Stopped Experiments: Minimax Regret, Aggregate-Feedback Bandits, and Reinforcement Learning

Yunbei Xu
National University of Singapore
yunbei@nus.edu.sg

Abstract

We introduce stopped experiments as a local complexity for repeated decisions with a fixed unknown environment. An experiment adaptively queries until posterior Bayes risk contracts by a constant factor or the experiment is abandoned; its cost is the risk scale times incurred regret, divided by its contraction probability. Uniform stopped bounds yield regret upper bounds, while pointwise stopped lower bounds yield calibrated finite-horizon lower bounds.

The conversion gives a new minimax result for stochastic top-k combinatorial bandits with aggregate feedback. With $2k$ independent Gaussian arms and only the selected sum observed, minimax regret is $\widetilde{\Theta}(\min\{kT, k^{3/2}\sqrt{T}\})$. The lower bound follows from the stopped cost of learning k paired signs and matches the Hadamard-design upper bound. This is where stopping is essential: fixed-confidence sample complexity alone does not charge the regret of the identifying subsets.

The framework also embeds episodic and preference-based reinforcement learning by treating a policy, or a policy pair, as a macro-action and a trajectory or preference as its observation. For compact Gaussian loss classes we prove an exact exponent relation. In bandit convex optimization we further separate observation-adaptive experiments from nonadaptive and fresh-direction designs on a dimple family.

1 Introduction

How much regret must be spent before observations make a decision problem meaningfully easier? Pure-exploration complexity counts observations; cumulative regret prices the decisions used to obtain them. Neither, by itself, identifies the local object that connects the two. The missing object must retain partial learning, allow later queries to depend on early observations, and charge only until the residual decision problem becomes easier. This paper proposes such an object and shows that it can close a minimax regret gap.

Stopped experiments. We begin with a fixed latent environment ω. An action a incurs a nonnegative decision gap $\Delta_\omega(a)$ and produces an observation. If Q_t is the posterior after t observations, its Bayes decision risk is

$$r(Q_t) = \inf_a \mathbb{E}_{Q_t} \Delta_\omega(a).$$

At a state where this risk is of order b, an adaptive experiment runs until it is abandoned or first reaches $r(Q_t) \le \gamma b$. If N is its duration and M the first-hit event, we define the pointwise stopped cost by

$$\mathsf{SC}_b(Q) = \inf_{\text{stopped experiments}} \frac{b\,\mathbb{E}_Q R_N}{\mathbb{P}_Q(\mathsf{M})}. \tag{1}$$

Thus $\mathsf{SC}_b(Q)$ is scaled regret per unit probability of a decision-relevant risk reduction. No dimension factor, random target, or information budget is included. A separate query charge is unnecessary: every query before the hit has expected gap at least γb, so $\mathbb{E}R_N \geq \gamma b\, \mathbb{E}N$.

A two-sided relation. Let $\mathsf{SC}^{\mathrm{pt}}(d, B)$ be the largest statewise value over risk scales $B^{-1} \leq b \leq 1$, and let $\mathsf{SC}^{\mathrm{unif}}(d, B)$ be the smallest bound attained by one measurable state-dependent experiment rule. The finite-horizon bounds have the form

$$\mathfrak{R}^{\mathrm{B}}(d, T) \lesssim_\gamma 1 + \sqrt{\mathsf{SC}^{\mathrm{unif}}(d, T)T}, \tag{2}$$

$$\mathfrak{R}^{\mathrm{B}}(d, n) \gtrsim_\gamma \sqrt{cn} \qquad \text{when } n \asymp_\gamma c/b^2, \tag{3}$$

for every $c < \mathsf{SC}^{\mathrm{pt}}(d, B)$, at some initial posterior and $b \in [B^{-1}, 1]$. The upper bound joins stopped experiments over geometric risk scales. For the lower bound, a policy that has not yet reduced risk pays at least γb per round, whereas reaching the lower-risk set with probability p costs at least cp/b. Balancing these two possibilities gives (3).

In terms of polynomial dimension exponents, the same theorem gives

$$\frac{\alpha^{\mathrm{pt}}_{\mathrm{stop}}}{2} \leq \alpha^{\mathrm{B}}_{\mathrm{reg}} \leq \frac{\alpha^{\mathrm{unif}}_{\mathrm{stop}}}{2}. \tag{4}$$

If the statewise experiments can be chosen measurably, the two stopped exponents agree and both inequalities are equalities. If Bayes and minimax regret also agree, then

$$\alpha^{\mathrm{M}}_{\mathrm{reg}} = \alpha^{\mathrm{B}}_{\mathrm{reg}} = \frac{\alpha_{\mathrm{stop}}}{2}.$$

These are exponent statements, not same-horizon equalities. They require nonnegative additive gaps, posterior sufficiency, and persistence of the same decision problem. They do not require convex losses or Gaussian observations.

A concrete minimax consequence. Consider $2k$ independent stochastic arms, choose exactly k each round, and observe only their aggregate reward. This full-bandit model occurs when individual attribution is unavailable because of aggregation, privacy, routing, or experimental design. The best known efficient algorithm has distribution-free regret $\widetilde{O}(k\sqrt{nT})$ (Rejwan and Mansour, 2020); for unrestricted top-k actions, the dependence on k was not matched by the standard semi-bandit lower bound.

We calculate the stopped cost of a paired $n = 2k$ family. A query is a linear measurement of k independent signs with noise variance k. Constant-factor posterior-risk contraction costs $\Theta(k/\varepsilon^2)$ queries, but, crucially, every pre-contraction query also incurs decision loss of order $k\varepsilon$. The stopped conversion preserves both facts and gives

$$\mathfrak{R}^{\star}_{2k,k}(T) = \widetilde{\Theta}\Big(\min\{kT, k^{3/2}\sqrt{T}\}\Big).$$

The lower bound is not obtained by appending a PAC lower bound to an explore–then–commit calculation. Its proof must stop an arbitrary regret policy at the first posterior-risk contraction and charge the regret on both the hitting and non-hitting branches. This is the first application in the paper for which the general conversion itself closes the rate. It also addresses the tightness discussion that has remained open since Rejwan and Mansour (2020).

Episodic reinforcement learning. Consider a fixed unknown finite-horizon MDP M that is reset from an initial distribution ρ at the beginning of every episode. Treat an entire within-episode policy π as the action and the resulting trajectory as its observation. The episode gap is

$$\Delta_M(\pi) = V_M^\star(\rho) - V_M^\pi(\rho) \geq 0.$$

The posterior risk

$$r(Q) = \inf_\pi \mathbb{E}_Q \Delta_M(\pi)$$

has exactly the persistence used above. Hence the stopped framework applies at the episode level. This macro-decision reduction includes standard episodic pseudo-regret. The finite-model corollary below gives automatic measurable selection and minimax closure when the model and trajectory alphabets are finite (Azar et al., 2017; Foster et al., 2021). Continuing RL is different. If the physical state is not reset, $r(Q_t, S_t)$ need not be a supermartingale: the next state may be harder even when no uncertainty is resolved. Such problems need an additional persistence condition or the full stateful framework of Xu (2026a).

Preferences rather than numerical rewards. Preference-based reinforcement learning replaces numerical trajectory rewards by comparisons that may better represent human judgments (Xu et al., 2020). At an episode boundary, choose a pair (π, π'), generate trajectories, and observe a stochastic preference. The macro-action gap is the average value loss of the two policies. The posterior over the fixed MDP and preference model is again sufficient, and the same problem recurs after reset. Thus a regret-efficient preference experiment has exactly the stopped form. This distinction matters for human or language-model feedback: a PAC comparison bound counts labels, whereas the stopped cost also prices how poor the displayed trajectories are while the reward model is being learned.

Why bandit convex optimization? Stochastic bandit convex optimization (BCO) is a particularly revealing case. An unknown continuous convex loss $F : C \to [0, 1]$ on a compact convex body $C \subset \mathbb{R}^d$ is fixed. The learner chooses X_t, observes

$$O_t = F(X_t) + \xi_t, \qquad \xi_t \overset{\text{iid}}{\sim} N(0, 1),$$

and pays $F(X_t) - \min_C F$. The same scalar query is therefore both an experiment and a costly decision. Moreover, convex geometry links near-optimal regions to the locations and directions of informative probes, and the useful direction of a later probe may depend on earlier observations.

BCO also has a concrete unresolved dimension gap. In the nontrivial long-horizon regime, the best general upper bound is $\widetilde{O}(d^{3/2}\sqrt{T})$ (Lattimore and György, 2023; Fokkema et al., 2024). The recent lower bound $\widetilde{\Omega}(d^{5/4}\sqrt{T})$ of Rajaraman (2026) is the first full-class lower bound with dimension dependence strictly larger than the linear-bandit rate. Carpentier (2025) develops a complementary simple-regret formulation and a direct localization method for noisy zeroth-order convex optimization. Neither cumulative-regret endpoint is known to be sharp.

For every Borel class of bounded continuous losses on a compact action space under Gaussian feedback, not necessarily convex, we prove that the pointwise experiments can be chosen measurably and that worst-prior Bayes regret equals minimax regret. Thus

$$\alpha_{\text{reg}} = \frac{\alpha_{\text{stop}}}{2}$$

without an information-matching or nondegeneracy assumption. For the full convex class, the known regret bounds are equivalent to

$$\frac{5}{2} \leq \alpha_{\text{stop}} \leq 3.$$

Convexity is used here to define the BCO class, invoke its known bounds, and construct the geometric example below. It is not used in the stopped reduction.

Observation-adaptive experiment design. For the dimple losses of Bakhtiari et al. (2025), at $\varepsilon = d^{-1}$ we prove

$$\mathsf{SC}_b^{[1]},\ \mathsf{SC}_b^{\mathrm{fresh}} = \Omega(d^3), \qquad \mathsf{SC}_b = \widetilde{O}(d^2).$$

Under the spherical prior, every nonadaptive batch needs size $\Omega(d^5)$, and every stopped policy restricted to fresh directions—independent Haar directions with rays starting at the origin—needs expected duration $\Omega(d^5)$, for a constant-probability contraction. The unrestricted experiment first obtains a weak direction, uses tangent differences chosen from the observed direction to amplify it, and then contracts the angular error geometrically. It uses $\widetilde{O}(d^4)$ queries and yields $\widetilde{O}(\min\{T/d, d\sqrt{T}\})$ regret on this family. This is a separation between experiment classes, not a lower bound against all algorithms for the dimple family. Its role is to show that the stopped experiment can use geometry revealed by early observations to guide later queries, beyond what is available from its first query alone.

Information and statistical testing. The stopped cost is a decision cost, not an information measure. Mutual information, Le Cam, Fano, Assouad, and interactive change of measure remain useful for bounding the cost of reaching the lower-risk set; our fresh-direction lower bound uses mutual information in this way. Maximal information gain may likewise help bound $\mathsf{SC}_b(Q)$ in a particular model. It need not enter the definition, because the full posterior records partial progress and governs later queries and stopping.

1.1 Related work

Blackwell comparison and the Bayesian theory of experiments evaluate observations through their effect on decision risk (Blackwell, 1953; Lindley, 1956; DeGroot, 1962). Classical sequential design optimizes both experiments and stopping (Chernoff, 1959; Naghshvar and Javidi, 2013). Here the stopping set is the first constant-factor reduction in Bayes decision risk, and the price is the regret incurred on the way to that set.

Information ratios, AIR, and DEC turn one-step regret–information or decision–estimation comparisons into sequential regret bounds (Russo and Van Roy, 2016, 2018; Lattimore and György, 2021; Foster et al., 2021, 2022, 2023; Xu and Zeevi, 2025; Neu et al., 2024). Chained ratios organize information across action scales (Gouverneur et al., 2024). These methods may re-optimize after every observation. Our distinction concerns the local object being optimized: the stopped cost evaluates an entire adaptive experiment, including how observations determine later queries, rather than assigning a value to one query at a time. Interactive Le Cam, Fano, and Assouad inequalities handle adaptive transcript laws directly (Chen et al., 2024).

