varepsilon \Delta}, \qquad R = \varepsilon r_\star. \tag{47}$$

Then $s \leq \varepsilon/8 \leq R/4$. Lemma H.5 and $r_\star \geq 1/2$ yield

$$M \geq \left(\frac{\varepsilon}{32 r_p \Delta} \right)^{(d-1)/2}. \tag{48}$$

For these centers v_j, define

$$\phi(z) = (1 - \|z\|^2)_+^3, \qquad f_j(x) = f_0(x) - \frac{\Delta}{24} \phi((x - v_j)/s). \tag{49}$$

The open supports $A_j = B(v_j, s)$ are disjoint. Indeed, $R \leq r_p \varepsilon = B/2$ and $s \leq \varepsilon/8 \leq B/16$, so they lie inside the quadratic region $B(0, B) \subset B_2^d$.

Lemma H.6 (Validity and comparator improvement)**.** *Each $f_j \in \mathcal{F}$, and*

$$v_\varepsilon(f_j) = b_\varepsilon - \Delta/(24p). \tag{50}$$

Proof. For $u = \|z\|^2 < 1$,

$$\nabla \phi = -6(1-u)^2 z, \qquad \nabla^2 \phi = -6(1-u)^2 I + 24(1-u) z z^\top \preceq 6I.$$

The function and its first two derivatives extend continuously by zero at the support boundary. On A_j,

$$\nabla^2 f_j \succeq I/(2B) - (\Delta/(24 s^2)) 6I = I/(4B),$$

and outside A_j, $f_j = f_0$. Hence f_j is globally convex. Its Lipschitz constant is at most $1/2 + (1/4)\sqrt{\Delta/B} \leq 3/4$, and its range remains in $[0, 1]$.

Since $2s < \varepsilon$, a feasible packing has at most one point in A_j, so its average loss is at least $b_\varepsilon - \Delta/(24p)$. Rotate $\varepsilon Q^\star$ so its first point is v_j. The rotated template is feasible, has baseline average b_ε, and only its first point is affected by the bump. It therefore attains the matching value $b_\varepsilon - \Delta/(24p)$. □

H.4 Blocking-aware change of measure

Let P_0, P_j be the full interaction laws for a fixed admissible policy and put $N_j = \sum_{t=1}^T \mathbf{1}\{x_t \in A_j\}$. Because each A_j has diameter less than ε, any two visits to it must be separated by at least p rounds. Thus

$$0 \leq N_j \leq \lceil T/p \rceil. \tag{51}$$

The Gaussian likelihood identity and Pinsker's inequality give

$$\mathrm{KL}(P_0 \| P_j) \leq \frac{\Delta^2}{2 \cdot 24^2} \mathbb{E}_0 N_j, \tag{52}$$

$$\mathbb{E}_j N_j \leq \mathbb{E}_0 N_j + \lceil T/p \rceil \frac{\Delta}{48} \sqrt{\mathbb{E}_0 N_j}. \tag{53}$$

Since the supports are disjoint,

$$\frac{1}{M} \sum_j \mathbb{E}_j N_j \leq \frac{T}{M} + \lceil T/p \rceil \frac{\Delta}{48} \sqrt{T/M}. \tag{54}$$

For $T \geq p$, the conditions

$$M \geq 4p, \qquad \Delta \sqrt{T/M} \leq 1 \tag{55}$$

imply that the right side of (54) is at most $T/(4p) + T/(24p) = 7T/(24p) < T/(2p)$, using $\lceil T/p \rceil \leq 2T/p$. Combining (45) and (50) gives

$$\mathfrak{R}_T \geq \frac{\Delta T}{48p} - C\varepsilon. \tag{56}$$

All actual observations, including transition observations, are included in the likelihood identity.

H.5 Choice of parameters and completion of the proof

Set

$$c_0 = \frac{1}{128 r_p \chi_{d,p}}, \qquad \Delta = c_0 \varepsilon^{(d-1)/(d+3)} T^{-2/(d+3)}. \tag{57}$$

If $\varepsilon^2 T \geq 1$, then $\Delta \leq c_0 \varepsilon \leq \varepsilon/(128 r_p)$. With $m = d-1$, (48) gives

$$M \geq (4\chi_{d,p})^{m/2} (\varepsilon^2 T)^{m/(m+4)} \geq 4p,$$

because $(4\chi_{d,p})^{m/2} \geq 4p$ by definition of $\chi_{d,p}$. Also

$$\Delta \sqrt{T/M} \leq c_0 (4\chi_{d,p})^{-m/4} \leq 1.$$

Thus for $T \geq p$ and $\varepsilon^2 T \geq 1$,

$$\mathfrak{R}_T \geq b \varepsilon^{\frac{d-1}{d+3}} T^{\frac{d+1}{d+3}} - C\varepsilon, \qquad b = \frac{c_0}{48p}. \tag{58}$$

We claim, uniformly for every $T \geq 1$,

$$\mathfrak{R}_T \geq \frac{\sqrt{T}}{64} + \frac{b}{2} \varepsilon^{\frac{d-1}{d+3}} T^{\frac{d+1}{d+3}} = \frac{\sqrt{T}}{64} + \frac{\varepsilon^{\frac{d-1}{d+3}} T^{\frac{d+1}{d+3}}}{12288 p^{1+1/d} \chi_{d,p}}. \tag{59}$$

If $T \geq 4p^2$ and $\varepsilon^2 T \geq 1$, use $C\varepsilon \leq p/16 \leq \sqrt{T}/32$ in (58) and average with $\mathfrak{R}_T \geq \sqrt{T}/16$. In the remaining cases, either $\varepsilon^2 T < 1$, in which case $\varepsilon^{\frac{d-1}{d+3}} T^{\frac{d+1}{d+3}} \leq \sqrt{T}$, or $T < 4p^2$. In the latter case $\varepsilon < 1/2$, so $\varepsilon^2 T < p^2$ and

$$\frac{\varepsilon^{\frac{d-1}{d+3}} T^{\frac{d+1}{d+3}}}{\sqrt{T}} = (\varepsilon^2 T)^{(d-1)/(2(d+3))} \leq p.$$

Since $b/2 = c_0/(96p) \leq 1/(96p)$, the right side of (59) is at most $(1/64 + 1/96)\sqrt{T} < \sqrt{T}/16$, so the simple affine bound proves it.

Finally average (59) with (42) and discard the extra positive $\sqrt{T}/128$ term. This yields the first inequality in (8). The second follows because $\min\{\sqrt{T}, d\} \geq 1$ and $24576 p^{1+1/d} \chi_{d,p} \geq 64$.