Near-optimal targets and localization appear in rate–distortion, satisficing, and scale-sensitive ratio analyses (Dong and Van Roy, 2018; Russo and Van Roy, 2022; Arumugam and Van Roy, 2021; Bubeck and Sellke, 2023). A distinct modern line develops data-dependent localization through pointwise complexities (Li and Xu, 2026; Xu, 2026c,b). There, "pointwise" refers to localization around a realized input, distribution, or geometric point. Here it means that the best stopped experiment may be chosen separately at each posterior state; the implementation question is whether those choices form one measurable rule. The two notions are complementary. Polynomial gains from observations that orient later probes also arise in structured best-arm identification (Maiti et al., 2026).

The Bellman-sufficient framework of Xu (2026a) is broader: it allows physical states, contexts, and other history-dependent variables, and links same-horizon upper and lower bounds through a chosen information index. Its two sides meet when the corresponding information scales agree. The present paper studies the narrower repeated-decision case in which the posterior is sufficient and the nonnegative gap persists from one round to the next. Here the same stopped cost enters both directions, but the lower bound is evaluated at a horizon calibrated to the risk scale. On posterior space, the stopped experiment is a stochastic shortest-path problem (Bertsekas and Tsitsiklis, 1991); the new part is its two-sided relation to cumulative regret.

2 Stopped experiments for repeated decisions

2.1 The repeated latent decision problem

We first work beyond convex optimization, but within a repeated decision problem with a fixed latent environment. This is the static-state specialization of the Bellman-sufficient framework of Xu (2026a): the pre-action state is the current posterior, the same action problem recurs, and the action gap below is the nonnegative one-step Bellman loss. General stateful control may also admit a Bellman-sufficient state, but it need not have the persistence property used to join successive risk scales.

Let Ω, $\mathcal{A}$, and $\mathcal{O}$ be standard Borel environment, action, and observation spaces, and let $\Delta_\omega(a) \in [0,1]$ be a nonnegative action gap. Querying a produces an observation in $\mathcal{O}$ from a kernel $P_\omega(\cdot \mid a)$. The environment $\omega \sim Q$ is drawn once and remains fixed. A policy chooses A_t from the pre-action history H_{t-1}, observes O_t, and incurs $\Delta_\omega(A_t)$. We include all independent exogenous policy randomization in H_0 and write $\mathcal{H}_t = \sigma(H_t)$ for the completed history filtration.

For a posterior Q, define the Bayes risk

$$r(Q) = \inf_{a \in \mathcal{A}} \mathbb{E}_Q \Delta_\omega(a).$$

We assume the posterior is a sufficient controlled state, r is measurable, and measurable ϵ-Bayes actions exist. These are standard in compact Borel models and are verified below for compact Gaussian loss classes. The theorem applies to static latent-model decision problems of the DEC/DMSO type. Neither the action space nor the map $a \mapsto \Delta_\omega(a)$ is assumed to be convex. The nonnegative gap assumption is substantive: with signed one-step increments, stopping can remove compensating future terms and elapsed time need not be paid by regret. Such problems require a separate query charge or a lower-bounded Bellman loss.

The repeated-decision structure implies that posterior Bayes risk is a nonnegative supermartingale. Indeed,

$$\mathbb{E}[r(Q_{t+1}) \mid \mathcal{H}_t] \le \inf_{a \in \mathcal{A}} \mathbb{E}[\mathbb{E}\{\Delta_\omega(a) \mid \mathcal{H}_{t+1}\} \mid \mathcal{H}_t] = r(Q_t). \tag{5}$$

This fact, rather than convexity or a particular observation law, controls returns to a risk scale in the upper bound.

Example 2.1 (Episodic reinforcement learning)**.** *Let M be a fixed unknown MDP with episode length H, rewards in $[0,1]$, and a common initial-state distribution ρ. One macro-action is a complete contingent policy π for an episode, and its observation is the resulting trajectory. After normalizing by H, put*

$$\Delta_M(\pi) = \frac{V_M^\star(\rho) - V_M^\pi(\rho)}{H} \in [0,1].$$

The framework then measures episode pseudo-regret; a realized reward difference need not be nonnegative. At episode boundaries,

$$r(Q) = \inf_{\pi} \mathbb{E}_Q \Delta_M(\pi)$$

satisfies (5). For a finite model class, finite state, action, and reward alphabets, and a finite episode horizon, the required measurable choices and finite-horizon minimax equality reduce to finite-dimensional statements. Stopping takes place between episodes.

This reduction relies on the reset to ρ. In continuing control, the physical state changes and $r(Q_t, S_t)$ need not be a supermartingale. The next state can be intrinsically harder even without new uncertainty. One must then assume the corresponding risk-persistence inequality or use a stateful Bellman formulation such as Xu (2026a).

Example 2.2 (Preference-based episodic reinforcement learning)**.** *Let M again reset from ρ, but suppose the learner observes a binary preference between two sampled trajectories rather than their numerical returns. A macro-action is a policy pair (π, π'); its observation contains the two trajectories and the preference label generated by a fixed comparison kernel. Normalize the query gap as*

$$\Delta_M(\pi, \pi') = \frac{2V_M^\star(\rho) - V_M^\pi(\rho) - V_M^{\pi'}(\rho)}{2H}.$$

This is nonnegative and charges the quality of both trajectories shown to the evaluator. A posterior on the MDP and comparison kernel is sufficient, and its Bayes risk is a supermartingale across episode pairs. The stopped theorems therefore apply verbatim when the model, policy, trajectory, and preference alphabets are finite.

This embedding is agnostic to the stochastic link between latent returns and preference probabilities; identifiability is reflected in whether the stopped cost is finite. It complements PAC preference-query guarantees such as Xu et al. (2020): comparison count controls only duration, whereas stopped cost also controls the value loss incurred in producing the comparisons. If preferences are generated by a changing evaluator or the learned reward is deployed in a different MDP, the fixed-environment reduction requires an augmented latent state.

Fix

$$0 < \gamma < \vartheta < 1. \tag{6}$$

A pair $s = (Q, b)$ is *active* when

$$\vartheta b < r(Q) \le b.$$

Starting from an active state, let

$$\tau_b = \inf\{t \ge 1 : r(Q_t) \le \gamma b\} \tag{7}$$

be the first posterior-risk contraction.

Definition 2.3 (Stopped cost)**.** *An admissible experiment consists of an adaptive query policy and a positive, possibly infinite, $(\mathcal{H}_t)$-stopping time σ. It runs for*

$$N = \sigma \wedge \tau_b$$

rounds and succeeds on $\mathsf{M} = \{\tau_b \le \sigma\}$. We require $N < \infty$ almost surely. Expectations below are initially extended-valued; Lemma 2.4 shows that every experiment with finite displayed ratio has $\mathbb{E}N < \infty$. Define

$$\mathsf{SC}_b(Q) := \inf_{\pi,\sigma} \frac{b\,\mathbb{E}_Q R_N}{\mathbb{P}_Q(\mathsf{M})}, \qquad R_N = \sum_{t=1}^{N} \Delta_\omega(A_t), \tag{8}$$

with ratio $+\infty$ when the success probability is zero.

The first-hit convention loses nothing: after reaching the lower-risk set, additional observations cannot improve the success event and only add nonnegative regret. Voluntary stopping is essential, since an experiment may abandon an unpromising branch.

Lemma 2.4 (Query cost is already charged)**.** *Every admissible experiment satisfies*

$$\mathbb{E}_Q R_N \geq \gamma b\, \mathbb{E}_Q N. \tag{9}$$

Consequently,

$$\mathsf{SC}_b(Q) \leq \inf_{\pi,\sigma} \frac{b^2 \mathbb{E}_Q N + b\mathbb{E}_Q R_N}{\mathbb{P}_Q(\mathsf{M})} \leq (1+\gamma^{-1})\mathsf{SC}_b(Q).$$

Proof. The event $\{t \leq N\}$ lies in $\mathcal{H}_{t-1}$, and on this event the pre-query posterior has not hit the lower-risk set. If π_t is the randomized action kernel, then

$$\mathbb{E}[\Delta_\omega(A_t) \mid \mathcal{H}_{t-1}] = \int \mathbb{E}_{Q_{t-1}} \Delta_\omega(a)\, \pi_t(da \mid H_{t-1}) \geq r(Q_{t-1}) > \gamma b.$$

Tonelli's theorem applied to the stopped sum proves (9); the comparison follows immediately. □

2.2 The Bellman form of a stopped experiment

The stopped cost is a stochastic shortest-path problem on posterior space. For a query cap L and price $q > 0$, put $W_0^q = 0$ and

$$\begin{aligned} J_k^q(Q,b) = \inf_{a\in\mathcal{A}} \Big\{ & b\, \mathbb{E}_Q \Delta_\omega(a) \\ & + \int \left[-q\, \mathbf{1}\{r(Q^{a,o}) \leq \gamma b\} + \mathbf{1}\{r(Q^{a,o}) > \gamma b\} W_{k-1}^q(Q^{a,o}, b) \right] P_Q(do \mid a) \Big\}, \\ W_k^q(Q,b) = \min\{0, J_k^q(Q,b)\}. & \end{aligned} \tag{10}$$

The root uses J_L^q, so the experiment makes at least one query; W_k^q permits voluntary stopping on a non-hitting branch. If $\mathsf{SC}_{b,L}$ is the infimum in (8) over experiments with $N \leq L$ almost surely, then

$$\mathsf{SC}_{b,L}(Q) < q \quad \Longleftrightarrow \quad J_L^q(Q,b) < 0, \qquad \mathsf{SC}_b(Q) = \inf_{L\geq 1} \mathsf{SC}_{b,L}(Q). \tag{11}$$

The first equivalence follows by subtracting $q\mathbb{P}(\mathsf{M})$ from the stopped cost; the second follows by monotone truncation. Early observations receive value through the continuation value at $Q^{a,o}$, even when the first query does not itself hit.

2.3 Pointwise and uniform values

Let $\mathfrak{D} = (\mathfrak{D}_d)$ be a sequence of such decision problems. For a scale window $B^{-1} \leq b \leq 1$, write $\mathsf{S}_{\mathfrak{D}}(d,B)$ for the active posterior states. Define

$$\mathsf{SC}_{\mathfrak{D}}^{\mathrm{pt}}(d,B) = \sup_{(Q,b)\in \mathsf{S}_{\mathfrak{D}}(d,B)} \inf_{\mathcal{E}} \frac{b\, \mathbb{E}_Q R_N}{\mathbb{P}_Q(\mathsf{M})}, \tag{12}$$

$$\mathsf{SC}_{\mathfrak{D}}^{\mathrm{unif}}(d,B) = \inf_{\Sigma \text{ universally measurable}} \sup_{(Q,b)\in \mathsf{S}_{\mathfrak{D}}(d,B)} \frac{b\, \mathbb{E}_Q^{\Sigma(Q,b)} R_N}{\mathbb{P}_Q^{\Sigma(Q,b)}(\mathsf{M})}. \tag{13}$$

The supremum of an empty set is zero. Always $\mathsf{SC}^{\mathrm{pt}} \le \mathsf{SC}^{\mathrm{unif}}$. The first quantity chooses the best experiment separately at each state. The second requires all statewise choices to form one universally measurable experiment rule.

Let

$$\mathfrak{R}^{\mathrm{B}}_{\mathfrak{D}}(d,T) = \sup_{Q} \inf_{\pi} \mathbb{E}^{\pi}_{Q} R_T$$

denote worst-prior Bayes regret, with any outer supremum over known action spaces understood.

Theorem 2.5 (Regret bounds from stopped experiments)**.** *For every $T \ge 2$ and $T^{-1} \le \eta < 1/2$,*

$$\mathfrak{R}^{\mathrm{B}}_{\mathfrak{D}}(d,T) \le \min\left\{T,\, C_{\gamma,\vartheta}\frac{\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,T)}{\eta} + 2T\eta\right\}. \tag{14}$$

In particular,

$$\mathfrak{R}^{\mathrm{B}}_{\mathfrak{D}}(d,T) \le C_{\gamma,\vartheta} \min\left\{T,\, 1 + \sqrt{\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,T)\, T}\right\}. \tag{15}$$

Conversely, for every $0 < c < \mathsf{SC}^{\mathrm{pt}}_{\mathfrak{D}}(d,B)$, some active initial posterior and scale $B^{-1} \le b \le 1$ satisfy, for every $n \ge 1$,

$$\mathfrak{R}^{\mathrm{B}}_{\mathfrak{D}}(d,n) \ge \frac{\gamma c b n}{c + \gamma b^2 n} \ge \frac{1}{2}\min\left\{\gamma b n, \frac{c}{b}\right\}. \tag{16}$$

If $c/(\gamma b^2) \ge 2$, then at

$$n = \left\lfloor \frac{c}{\gamma b^2} \right\rfloor$$

one has

$$\mathfrak{R}^{\mathrm{B}}_{\mathfrak{D}}(d,n) \ge \frac{\sqrt{\gamma}}{4}\sqrt{cn}. \tag{17}$$

Proof. If $\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,T) = +\infty$, the upper statement is the trivial bound T. Otherwise fix one universally measurable selector Σ whose worst-state ratio is at most $\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,T) + \epsilon$. For the upper bound, use geometric risk buckets $b_j = \vartheta^j$. Whenever $r(Q_t) > 2\eta$, invoke $\Sigma(Q_t, b_j)$ at the unique active scale; otherwise play a measurable δ_t-Bayes action, where $\sum_t \delta_t \le \delta$. Virtually complete the last block if it crosses the horizon. Every actual block makes at least one query, so only this final block can cross T. At scale b_j, the selector's ratio bound gives

$$b_j\, \mathbb{E}[R_k \mid H_{S_k}] \le \{\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,T) + \epsilon\} p_k,$$

where p_k is the conditional hit probability.

After the k-th hit in bucket j, at time U_k, one has $r(Q_{U_k}) \le \gamma b_j$. By (5) and Doob's maximal inequality, applied first over finite horizons after U_k and then by monotone convergence,

$$\mathbb{P}\left\{\sup_{t \ge U_k} r(Q_t) > \vartheta b_j \,\middle|\, \mathcal{H}_{U_k}\right\} \le \frac{\gamma}{\vartheta}.$$

A later return to the same active bucket requires the event inside this probability. The number of hits in one bucket is therefore geometrically dominated and has expectation at most $(1 - \gamma/\vartheta)^{-1}$. Hence

$$\mathbb{E}R_{\mathrm{explore}} \le \frac{\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,T) + \epsilon}{1 - \gamma/\vartheta} \sum_{b_j \ge 2\eta} \frac{1}{b_j} \le \frac{\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,T) + \epsilon}{2(1-\vartheta)(1-\gamma/\vartheta)\eta}.$$

The non-experiment rounds contribute at most $2T\eta + \delta$. Letting $\delta, \epsilon \downarrow 0$ proves (14); optimizing η, with the boundary choices and the trivial bound T, proves (15).

For the converse, choose an active state with $\mathsf{SC}_b(Q) > c$. Run an arbitrary n-round policy, stop it at $N = \tau_b \wedge n$, and put $p = \mathbb{P}_Q(\tau_b \le n)$. The stopped-cost premise and nonnegativity give

$$\mathbb{E}_Q R_n \ge \mathbb{E}_Q R_N \ge \frac{c}{b}\, p.$$

By the same pre-hit argument as in Lemma 2.4,

$$\mathbb{E}_Q R_n \ge \mathbb{E}_Q R_N \ge \gamma b\, \mathbb{E}_Q N \ge \gamma b n (1 - p).$$

Minimizing the maximum of these two affine functions over $p \in [0, 1]$ gives (16). At the displayed calibrated horizon, $n \ge c/(2\gamma b^2)$, and direct substitution gives $\mathbb{E}_Q R_n \ge c/(4b) \ge (\sqrt{\gamma}/4)\sqrt{cn}$. □

2.4 Polynomial exponents

The finite-value bounds compare different scales and, on the lower side, different horizons. Polynomial exponents remove this calibration while retaining the distinction between statewise and measurably implementable experiments. Define

$$\alpha_{\mathfrak{D}}^{\mathrm{stop,pt}} = \sup_{m \in \mathbb{N}} \limsup_{d \to \infty} \frac{\log\{1 + \mathsf{SC}_{\mathfrak{D}}^{\mathrm{pt}}(d, d^m)\}}{\log d}, \tag{18}$$

$$\alpha_{\mathfrak{D}}^{\mathrm{stop,unif}} = \sup_{m \in \mathbb{N}} \limsup_{d \to \infty} \frac{\log\{1 + \mathsf{SC}_{\mathfrak{D}}^{\mathrm{unif}}(d, d^m)\}}{\log d}, \tag{19}$$

$$\alpha_{\mathfrak{D}}^{\mathrm{B}} = \sup_{M \in \mathbb{N}} \limsup_{d \to \infty} \sup_{2 \le T \le d^M} \frac{\log\{1 + \mathfrak{R}_{\mathfrak{D}}^{\mathrm{B}}(d, T)/\sqrt{T}\}}{\log d}. \tag{20}$$

If a frequentist minimax value $\mathfrak{R}_{\mathfrak{D}}^{\mathrm{M}}$ is defined, let $\alpha_{\mathfrak{D}}^{\mathrm{M}}$ be given by the same display.

Theorem 2.6 (Regret and stopped exponents)**.** *Every sequence of repeated latent decision problems satisfying the assumptions of this section obeys*

$$\frac{\alpha_{\mathfrak{D}}^{\mathrm{stop,pt}}}{2} \le \alpha_{\mathfrak{D}}^{\mathrm{B}} \le \frac{\alpha_{\mathfrak{D}}^{\mathrm{stop,unif}}}{2}. \tag{21}$$

Neither inequality requires convexity or a particular observation law.

Suppose, in addition, that for every finite d, B and every $\epsilon > 0$, the statewise experiments have a universally measurable choice whose ratio at every active (Q, b) is at most $\mathsf{SC}_b(Q) + \epsilon$. Then

$$\mathsf{SC}_{\mathfrak{D}}^{\mathrm{pt}}(d, B) = \mathsf{SC}_{\mathfrak{D}}^{\mathrm{unif}}(d, B) =: \mathsf{SC}_{\mathfrak{D}}(d, B),$$

and

$$\alpha_{\mathfrak{D}}^{\mathrm{B}} = \frac{\alpha_{\mathfrak{D}}^{\mathrm{stop}}}{2}. \tag{22}$$

If finite-horizon Bayes–minimax equality also holds, then

$$\alpha_{\mathfrak{D}}^{\mathrm{M}} = \alpha_{\mathfrak{D}}^{\mathrm{B}} = \frac{\alpha_{\mathfrak{D}}^{\mathrm{stop}}}{2}. \tag{23}$$

Proof. For $T \le d^M$, Theorem 2.5 and monotonicity of the scale window give

$$\frac{\mathfrak{R}^{\mathrm{B}}_{\mathfrak{D}}(d,T)}{\sqrt{T}} \le C_{\gamma,\vartheta}\left\{T^{-1/2} + \sqrt{\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{D}}(d,d^M)}\right\}.$$

Taking exponents proves the upper inequality in (21).

For the reverse inequality, first suppose $\alpha^{\mathrm{stop,pt}}_{\mathfrak{D}} > 0$, and fix

$$0 < q < q' < \alpha^{\mathrm{stop,pt}}_{\mathfrak{D}}.$$

For some fixed m and an infinite sequence of dimensions,

$$1 + \mathsf{SC}^{\mathrm{pt}}_{\mathfrak{D}}(d,d^m) \ge d^{q'}.$$

For all sufficiently large dimensions in this sequence, there is an active state (Q_d, b_d), with $b_d \ge d^{-m}$, such that

$$\mathsf{SC}_{b_d}(Q_d) > c_d, \qquad c_d = d^q.$$

Set

$$n_d = \left\lfloor \frac{c_d}{\gamma b_d^2} \right\rfloor.$$

Then $c_d/(\gamma b_d^2) \ge 2$ for large d, and

$$n_d \le \gamma^{-1} d^{q+2m} \le d^M$$

for one fixed integer M. The calibrated lower bound in Theorem 2.5 yields

$$\frac{\mathfrak{R}^{\mathrm{B}}_{\mathfrak{D}}(d,n_d)}{\sqrt{n_d}} \ge \frac{\sqrt{\gamma}}{4}\sqrt{c_d} = \frac{\sqrt{\gamma}}{4} d^{q/2}.$$

Thus $\alpha^{\mathrm{B}}_{\mathfrak{D}} \ge q/2$. Let $q \uparrow \alpha^{\mathrm{stop,pt}}_{\mathfrak{D}}$; if the latter exponent is infinite, take arbitrary finite q. When it is zero, the lower inequality is immediate.

The measurable-choice assumption identifies the two finite stopped values, and hence their exponents. Bayes–minimax equality identifies the remaining regret exponents. □

The equality does not require information-scale matching or nondegeneracy; it still assumes nonnegative gaps, posterior sufficiency, recurrence of the same decision problem, and measurable selection.

Corollary 2.7 (Finite episodic models)**.** *Consider any sequence of reset episodic reinforcement-learning problems as in Example 2.1, with a finite model class, finite state, action, and reward alphabets, and a finite episode horizon. Using the normalized episode gap,*

$$\alpha^{\mathrm{M}}_{\mathrm{RL}} = \alpha^{\mathrm{B}}_{\mathrm{RL}} = \frac{\alpha^{\mathrm{stop}}_{\mathrm{RL}}}{2}. \tag{24}$$

For fixed H, normalization does not change the dimension exponent. If H grows with d, its dependence must be restored when translating back to unnormalized regret.

Proof. The complete contingent policy space and the trajectory space are finite. The posterior belongs to a finite-dimensional simplex, the capped stopped problems admit measurable statewise choices, and truncation removes the cap. At every finite outer horizon there are finitely many pure nonanticipating policies, and randomized policies generate their compact convex hull. The finite minimax theorem gives Bayes–minimax equality. Apply Theorem 2.6. □

3 Aggregate-feedback top-k bandits

We now use the conversion to obtain a finite-horizon minimax result. There are n independent arms with unit-variance Gaussian rewards. At round t, the learner selects $S_t \subset [n]$, $|S_t| = k$, and observes only

$$Y_t = \sum_{j \in S_t} X_{t,j}.$$

The reward is the same sum. Thus this is full-bandit, rather than semi-bandit, feedback. Write $\mathfrak{R}^\star_{n,k}(T)$ for minimax expected regret over mean vectors in $[0,1]^n$.

Theorem 3.1 (Dense top-k aggregate-feedback regret)**.** *There are universal constants $c, C > 0$ such that, for $n = 2k$,*

$$c \min\{kT, k^{3/2}\sqrt{T}\} \le \mathfrak{R}^\star_{2k,k}(T) \le C \min\{kT, k^{3/2}\sqrt{T \log T}\}. \tag{25}$$

Consequently the minimax rate is $\widetilde{\Theta}(\min\{kT, k^{3/2}\sqrt{T}\})$.

The upper bound is the minimum of the trivial kT bound and the distribution-free CSAR guarantee $\widetilde{O}(k\sqrt{nT})$ of Rejwan and Mansour (2020). We prove the lower bound by calculating a stopped cost. Pair the arms and draw $\theta = (\theta_1, \dots, \theta_k)$ uniformly from $\{-1, +1\}^k$. For $i \in [k]$, set

$$\mu_{i,+} = \frac{1}{2} + \varepsilon\theta_i, \qquad \mu_{i,-} = \frac{1}{2} - \varepsilon\theta_i, \qquad 0 < \varepsilon \le \frac{1}{4}. \tag{26}$$

For a selected k-set S, put

$$a_i(S) = \mathbf{1}\{(i,+) \in S\} - \mathbf{1}\{(i,-) \in S\} \in \{-1, 0, 1\}.$$

Then

$$Y = \frac{k}{2} + \varepsilon\langle a(S), \theta\rangle + \xi, \qquad \xi \sim N(0, k), \tag{27}$$

and the unnormalized gap is

$$\Delta^{\mathrm{raw}}_\theta(S) = \varepsilon\{k - \langle a(S), \theta\rangle\}. \tag{28}$$

We apply the general theory to $\Delta_\theta = \Delta^{\mathrm{raw}}_\theta / k \in [0, 1]$.

Let $m_i(H) = \mathbb{E}[\theta_i \mid H]$. Because one posterior-preferred arm from each pair is Bayes optimal, the posterior risk is exactly

$$r(Q_H) = \frac{\varepsilon}{k} \sum_{i=1}^{k} \{1 - |m_i(H)|\}. \tag{29}$$

In particular the product prior Q_0 is active at scale $b = \varepsilon$.

Proposition 3.2 (Stopped cost of the paired family)**.** *For every fixed $0 < \gamma < 1$, uniformly over $0 < \varepsilon \le 1/4$,*

$$\mathsf{SC}_\varepsilon(Q_0) = \Theta_\gamma(k). \tag{30}$$

Proof. For the upper bound, use $N = C_\gamma k/\varepsilon^2$ predetermined balanced sign queries $a_t \in \{-1, +1\}^k$, chosen as repeated orthogonal rows when available and otherwise as a Rademacher design. Least squares under (27) estimates a constant fraction of the signs with error probability small enough that (29) falls below $\gamma\varepsilon$ with probability bounded away from zero. Each query has prior expected normalized gap exactly ε. Stopping at the first contraction can only reduce regret, and hence

$$\frac{\varepsilon\, \mathbb{E}R_N}{\mathbb{P}(\mathsf{M})} \lesssim_\gamma \varepsilon \frac{(k/\varepsilon^2)\varepsilon}{1} \lesssim_\gamma k.$$

For the reverse inequality, the stopped direct-product lemma (Lemma A.1) applied to (27) gives, for every experiment with success probability p,

$$\mathbb{E}N \geq c_\gamma \frac{pk}{\varepsilon^2}. \tag{31}$$

The point is that a successful transcript has $\sum_i |m_i| \geq (1-\gamma)k$, whereas flipping coordinate i changes one conditional Gaussian mean by $2\varepsilon a_i$, and $\sum_i a_i^2 \leq k$ cancels the aggregate variance k. Lemma 2.4 now yields

$$\frac{\varepsilon\, \mathbb{E}R_N}{p} \geq \gamma\varepsilon^2 \frac{\mathbb{E}N}{p} \geq c_\gamma k.$$

□

Proof of Theorem 3.1. The upper bound was noted above. For the lower bound, apply Theorem 2.5 to Proposition 3.2. For $T \gtrsim k$, choose $\varepsilon \asymp \sqrt{k/T}$; the calibrated-horizon bound gives $\Omega(\sqrt{kT})$ normalized regret. Multiplication by k restores sum-reward units and gives $\Omega(k^{3/2}\sqrt{T})$. For $T \lesssim k$, take a sufficiently small constant ε. The non-hitting term in (16) gives $\Omega(T)$ normalized regret and hence $\Omega(kT)$ in raw units. □

Why the stopped conversion is essential here. The paired family also has PAC identification complexity $\Theta(k/\varepsilon^2)$. That fact alone does not imply (25): a sample may be statistically informative while having very different decision loss, and a finite-horizon policy may learn only on selected transcript branches. Proposition 3.2 couples information, regret, and the probability of posterior contraction for the same stopped policy. Theorem 2.5 then handles both the branch that contracts and the branch that continues to pay risk. This is precisely the step absent from a fixed-confidence sample-complexity argument.

Relation to extreme contextual recommendation. Top-k contextual bandits are used for advertising and recommendation when the arm set is extremely large. Sen et al. (2021) use an arm hierarchy to reduce millions of arms to a context-dependent candidate set. Their feedback model may reveal all or some selected-arm rewards, whereas Theorem 3.1 assumes only the aggregate. The theorem therefore does not lower-bound their algorithm. It identifies a complementary bottleneck: after computational localization has produced a candidate set, loss of individual attribution can add a k-dependent statistical price. A contextual stopped complexity would condition on the context and posterior, charge the regret of aggregate probes, and stop when the context-averaged top-k decision risk contracts.

4 Gaussian feedback on compact action spaces

We next give a broad setting in which the statewise stopped experiments form one measurable rule and Bayes regret equals minimax regret. Let $\mathcal{C}_d$ be a collection of nonempty compact metric action spaces. For $C \in \mathcal{C}_d$, let $\mathfrak{F}_d(C)$ be a nonempty Borel subset of $\mathcal{C}(C, [0,1])$, with the uniform topology. If f is fixed, querying $x \in C$ reveals

$$O = f(x) + \xi, \qquad \xi \sim N(0,1),$$

and incurs the gap $f(x) - \min_C f$. Neither C nor f is assumed to be convex in this section. Write

$$\mathfrak{R}^{\mathrm{M}}_{\mathfrak{F}}(d,T) = \sup_{C\in\mathcal{C}_d} \inf_{\pi} \sup_{f\in\mathfrak{F}_d(C)} \mathbb{E}^{\pi}_{f} R_T, \tag{32}$$

$$\mathfrak{R}^{\mathrm{B}}_{\mathfrak{F}}(d,T) = \sup_{C\in\mathcal{C}_d} \sup_{Q \text{ on } \mathfrak{F}_d(C)} \inf_{\pi} \mathbb{E}^{\pi}_{Q} R_T. \tag{33}$$

Both infima are over Borel randomized nonanticipating policies. The values $\mathsf{SC}^{\mathrm{pt}}_{\mathfrak{F},C}$ and $\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{F},C}$ are defined by (12)–(13); an outer supremum over $C \in \mathcal{C}_d$ gives $\mathsf{SC}^{\mathrm{pt}}_{\mathfrak{F}}$ and $\mathsf{SC}^{\mathrm{unif}}_{\mathfrak{F}}$.

Theorem 4.1 (Gaussian statewise selection and minimax equality)**.** *For every finite d, T, B,*

$$\mathfrak{R}^{\mathrm{M}}_{\mathfrak{F}}(d,T) = \mathfrak{R}^{\mathrm{B}}_{\mathfrak{F}}(d,T), \tag{34}$$

$$\mathsf{SC}^{\mathrm{pt}}_{\mathfrak{F},C}(d,B) = \mathsf{SC}^{\mathrm{unif}}_{\mathfrak{F},C}(d,B). \tag{35}$$

The supremum in (34) may be restricted to finitely supported priors. For every $\epsilon > 0$, the second equality is implemented by a universally measurable state-dependent experiment whose finite deterministic cap may depend on the state.

The first equality is a finite-horizon minimax theorem, not an asymptotic interchange. The second identifies the statewise and uniform stopped values. Convexity is absent: compactness and continuity give Bayes-action minimizers, lower-semianalytic dynamic programming gives universally measurable near-minimizers, and the positive Gaussian density gives a Borel posterior update. Write their common worst-space value as $\mathsf{SC}_{\mathfrak{F}}(d,B)$, and use the exponents from (18)–(20) with $\mathfrak{D} = \mathfrak{F}$.

Corollary 4.2 (Regret and stopped exponents under Gaussian feedback)**.** *For every sequence of the Gaussian loss models above,*

$$\alpha^{\mathrm{M}}_{\mathfrak{F}} = \alpha^{\mathrm{B}}_{\mathfrak{F}} = \frac{\alpha^{\mathrm{stop}}_{\mathfrak{F}}}{2}. \tag{36}$$

No information-scale matching or nondegeneracy assumption is required.

Proof. Apply Theorem 2.6 and (34)–(35). □

4.1 The full convex class

We now return to stochastic BCO: $C \subset \mathbb{R}^d$ is a compact convex body and $\mathfrak{F}_d(C)$ is the full class of continuous $[0,1]$-valued convex losses. Convexity first becomes essential here, through the model class and the BCO bounds used below.

Corollary 4.3 (Full-class exponents)**.** *For the full class of continuous $[0,1]$-valued convex losses,*

$$\frac{5}{2} \le \alpha^{\mathrm{stop}}_{\mathrm{full}} \le 3, \qquad \frac{5}{4} \le \alpha^{\mathrm{reg}}_{\mathrm{full}} \le \frac{3}{2}. \tag{37}$$

Determining either exponent determines the other.

Proof. The upper endpoint is the current general BCO upper bound (Lattimore and György, 2023; Fokkema et al., 2024); the lower endpoint is the minimax theorem of Rajaraman (2026). Its 1-Lipschitz hard family embeds in the present bounded unit-noise model by the affine rescaling $g = (f+1)/2$: from an observation $f(x) + \xi$, return $(f(x) + \xi + 1)/2 + \zeta$, where $\zeta \sim N(0, 3/4)$ is independent. The result is $g(x) + \xi'$ with $\xi' \sim N(0,1)$, and regret changes only by a factor two. Apply Corollary 4.2. □

5 Observation-adaptive experiments: the dimple family

Let $C = B_2^d(1)$. Fix

$$d^{-1} \le \varepsilon \le \varepsilon_0, \qquad s = \sqrt{1+2\varepsilon},$$

where $\varepsilon_0 > 0$ is a sufficiently small constant. For $\theta \in S_2^{d-1}(1)$, define

$$F_\theta(x) = \varepsilon + \frac{1}{2}\|x\|^2 - \frac{1}{2}\big(s - \|\theta - x\|\big)_+^2. \tag{38}$$

These functions are continuous, convex, and $[0,1]$-valued after reducing ε_0 if necessary (Bakhtiari et al., 2025, Lemma 10). Their unique minimizers are θ, with $F_\theta(\theta) = 0$. At the origin,

$$b_0 := F_\theta(0) = s - 1 = \frac{2\varepsilon}{\sqrt{1+2\varepsilon}+1} \asymp \varepsilon. \tag{39}$$

Bakhtiari et al. (2025) use this family to exhibit an obstruction to standard one-step information-ratio analysis. We show that the same geometry admits a cheaper multistep experiment. The lower bound below uses mutual information to show that an unoriented ray reveals very little; the upper construction uses early observations to orient later tangent probes.

Use the rotation-invariant prior $\Theta \sim \mathrm{Unif}(S_2^{d-1}(1))$. Since $\nabla F_\theta(0) = -b_0\theta$, convexity gives

$$F_\theta(x) \ge b_0\{1 - \langle \theta, x\rangle\}.$$

The origin is therefore the prior Bayes action and $r(Q_0) = b_0$. Write

$$\tau_\gamma = \inf\{t \ge 1 : r(Q_t) \le \gamma b_0\}. \tag{40}$$

Definition 5.1 (Fresh directions)**.** *A policy uses* fresh directions *if, at round t, it first draws $B_t \sim \mathrm{Unif}(S_2^{d-1}(1))$, independently of (Θ, H_{t-1}), then chooses an arbitrary $\sigma(H_{t-1}, B_t)$-measurable radius $R_t \in [0,1]$ and queries $X_t = R_t B_t$.*

The radius may use the complete past. Only the direction must be fresh. Every ray starts at the origin. The comparison excludes shifted or paired probes, anisotropic directions, and reuse of a learned direction.

Theorem 5.2 (Fresh-direction information bound)**.** *For $d \ge 2$, there is a universal $C < \infty$ such that, for every posterior ν of Θ, every fresh $B \sim \mathrm{Unif}(S_2^{d-1}(1))$, every measurable $R = R(B) \in [0,1]$, and independent $\xi \sim N(0,1)$,*

$$I_\nu(\Theta; B, F_\Theta(RB) + \xi) \le C(\varepsilon^4 + d^{-4}). \tag{41}$$

Under the spherical prior, every nonadaptive batch $X_1, \dots, X_n$, possibly randomized but independent of Θ and fixed before its observations, satisfies

$$I(\Theta; O_{1:n} \mid X_{1:n}) \le Cn(\varepsilon^4 + d^{-4}). \tag{42}$$

For any initial prior, every stopped fresh-direction experiment with duration N and $\mathbb{E}N < \infty$ satisfies

$$I(\Theta; H_N) \le C(\varepsilon^4 + d^{-4})\,\mathbb{E}N. \tag{43}$$

Conversely, under the spherical initial prior, if $\mathsf{M} \in \sigma(H_N)$ is an event on which the terminal posterior Bayes risk is at most γb_0, then

$$I(\Theta; H_N) \ge \frac{d-1}{2}(1-\gamma)^2\,\mathbb{P}(\mathsf{M}). \tag{44}$$

Consequently, every stopped fresh-direction experiment under the spherical prior with $\mathbb{P}(\mathsf{M}) > 0$ *satisfies*

$$\frac{\mathbb{E}N}{\mathbb{P}(\mathsf{M})} \geq c_\gamma \frac{d}{\varepsilon^4 + d^{-4}}. \tag{45}$$

Under the spherical prior, the same bound holds with $N = n$ *for an arbitrary nonadaptive batch of size* n*. Thus, under that prior, constant-probability contraction requires batch size, or expected fresh-direction duration,* $\Omega_\gamma(d/(\varepsilon^4 + d^{-4}))$*.*

Fix a sufficiently large $C_L = C_L(\gamma)$, and for $0 < a \leq b_0$ put

$$L_{d,a} = 1 + \log\left\{\frac{C_L d(1 + \log(d/a))}{a}\right\}. \tag{46}$$

Theorem 5.3 (Observation-adaptive directional experiment)**.** *There are constants* $\gamma_0, c_0 > 0$*, with* $\gamma_0 + c_0 < \gamma$*, and* $C_\gamma < \infty$ *such that, for every sufficiently large* d*, every* $d^{-1} \leq \varepsilon \leq \varepsilon_0$*, and every* $0 < a \leq b_0$*, an observation-adaptive experiment returns* $\widehat{x} \in B_2^d(1)$ *with*

$$\sup_{\theta \in S_2^{d-1}(1)} \mathbb{P}_\theta\{F_\theta(\widehat{x}) > \gamma_0 a\} \leq c_0 a, \tag{47}$$

$$N \leq \frac{C_\gamma d^2 L_{d,a}^2}{a^2}, \tag{48}$$

$$\sup_{\theta \in S_2^{d-1}(1)} \mathbb{E}_\theta R_N \leq \frac{C_\gamma d^2 L_{d,a}^2}{a}. \tag{49}$$

The query cap is deterministic. The conditional distributions and subspaces of the directions in the last two stages depend measurably on earlier observations. Under every prior on Θ*,*

$$\mathbb{P}\left\{\mathbb{E}[F_\Theta(\widehat{x}) \mid H_N] \leq \gamma a\right\} \geq 1 - \frac{\gamma_0 + c_0}{\gamma} > 0. \tag{50}$$

The construction has three stages. The table suppresses logarithmic factors.

stage	progress	(queries, regret)
orthogonal scan	obtain v_0 with $\langle \theta, v_0 \rangle = \widetilde{\Omega}(\max\{\sqrt{\varepsilon}, d^{-1/2}\})$	$(d^2/\varepsilon^2,\ d^2/\varepsilon)$
tangent amplification	choose tangent directions from the current estimate and enter a fixed angular cone	$(d^2/\varepsilon^2,\ d^2/\varepsilon)$
angular contraction	invert boundary observations and repeat $q \mapsto q/2$ until $q = O(a)$	$(d^2/a^2,\ d^2/a)$

Only the last two stages orient later probes using earlier observations. Their detailed analysis is in Appendix D.

For the next statement, $\mathsf{SC}_b^{[1]}$ restricts (8) to one query, and $\mathsf{SC}_b^{\mathrm{fresh}}$ restricts it to fresh-direction policies.

Corollary 5.4 (Polynomial separations)**.** *For all sufficiently large* d*, at* $\varepsilon = d^{-1}$ *and* $b = b_0 \asymp d^{-1}$*, the initial spherical state satisfies*

$$\mathsf{SC}_b^{[1]} = \Omega_\gamma(d^3), \qquad \mathsf{SC}_b^{\mathrm{fresh}} = \Omega_\gamma(d^3), \qquad \mathsf{SC}_b = \widetilde{O}_\gamma(d^2). \tag{51}$$

At the same scale and under the spherical prior, a nonadaptive batch with constant contraction probability has size $\Omega_\gamma(d^5)$, and every stopped fresh-direction experiment with $\mathbb{P}(\mathsf{M}) > 0$ satisfies

$$\frac{\mathbb{E}N}{\mathbb{P}(\mathsf{M})} \geq c_\gamma d^5.$$

In particular, constant success probability requires $\mathbb{E}N = \Omega_\gamma(d^5)$. The observation-adaptive construction contracts risk with constant probability using $\widetilde{O}_\gamma(d^4)$ queries.

Let

$$L_{d,T} = 1 + \log\left\{C_L d \max\{T, d\}[1 + \log\{d \max\{T, d\}\}]\right\}.$$

Under the spherical prior, every fresh-direction policy satisfies

$$\mathbb{E}R_T \geq c_\gamma \min\left\{\frac{T}{d}, d^4\right\}, \tag{52}$$

while an unrestricted policy satisfies, uniformly in θ,

$$\mathbb{E}_\theta R_T \leq C_\gamma L_{d,T} \min\left\{\frac{T}{d}, d\sqrt{T}\right\}. \tag{53}$$

At $T = d^5$, these bounds are $\Omega(d^4)$ and $\widetilde{O}(d^{7/2})$, respectively.

Proof of (51). For a one-query or fresh-direction first-hit experiment, let $p = \mathbb{P}(\mathsf{M})$. Theorem 5.2 gives

$$C(\varepsilon^4 + d^{-4})\,\mathbb{E}N \geq I(\Theta; H_N) \geq c_\gamma d\, p.$$

Before the hit, Lemma 2.4 gives $\mathbb{E}R_N \geq \gamma b\,\mathbb{E}N$. Therefore

$$\frac{b\,\mathbb{E}R_N}{p} \geq \frac{c_\gamma b^2 d}{\varepsilon^4 + d^{-4}}.$$

At $b \asymp \varepsilon = d^{-1}$, this is $\Omega_\gamma(d^3)$. The same argument covers the one-query restriction.

Couple the capped routine from Theorem 5.3, run at $a = b_0$, to its first-hit truncation. The two policies agree until the hit, and (50) implies that the virtual capped path hits with probability at least $p_\gamma := 1 - (\gamma_0 + c_0)/\gamma > 0$. Truncation can only decrease regret, so

$$\frac{b\,\mathbb{E}R_{\tau_\gamma \wedge N}}{\mathbb{P}(\tau_\gamma \leq N)} \leq \frac{C_\gamma}{p_\gamma} b \frac{d^2 L_{d,b}^2}{b} = \widetilde{O}_\gamma(d^2).$$

This proves the upper bound. □

6 Discussion

The stopped cost asks how much regret must be spent before observations make a repeated decision problem easier. The aggregate-feedback theorem shows that this question can determine a minimax rate. On the paired top-k family, posterior contraction requires learning many signs through noisy sums, and every query used to do so has a decision cost. The stopped lower-bound conversion is indispensable because it applies to an arbitrary finite-horizon policy, including policies that contract only on selected transcript branches. The matching upper rate comes from an existing Hadamard-design algorithm; stopped experiments certify its local cost but are not needed to run it.

Bandit convex optimization illustrates a different, genuinely multistep role. On the dimple family, an unoriented ray carries little information about the hidden direction, but a weak initial direction can be reused to place tangent probes, amplify alignment, and contract angular error. This gives a polynomial separation between spherical-prior nonadaptive experiments or fresh-direction experiments and an observation-adaptive experiment. It does not give a lower bound for unrestricted BCO.

For the full convex class, determining $\alpha_{\mathrm{full}}^{\mathrm{stop}} \in [5/2, 3]$ is equivalent, at the level of polynomial dimension exponents, to closing the current BCO regret gap. An upper bound must construct low-cost posterior experiments uniformly over states and scales; a lower bound must rule out their full adaptive choice of later queries.

For episodic reinforcement learning, the present contribution is a conversion theorem rather than a new tabular minimax rate. The macro-action reduction makes the scope exact and shows what an application must prove: a policy-search or preference-query routine must reduce posterior decision risk while charging the value loss of the trajectories it generates. Preference-based RL is especially natural because numerical rewards are unavailable and comparison complexity alone can hide poor exploration trajectories. Establishing a stopped-cost bound for a structured preference model would immediately yield a regret guarantee, but no such new rate is claimed here.

The top-k contextual extension has two separate scales. Arm hierarchies can make search over an extreme catalog computationally feasible (Sen et al., 2021); aggregate feedback can still make attribution statistically difficult within the localized candidate set. A useful next theorem would calculate a context-averaged stopped profile that combines these reductions. The dense noncontextual theorem proves that the second price can be real, but does not transfer automatically to semi-bandit feedback or to a realizable contextual regression class.

A A stopped direct-product lemma

Lemma A.1 (Stopped Assouad bound for Gaussian sums)**.** *Let θ be uniform on $\{-1,+1\}^k$. At time t, an adaptive policy chooses $a_t \in \{-1,0,1\}^k$, with $\|a_t\|_2^2 \le k$, and observes*

$$Y_t = \varepsilon\langle a_t, \theta\rangle + \xi_t, \qquad \xi_t \overset{\text{iid}}{\sim} N(0,k).$$

Let N be the first time that $\sum_i (1 - |\mathbb{E}[\theta_i \mid H_N]|) \le \gamma k$, or an earlier voluntary stopping time, and let M be the first alternative. If $p = \mathbb{P}(\mathsf{M})$, then

$$\mathbb{E}N \ge c_\gamma \frac{pk}{\varepsilon^2}.$$

Proof. For $i \in [k]$, pair each parameter θ with $\theta^{(i)}$, obtained by flipping its i-th coordinate, and run the two transcript laws with the same policy kernel. The stopped Gaussian chain rule gives

$$D_{\mathrm{KL}}(P_\theta^N \| P_{\theta^{(i)}}^N) = \frac{2\varepsilon^2}{k}\mathbb{E}_\theta \sum_{t \le N} a_{t,i}^2. \tag{54}$$

This identity follows first for $N \wedge L$ and then by monotone convergence. Averaging over θ, summing over i, and using $\sum_i a_{t,i}^2 \le k$ yields

$$2^{-k}\sum_\theta \sum_{i=1}^k D_{\mathrm{KL}}(P_\theta^N \| P_{\theta^{(i)}}^N) \le 2\varepsilon^2 \mathbb{E}N. \tag{55}$$

For each i, let P_i^+ and P_i^- be the transcript laws conditional on $\theta_i = +1$ and $\theta_i = -1$, respectively. Joint convexity of relative entropy and the pairing above give

$$\sum_i \{D_{\mathrm{KL}}(P_i^+ \| P_i^-) + D_{\mathrm{KL}}(P_i^- \| P_i^+)\} \le 2^{1-k}\sum_\theta \sum_i D_{\mathrm{KL}}(P_\theta^N \| P_{\theta^{(i)}}^N).$$

Let $P = (P_i^+ + P_i^-)/2$, which is the same unconditional transcript law for every i. Bayes' rule gives

$$m_i(h) = \frac{dP_i^+ - dP_i^-}{dP_i^+ + dP_i^-}(h).$$

Consequently the i-th Jeffreys divergence is

$$D_{\mathrm{KL}}(P_i^+ \| P_i^-) + D_{\mathrm{KL}}(P_i^- \| P_i^+) = 4\mathbb{E}_P[m_i(H_N)\operatorname{arctanh} m_i(H_N)]\,.$$

Since $x \operatorname{arctanh} x \ge x^2$, if a stopped-transcript event M satisfies

$$\sum_i (1 - |\mathbb{E}[\theta_i \mid H_N]|) \le \gamma k \quad \text{on } M,$$

then Cauchy–Schwarz yields, pointwise on M,

$$\sum_i m_i(H_N)^2 \ge k^{-1}\left(\sum_i |m_i(H_N)|\right)^2 \ge (1-\gamma)^2 k.$$

It follows that

$$2^{-k}\sum_\theta \sum_{i=1}^k D_{\mathrm{KL}}(P_\theta^N \| P_{\theta^{(i)}}^N) \ge 2(1-\gamma)^2 k\,\mathbb{P}(M). \tag{56}$$

Take $M = \mathsf{M}$ and combine (55)–(56). □

B The fresh-direction lower bound

Proof of Theorem 5.2. Put

$$q(x) = \varepsilon + \frac{1}{2}\|x\|^2.$$

For $B \in S_2^{d-1}(1)$, $r \in [0,1]$, and $u = \langle \theta, B \rangle$,

$$|F_\theta(rB) - q(rB)| = \frac{1}{2}\left(s - \sqrt{1 + r^2 - 2ru}\right)_+^2.$$

Minimizing the square root over $r \in [0,1]$ shows that the supremum of the right side is $(s-1)^2/2$ when $u \le 0$, and

$$\frac{1}{2}\left(s - \sqrt{1-u^2}\right)^2$$

when $u > 0$. On $|u| \le 1/2$, its square is at most $C(\varepsilon^4 + u^8)$; the complementary spherical cap has exponentially small probability. Rotation invariance and

$$\mathbb{E}_B \langle \theta, B \rangle^8 = \frac{105}{d(d+2)(d+4)(d+6)}$$

therefore give, uniformly in θ,

$$\mathbb{E}_B \sup_{0 \le r \le 1} |F_\theta(rB) - q(rB)|^2 \le C(\varepsilon^4 + d^{-4}). \tag{57}$$

Now condition on an arbitrary posterior ν. The fresh Haar direction B is independent of Θ, and $R(B)$ is known after conditioning on B. Comparing the Gaussian observation channel with the Θ-independent reference $q(R(B)B) + \xi$, the information-radius bound gives

$$I_\nu(\Theta; B, F_\Theta(RB) + \xi) \le \frac{1}{2}\mathbb{E}_{\Theta \sim \nu, B} |F_\Theta(RB) - q(RB)|^2.$$

Equation (57) proves (41).

Under the spherical prior, for a fixed nonadaptive design point $x = ru$, rotation invariance interchanges a fixed direction u under uniform Θ with a uniform direction B under fixed θ. Thus (57) also gives

$$\mathbb{E}_\Theta |F_\Theta(x) - q(x)|^2 \le C(\varepsilon^4 + d^{-4}).$$

Condition on a possibly randomized design and compare its joint Gaussian observation law with the product reference $(q(X_i) + \xi_i)_{i \le n}$. Additivity of the reference KL divergence proves (42).

For a stopped fresh-direction experiment, let $N_n = N \wedge n$ and use the cemetery-padded transcript. Since $\{N \ge t\}$ is H_{t-1}-measurable, conditional application of (41) and the chain rule give

$$I(\Theta; H_{N_n}) \le C(\varepsilon^4 + d^{-4}) \sum_{t=1}^{n} \mathbb{P}(N \ge t) = C(\varepsilon^4 + d^{-4})\, \mathbb{E}N_n.$$

Continuity of relative entropy along the increasing stopped σ-fields, followed by $n \to \infty$, proves (43).

For the converse, convexity at the origin and $\nabla F_\theta(0) = -b_0\theta$ give

$$F_\theta(x) \ge b_0\{1 - \langle \theta, x \rangle\}. \tag{58}$$

With $m_N = \mathbb{E}[\Theta \mid H_N]$, taking the infimum over x in (58) gives directly

$$r(Q_{H_N}) \geq b_0(1 - \|m_N\|).$$

Hence $r(Q_{H_N}) \leq \gamma b_0$ on M implies $\|m_N\| \geq 1 - \gamma$. The uniform law σ on $S_2^{d-1}(1)$ is $1/(d-1)$-sub-Gaussian. The variational formula for relative entropy therefore yields

$$D_{\mathrm{KL}}(\nu\|\sigma) \geq \frac{d-1}{2}\|\mathbb{E}_\nu\Theta\|^2.$$

Apply this to each posterior and average on M:

$$I(\Theta; H_N) = \mathbb{E}D_{\mathrm{KL}}(Q_{H_N}\|\sigma) \geq \frac{d-1}{2}(1-\gamma)^2\mathbb{P}(\mathsf{M}).$$

Combining this inequality with (43) proves (45). □

Proof of the lower bound in Corollary 5.4. Let $N = \tau_\gamma \wedge T$ and $p = \mathbb{P}(\tau_\gamma \leq T)$. Before the contraction time, every action has conditional expected regret at least $r(Q_{t-1}) > \gamma b_0$. Hence

$$\mathbb{E}R_T \geq \gamma b_0\, \mathbb{E}N. \tag{59}$$

Moreover,

$$\mathbb{E}N \geq T(1-p).$$

Apply Theorem 5.2 to the stopped transcript H_N. On $\{\tau_\gamma \leq T\}$, its terminal posterior risk is at most γb_0, and therefore

$$Cd^{-4}\mathbb{E}N \geq I(\Theta; H_N) \geq c_\gamma d\, p$$

when $\varepsilon = d^{-1}$. Thus

$$\mathbb{E}N \geq c_\gamma d^5 p.$$

Minimizing $\max\{T(1-p), c_\gamma d^5 p\}$ over $p \in [0,1]$ gives $\mathbb{E}N \geq c_\gamma \min\{T, d^5\}$. Since $b_0 \asymp d^{-1}$, (59) proves (52). □

C The dimple regret bounds

Proof of the upper statements in Corollary 5.4. The same adaptive experiment, run to accuracy a and followed by commitment, gives the family upper bound. At $a = b_0 \asymp \varepsilon$, Theorem 5.3 uses $\widetilde{O}(d^2/\varepsilon^2)$ queries and $\widetilde{O}(d^2/\varepsilon)$ regret. Equation (50) proves the adaptive part of the query separation. Since $\varepsilon^4 + d^{-4} \leq 2\varepsilon^4$ in the stated range, Theorem 5.2 proves the lower part for nonadaptive experiments under the spherical prior and for fresh-direction experiments. Specializing to $\varepsilon = d^{-1}$ gives the displayed endpoint.

For the regret upper bound, first note that playing the origin for all T rounds costs $b_0 T \leq CT/d$. When $T \geq C_\gamma d^4 L_{d,T}^2$, choose

$$a = A_\gamma \frac{dL_{d,T}}{\sqrt{T}}$$

with A_γ large enough. Then $a \leq b_0$, $L_{d,a} \leq CL_{d,T}$, and (48) uses at most $T/2$ rounds. Indeed, enlarging the crossover constant makes $a \leq c/d \leq b_0$, while $1/a \leq \sqrt{T}/d$ and $\log(d/a) \leq \log(dT)$ give the displayed logarithmic comparison directly from (46). The exploration regret is at most

$$\frac{C_\gamma d^2 L_{d,a}^2}{a} \leq C_\gamma d L_{d,T}\sqrt{T}.$$

Play $\widehat{x}$ for the remaining rounds. Since losses are bounded by one, (47) gives

$$\sup_{\theta} \mathbb{E}_\theta F_\theta(\widehat{x}) \le (\gamma_0 + c_0)a \le C_\gamma a.$$

The exploitation regret is therefore at most $C_\gamma aT \le C_\gamma dL_{d,T}\sqrt{T}$. Below the displayed crossover, use the origin strategy. If $T \le d^4$, its cost CT/d is the smaller term in (53). If $d^4 < T < C_\gamma d^4 L_{d,T}^2$, then

$$\frac{T}{d} \le C_\gamma L_{d,T}\, d\sqrt{T}.$$

These two cases prove (53) throughout the remaining range. □

D Proof of the observation-adaptive experiment

Throughout this appendix, reduce ε_0, if necessary, so that $\varepsilon_0 \le 1/8$. Put

$$s_+ = \sqrt{1 + 2\varepsilon_0}, \qquad \beta_- = \frac{2}{s_+ + 1}.$$

Then, uniformly over the parameter range,

$$1 \le s \le s_+, \qquad \beta_-\varepsilon \le b_0 = s - 1 \le \varepsilon. \tag{60}$$

Fix

$$\gamma_0 = \frac{\gamma}{4}, \qquad c_\delta = \frac{\gamma}{4}, \qquad c_0 = c_\delta, \qquad \delta = c_\delta a,$$

so that $\gamma_0 + c_0 = \gamma/2 < \gamma$, and put

$$\mathcal{L} = 1 + \log\{16d(1 + \log(d/a))/\delta\}.$$

Thus $\mathcal{L} \le C_\gamma L_{d,a}$ after increasing $C_L(\gamma)$.

We now record the order in which the remaining constants are chosen. First choose a constant A_{H} large enough for all the Haar-coordinate bounds below. Next fix $C_{\mathrm{geo}} \ge 4$, and then choose

$$0 < q_\star \le \min\left\{\frac{1}{8}, \frac{1}{4C_{\mathrm{geo}}}\right\}. \tag{61}$$

In particular,

$$q_\star < s/4, \qquad C_{\mathrm{geo}} q_\star \le s/2$$

uniformly in ε. The initialization step c_h and its repetition constant are then chosen; this produces the numerical correlation constant $\kappa_{\mathrm{init}} > 0$ in (70) below. After κ_{init} is fixed, choose c_2; this fixes the geometric constants B_λ, c_4, k_0. Next choose k and $\rho = \sqrt{1 + k^2}$, and only then choose the amplification repetition constants C_{gate} and C_{amp}, whose simultaneous confidence bounds depend on the resulting deterministic iteration cap. Finally choose c_5, then c_{term}, then c_6, and last C_{con}. The compatible inequalities imposed on these constants are displayed at the point where each is used.

We use the following conditional form of sphere concentration. If $(u_i)_{i=1}^k$ is a Haar orthonormal basis of a fixed k-dimensional subspace and w is a fixed unit vector in that subspace, then, after increasing A_{H},

$$\max_{1 \le i \le k} |\langle w, u_i \rangle| \le A_{\mathrm{H}} \sqrt{\mathcal{L}/k} \tag{62}$$

except on an event of probability at most $\exp(-4\mathcal{L})$, provided $\mathcal{L} \le k$. When $\mathcal{L} > k$, we instead use the deterministic bound one.

Every basis is drawn conditionally on the preceding transcript. Under $\mathbb{P}_\theta$, the current iterate and the tangent component of the fixed vector θ are then fixed, so (62) applies conditionally after normalizing a nonzero tangent component. Every empirical mean uses fresh Gaussian noise. The initialization, amplification, and contraction stages contain at most $C(1+\log(d/a))$ concentration calls in total. The definition of $\mathcal{L}$, together with sufficiently large repetition constants, therefore gives events $\mathcal{H}_{\mathrm{init}}, \mathcal{H}_{\mathrm{amp}}, \mathcal{H}_{\mathrm{con}}$ such that, uniformly in θ,

$$\mathbb{P}_\theta(\mathcal{H}_{\mathrm{init}}^c) \le \delta/3, \quad \mathbb{P}_\theta(\mathcal{H}_{\mathrm{amp}}^c \mid \mathcal{H}_{\mathrm{init}}) \le \delta/3, \quad \mathbb{P}_\theta(\mathcal{H}_{\mathrm{con}}^c \mid \mathcal{H}_{\mathrm{init}} \cap \mathcal{H}_{\mathrm{amp}}) \le \delta/3.$$

Write $\mathcal{H} = \mathcal{H}_{\mathrm{init}} \cap \mathcal{H}_{\mathrm{amp}} \cap \mathcal{H}_{\mathrm{con}}$. All three stages have deterministic caps. Off $\mathcal{H}$, the same capped routine is continued, returning the origin whenever a normalization is undefined and taking the stopping branch at every gate equality. These conventions make the complete experiment total and Borel measurable.

Whenever $r = \|\theta - x\| < s$, expanding (38) gives

$$F_\theta(x) = s\|\theta - x\| + \langle \theta, x\rangle - 1. \tag{63}$$

D.1 Orthogonal initialization

Lemma D.1. *With probability at least $1 - \delta/3$, the construction below returns a unit vector v_0 satisfying*

$$\langle \theta, v_0\rangle \ge c \min\left\{1, \max\left(\sqrt{\varepsilon}, \sqrt{\mathcal{L}/d}\right)\right\}. \tag{64}$$

It uses at most $Cd^2\mathcal{L}^2/\varepsilon^2$ queries and incurs at most $Cd^2\mathcal{L}^2/\varepsilon$ regret on the success event.

Proof. Draw a Haar basis $(u_i)_{i=1}^d$, and set

$$h_0 = c_h \min\{\sqrt{\varepsilon}, \varepsilon\sqrt{d/\mathcal{L}}\}, \qquad n_0 = \left\lceil C_{\mathrm{init}} d\mathcal{L}^2/\varepsilon^2 \right\rceil. \tag{65}$$

Estimate every centered difference

$$D_i^0 = \frac{F_\theta(h_0 u_i) - F_\theta(-h_0 u_i)}{2h_0}$$

using n_0 repetitions at each of the two points. Let A_0 be a universal constant for the Taylor estimate used below. If $\mathcal{L} \le d$, then $h_0 = c_h\varepsilon/t$, where

$$t = \max\{\sqrt{\varepsilon}, \sqrt{\mathcal{L}/d}\},$$

and (62) gives

$$h_0^2 + 2h_0 \max_i |\langle \theta, u_i\rangle| \le (c_h^2 + 2A_{\mathrm{H}} c_h)\varepsilon.$$

If $\mathcal{L} > d$, then $h_0 = c_h\varepsilon\sqrt{d/\mathcal{L}}$, and the deterministic coordinate bound gives the same inequality with A_{H} replaced by one. Choose $c_h \le 1/8$ so that

$$c_h^2 + 2\max\{A_{\mathrm{H}}, 1\}c_h \le 1, \qquad A_0 c_h^2 \le \frac{\beta_-}{8}. \tag{66}$$

Both query points are then strictly in the active dimple. Equation (63) gives

$$D_i^0 = -\lambda_i \langle \theta, u_i\rangle, \qquad \lambda_i = \frac{2s}{r_i^+ + r_i^-} - 1, \qquad |\lambda_i - b_0| \le A_0 h_0^2, \tag{67}$$

where $r_i^{\pm} = \|\theta \mp h_0 u_i\|$. Hence the noiseless difference vector is

$$D = -b_0\theta + e_0, \qquad \|e_0\| \le A_0 h_0^2 \le \frac{\beta_-}{8}\varepsilon.$$

Write $\widehat{D} = D + \zeta$. Conditional on the basis, ζ is an isotropic Gaussian vector with coordinate variance $(2n_0h_0^2)^{-1}$. On $\mathcal{H}_{\text{init}}$,

$$|\langle\theta, \zeta\rangle| \le \frac{C\varepsilon}{\sqrt{C_{\text{init}}}\, h_0\sqrt{d\mathcal{L}}}, \tag{68}$$

$$\|\zeta\| \le \frac{C\varepsilon}{\sqrt{C_{\text{init}}}\, h_0\mathcal{L}}\left(1 + \sqrt{\mathcal{L}/d}\right). \tag{69}$$

Set

$$t = \min\{1, \max(\sqrt{\varepsilon}, \sqrt{\mathcal{L}/d})\}.$$

If $\mathcal{L} \le d$, then $h_0 = c_h\varepsilon/t$, and

$$\frac{t}{\sqrt{d\mathcal{L}}} \le \varepsilon, \qquad \frac{t^2}{\mathcal{L}}\left(1 + \sqrt{\mathcal{L}/d}\right) \le 2\varepsilon.$$

These inequalities follow by considering separately $t^2 = \varepsilon$ and $t^2 = \mathcal{L}/d$. If $\mathcal{L} > d$, then $t = 1$, $h_0 = c_h\varepsilon\sqrt{d/\mathcal{L}}$, and

$$\frac{1}{d} \le \varepsilon, \qquad \frac{1 + \sqrt{\mathcal{L}/d}}{\sqrt{d\mathcal{L}}} \le \frac{2}{d} \le 2\varepsilon.$$

Consequently, after choosing C_{init} sufficiently large,

$$|\langle\theta, \zeta\rangle| \le \frac{\beta_-}{8}\varepsilon, \qquad \|\zeta\| \le B_{\text{init}}\frac{\varepsilon}{t}$$

for a fixed universal B_{init}. Together with (60) and (66), this yields

$$\langle\theta, -\widehat{D}\rangle \ge \frac{\beta_-}{2}\varepsilon, \qquad \|\widehat{D}\| \le (2 + B_{\text{init}})\frac{\varepsilon}{t}.$$

Thus $\widehat{D} \ne 0$, and, with

$$\kappa_{\text{init}} = \frac{\beta_-}{2(2 + B_{\text{init}})}, \qquad c_{\text{init}} := \kappa_{\text{init}}t,$$

the vector $v_0 = -\widehat{D}/\|\widehat{D}\|$ satisfies

$$\langle\theta, v_0\rangle \ge c_{\text{init}} = \kappa_{\text{init}} \min\{1, \max(\sqrt{\varepsilon}, \sqrt{\mathcal{L}/d})\}. \tag{70}$$

This implies (64).

The stage uses $2dn_0 = O(d^2\mathcal{L}^2/\varepsilon^2)$ queries. Every query point has loss $O(\varepsilon)$, so its regret is $O(d^2\mathcal{L}^2/\varepsilon)$. □

D.2 Tangent amplification

Lemma D.2. *Conditionally on the initialization event, a capped tangent routine reaches $\|\theta - v\| \le q_\star$ with failure probability at most $\delta/3$. With c_{init} defined in (70), the stage uses*

$$N_{\mathrm{amp}} \le \frac{Cd^2\mathcal{L}}{c_{\mathrm{init}}^4}, \qquad R_{\mathrm{amp}} \le \frac{Cd^2\mathcal{L}}{c_{\mathrm{init}}^2}, \tag{71}$$

up to lower-order $C\mathcal{L}(1 + \log(1/c_{\mathrm{init}}))$ gate costs.

Proof. Let

$$\phi_\varepsilon(q) = \varepsilon + \frac{1}{2} - \frac{1}{2}(s-q)_+^2, \qquad \Delta_\star = \phi_\varepsilon(q_\star) - \phi_\varepsilon(q_\star/2).$$

Because $q_\star < s$,

$$\Delta_\star = \frac{sq_\star}{2} - \frac{3q_\star^2}{8} \ge \frac{q_\star}{2} - \frac{3q_\star^2}{8} > 0. \tag{72}$$

The lower bound is universal.

The routine maintains a unit vector v_j and a deterministic number $\underline{c}_j$ satisfying

$$\langle \theta, v_j \rangle \ge \underline{c}_j, \qquad \underline{c}_j = \rho^j c_{\mathrm{init}}, \tag{73}$$

where $\rho > 1$ is fixed below. If $\underline{c}_j > 1 - q_\star^2/8$, then

$$\|\theta - v_j\|^2 = 2\{1 - \langle \theta, v_j \rangle\} < q_\star^2/4,$$

and the routine stops.

Otherwise, average $m_{\mathrm{gate}} = \lceil C_{\mathrm{gate}}\mathcal{L} \rceil$ observations at v_j, and compare the average with

$$T_{\mathrm{gate}} = \frac{\phi_\varepsilon(q_\star/2) + \phi_\varepsilon(q_\star)}{2}.$$

The routine stops when the average is at most T_{gate}, and continues otherwise. After k and ρ are fixed below, choose C_{gate} large enough so that, simultaneously at every gate,

$$|\overline{Y}_j - F_\theta(v_j)| \le \Delta_\star/4$$

on $\mathcal{H}_{\mathrm{amp}}$. Since $F_\theta(v_j) = \phi_\varepsilon(\|\theta - v_j\|)$ and ϕ_ε is nondecreasing, stopping implies

$$\|\theta - v_j\| \le q_\star, \tag{74}$$

whereas continuing implies

$$\|\theta - v_j\| \ge q_\star/2. \tag{75}$$

Suppose the gate continues. Write

$$c = \underline{c}_j, \qquad v = v_j, \qquad p = \langle \theta, v \rangle, \qquad x = (c/2)v, \qquad r_0 = \|\theta - x\|.$$

Then $p \ge c$, and

$$r_0^2 = 1 + \frac{c^2}{4} - cp \le 1 - \frac{3c^2}{4}, \qquad r_0 \ge 1 - \frac{c}{2} \ge \frac{1}{2}. \tag{76}$$

$$c^2 \ge c_{\mathrm{init}}^2 \ge \kappa_{\mathrm{init}}^2 \varepsilon. \tag{77}$$

Draw a Haar orthonormal basis $(u_i)_{i=1}^{d-1}$ of $v^\perp$, and write $q_i = \langle \theta, u_i \rangle$. When $\mathcal{L} \le d-1$, set $h = c_2 c$. On the conditional Haar event,

$$|q_i| \le A_{\mathrm{H}} \sqrt{\mathcal{L}/(d-1)} \le \frac{2A_{\mathrm{H}}}{\kappa_{\mathrm{init}}}\, c.$$

When $\mathcal{L} > d-1$, set $h = c_2 c^2$ and use $|q_i| \le 1$; in this case $\mathcal{L}/d \ge 1/2$, so $c \ge c_{\mathrm{init}} \ge \kappa_{\mathrm{init}}/\sqrt{2}$. Choose $c_2 \le 1/8$ so that

$$c_2^2 + \frac{4A_{\mathrm{H}}}{\kappa_{\mathrm{init}}} c_2 \le \frac{3}{8}, \qquad c_2^2 + 2c_2 \le \frac{3}{8}. \tag{78}$$

In the two cases, respectively,

$$h^2 + 2h|q_i| \le \frac{3c^2}{8}.$$

Together with (76), this shows that $x \pm h u_i$ lie in the active dimple. Their norms are at most one after decreasing c_2, if necessary.

For $r_i^{\pm} = \|\theta - x \mp h u_i\|$, the exact centered difference is

$$D_i = \frac{F_\theta(x + h u_i) - F_\theta(x - h u_i)}{2h} = -\lambda_i q_i, \qquad \lambda_i = \frac{2s}{r_i^+ + r_i^-} - 1. \tag{79}$$

Let $\lambda_0 = s/r_0 - 1$. From (76), for a universal constant $B_0 < \infty$,

$$\lambda_0 \ge 1 - r_0 \ge \frac{3c^2}{8}, \qquad \lambda_0 \le B_0 c, \tag{80}$$

where the upper bound uses $s - 1 \le \varepsilon \le c^2/\kappa_{\mathrm{init}}^2$ and $r_0 \ge 1/2$. A Taylor expansion of $\sqrt{r_0^2 + z}$, uniform over the preceding activity margin, gives

$$|\lambda_i - \lambda_0| \le A_\lambda h^2$$

for a universal A_λ. Decrease c_2 further so that

$$A_\lambda c_2^2 \le \frac{3}{16}. \tag{81}$$

Since $c \le 1$, both choices of h then imply

$$\frac{3c^2}{16} \le \lambda_i \le B_\lambda c \tag{82}$$

for a universal B_λ.

Set

$$g = -\sum_{i=1}^{d-1} D_i u_i = \sum_{i=1}^{d-1} \lambda_i q_i u_i.$$

If $q = \|\theta - v\|$, then $p = 1 - q^2/2$, and (75) gives

$$\sum_i q_i^2 = 1 - p^2 = q^2(1 - q^2/4) \ge q_\star^2/8.$$

Therefore (82) implies

$$\langle \theta, g \rangle \ge c_4 c^2, \qquad \|g\| \le B_\lambda c, \qquad c_4 := \frac{3q_\star^2}{128}. \tag{83}$$

Estimate every D_i with

$$n_c = \lceil C_{\mathrm{amp}} d\mathcal{L}/c^4 \rceil$$

repetitions at each of $x + hu_i$ and $x - hu_i$. With empirical differences $\widehat{D}_i$, define

$$\widehat{g} = -\sum_{i=1}^{d-1} \widehat{D}_i u_i = g + \zeta.$$

Conditional on the basis, ζ is isotropic in $v^\perp$, with coordinate variance $(2n_c h^2)^{-1}$. For $\mathcal{L} \le d-1$, this variance is at most

$$\frac{Cc^2}{C_{\mathrm{amp}} d\mathcal{L}};$$

for $\mathcal{L} > d-1$, it is at most

$$\frac{C}{C_{\mathrm{amp}} d\mathcal{L}},$$

and c is bounded below by a universal multiple of κ_{init}. After k and ρ are fixed below, choose C_{amp} large enough so that, simultaneously at all stages,

$$\|\zeta\| \le c, \qquad |\langle \theta, \zeta \rangle| \le \frac{c_4}{2} c^2 \tag{84}$$

on $\mathcal{H}_{\mathrm{amp}}$.

It follows that $\widehat{g} \neq 0$, and for $u = \widehat{g}/\|\widehat{g}\|$,

$$\langle \theta, u \rangle \ge k_0 c, \qquad k_0 := \frac{c_4}{2(B_\lambda + 1)}.$$

Choose

$$0 < k \le \min\{k_0/2, q_\star/4\}, \qquad \rho = \sqrt{1+k^2} > 1, \tag{85}$$

and set

$$v_{j+1} = \frac{v_j + ku}{\sqrt{1+k^2}}, \qquad \underline{c}_{j+1} = \rho \underline{c}_j.$$

Because $u \perp v_j$,

$$\langle \theta, v_{j+1} \rangle \ge \frac{c + k k_0 c}{\sqrt{1+k^2}} \ge \sqrt{1+k^2}\, c = \underline{c}_{j+1}.$$

Thus (73) is preserved. The restriction $k \le q_\star/4$ also ensures $\rho(1 - q_\star^2/8) < 1$, so the deterministic lower bounds never exceed one before the next stopping check.

After at most

$$1 + \left\lceil \frac{\log\{(1 - q_\star^2/8)/c_{\mathrm{init}}\}}{\log \rho} \right\rceil = O(1 + \log(1/c_{\mathrm{init}}))$$

iterations, either a gate has stopped or the deterministic correlation threshold has been crossed. In both cases the returned vector satisfies $\|\theta - v\| \le q_\star$.

At correlation level c, the tangent queries number $O(d^2\mathcal{L}/c^4)$. By (77),

$$F_\theta(x \pm hu_i) \le \varepsilon + \frac{1}{2}\|x \pm hu_i\|^2 \le Cc^2,$$

so their regret is $O(d^2\mathcal{L}/c^2)$. The gate uses $O(\mathcal{L})$ queries and regret per iteration. Geometric summation over $c_j = \rho^j c_{\mathrm{init}}$ gives

$$N_{\mathrm{amp}} \le \frac{Cd^2\mathcal{L}}{c_{\mathrm{init}}^4} + C\mathcal{L}(1 + \log(1/c_{\mathrm{init}})),$$

and

$$R_{\mathrm{amp}} \le \frac{Cd^2\mathcal{L}}{c_{\mathrm{init}}^2} + C\mathcal{L}(1 + \log(1/c_{\mathrm{init}})).$$

Finally,

$$c_{\mathrm{init}} \ge \kappa_{\mathrm{init}} \max\{\sqrt{\varepsilon}, \sqrt{\mathcal{L}/d}\}$$

unless it is already a fixed constant. This yields (71) and the simplified bounds

$$O(d^2\mathcal{L}/\varepsilon^2) \quad \text{and} \quad O(d^2\mathcal{L}/\varepsilon),$$

respectively. □

D.3 Angular contraction

Lemma D.3. *Conditionally on the first two stage events, the capped contraction stage returns $\widehat{x}$ with $F_\theta(\widehat{x}) \le \gamma_0 a$ outside probability $\delta/3$. It has deterministic query cap*

$$C_\gamma d^2\mathcal{L}/a^2$$

and success-event regret at most

$$C_\gamma d^2\mathcal{L}/a.$$

Proof. Choose

$$0 < c_5 \le \frac{1}{8}, \qquad 0 < c_{\mathrm{term}} \le \frac{\gamma_0}{s_+}. \tag{86}$$

Let

$$Q_j = q_\star 2^{-j}, \qquad J = \min\{j \ge 0 : Q_j \le c_{\mathrm{term}} a\}.$$

The routine maintains the invariant

$$\|v_j\| = 1, \qquad \|\theta - v_j\| \le Q_j. \tag{87}$$

The amplification lemma establishes it at $j = 0$. For each $j < J$, write $v = v_j$, $Q = Q_j$, set $h = c_5 Q$, and take a measurably chosen orthonormal basis $(u_i)_{i=1}^{d-1}$ of $v^\perp$. Define

$$\alpha_h = (1 + h^2)^{-1/2}, \qquad t = h(1 + h^2)^{-1/2}, \qquad x_i^\pm = \alpha_h v \pm t u_i.$$

Then $x_i^\pm$ are unit vectors, $t \ge c_5 Q/\sqrt{2}$, and

$$\|x_i^\pm - v\|^2 = 2(1 - \alpha_h) \le h^2.$$

Consequently,

$$\|\theta - x_i^\pm\| \le (1 + c_5)Q \le C_{\mathrm{geo}} Q \le C_{\mathrm{geo}} q_\star \le s/2. \tag{88}$$

Thus every probe lies in the active dimple.

Write

$$\theta = c_v v + w, \qquad c_v = \langle \theta, v \rangle = 1 - \frac{1}{2}\|\theta - v\|^2, \qquad w = \sum_i q_i u_i.$$

Since $Q \le q_\star \le 1/8$, one has $c_v > 0$,

$$c_v = \sqrt{1 - \|w\|^2}, \qquad \|w\| \le \|\theta - v\| \le Q < 1/2. \tag{89}$$

Let

$$\psi(r) = sr - \frac{1}{2}r^2, \qquad G(p) = \psi(\sqrt{2-2p}).$$

Define

$$I_Q = \{p \in [-1,1] : \sqrt{2-2p} \le C_{\mathrm{geo}} Q\}.$$

By (88), the inner products

$$p_i^{\pm} = \langle \theta, x_i^{\pm} \rangle = \alpha_h c_v \pm t q_i$$

belong to I_Q, and $F_\theta(x_i^{\pm}) = G(p_i^{\pm})$. The restriction $G_Q := G|_{I_Q}$ is strictly decreasing. Moreover, for $p \in I_Q$,

$$|(G_Q^{-1})'(G(p))| = \frac{\sqrt{2-2p}}{s - \sqrt{2-2p}} \le 2C_{\mathrm{geo}} Q. \tag{90}$$

The right side tends to zero at $p = 1$, so this estimate includes the endpoint by continuity. Set

$$B_q := \frac{2\sqrt{2} C_{\mathrm{geo}}}{c_5},$$

and choose $c_6 > 0$ so that

$$2B_q c_6 \le \frac{1}{2}. \tag{91}$$

Let $\overline{Y}_i^{\pm}$ be the sample means at $x_i^{\pm}$, and let $\widetilde{Y}_i^{\pm}$ be their Euclidean projections onto the interval $G(I_Q)$. Define

$$\widehat{q}_i = \frac{G_Q^{-1}(\widetilde{Y}_i^{+}) - G_Q^{-1}(\widetilde{Y}_i^{-})}{2t}, \qquad \widehat{w} = \sum_i \widehat{q}_i u_i.$$

Use

$$n_Q = \lceil C_{\mathrm{con}} d\mathcal{L}/Q^2 \rceil$$

repetitions at each site. Choose C_{con} so that, simultaneously for all i and all $j < J$,

$$|\overline{Y}_i^{\pm} - F_\theta(x_i^{\pm})| \le \frac{c_6 Q}{\sqrt{d}} \tag{92}$$

on $\mathcal{H}_{\mathrm{con}}$. Projection onto $G(I_Q)$ cannot increase the error because the true value belongs to this interval. Combining (90), (92), and $t \ge c_5 Q/\sqrt{2}$ gives

$$|\widehat{q}_i - q_i| \le B_q \frac{c_6 Q}{\sqrt{d}},$$

and hence

$$\|\widehat{w} - w\| \le B_q c_6 Q. \tag{93}$$

Let $\widetilde{w}$ be the Euclidean projection of $\widehat{w}$ onto the ball of radius $1/2$. Since the true w belongs to this ball by (89),

$$\|\widetilde{w} - w\| \le \|\widehat{w} - w\|.$$

Set

$$v_{j+1} = \sqrt{1 - \|\widetilde{w}\|^2}\, v + \widetilde{w}.$$

The map

$$z \longmapsto \sqrt{1 - \|z\|^2}\, v + z$$

is 2-Lipschitz on the ball of radius 1/2. Using $\theta = \sqrt{1 - \|w\|^2}\, v + w$, we obtain

$$\|\theta - v_{j+1}\| \le 2\|\widetilde{w} - w\| \le Q/2 = Q_{j+1}.$$

This proves (87) by induction.

The number of queries at scale Q_j is $O(d^2\mathcal{L}/Q_j^2)$. By (88),

$$F_\theta(x_i^\pm) = \psi(\|\theta - x_i^\pm\|) \le s_+ C_{\mathrm{geo}} Q_j,$$

so the corresponding regret is $O(d^2\mathcal{L}/Q_j)$. If $J = 0$, this stage makes no queries. Otherwise, Q_j decreases geometrically and $Q_{J-1} > c_{\mathrm{term}} a$, so

$$\sum_{j<J} \frac{d^2\mathcal{L}}{Q_j^2} \le \frac{Cd^2\mathcal{L}}{c_{\mathrm{term}}^2 a^2}, \qquad \sum_{j<J} \frac{d^2\mathcal{L}}{Q_j} \le \frac{Cd^2\mathcal{L}}{c_{\mathrm{term}} a}.$$

The output $\widehat{x} = v_J$ then satisfies

$$F_\theta(\widehat{x}) = \psi(\|\theta - v_J\|) \le s_+\|\theta - v_J\| \le s_+ c_{\mathrm{term}} a \le \gamma_0 a.$$

The number J and every repetition count are deterministic. Continuing the same clipped updates off $\mathcal{H}_{\mathrm{con}}$ therefore gives the claimed deterministic cap. □

Proof of Theorem 5.3. Run Lemmas D.1–D.3 in order, assigning failure probability $\delta/3$ to each. The three conditional failure bounds imply $\sup_\theta \mathbb{P}_\theta(\mathcal{H}^c) \le \delta = c_0 a$. The amplification lower bounds and the angular invariant are asserted only on $\mathcal{H}$; the deterministic caps, clipping rules, and fixed iteration limits define the same routine on $\mathcal{H}^c$. Because $a \le b_0 \asymp \varepsilon$, their deterministic caps and geometric sums give

$$N \le C_\gamma d^2 L_{d,a}^2 / a^2$$

and, on $\mathcal{H}$,

$$R_N \le C_\gamma d^2 L_{d,a}^2 / a.$$

On $\mathcal{H}^c$, the capped routine uses no more than its displayed query budget and incurs at most one unit of regret per query. Since $\mathbb{P}(\mathcal{H}^c) \le \delta = c_\delta a$, this adds at most

$$C_\gamma d^2 L_{d,a}^2 / a$$

to expected regret. The accuracy statement follows from Lemma D.3 and the union bound. Since losses are bounded by one, the same statement gives, under every prior,

$$\mathbb{E} F_\Theta(\widehat{x}) \le (\gamma_0 + c_0) a.$$

Apply Markov's inequality to the nonnegative random variable $\mathbb{E}[F_\Theta(\widehat{x}) \mid H_N]$ to obtain (50). On the displayed event, the posterior Bayes risk is no larger than the posterior loss of $\widehat{x}$, and hence is at most γa. □

E Gaussian models and Bayes–minimax equality

This appendix proves Theorem 4.1. It establishes measurable implementation and the order of the Bayes and minimax optimizations; the two-sided stopped reduction itself was proved in Section 2.

Lemma E.1 (Prior-wise Borel realization)**.** *Fix a Borel prior Q and a finite horizon T. On standard Borel history, action, and observation spaces, every universally measurable randomized nonanticipating policy has a Borel randomized nonanticipating policy inducing the same joint law of (F, H_T) under Q. The same holds when the action space is augmented by a stopping symbol.*

Proof. At time t, represent the action kernel as a universally measurable map from histories into the standard Borel space of action distributions. Under the reached history law, this map has a Borel version. Replace it by that version; the strategic law through time t is unchanged. Induction over $t = 1, \dots, T$ proves the claim. Adding one isolated stopping symbol preserves the standard Borel property. □

E.1 Finite-horizon Bayes–minimax identity

Let $\nu = N(0,1)$, let $\mathsf{H}_T = (C \times \mathbb{R})^T$, and equip $\mathcal{P}(\mathsf{H}_T)$ with the topology of weak convergence. Let $\mathcal{P}_T$ be the set of reference strategic measures

$$P(dh_T) = \prod_{t=1}^{T} \pi_t(dx_t \mid h_{t-1})\, \nu(dy_t)$$

generated by Borel randomized nonanticipating policies.

Lemma E.2 (Compactness of reference strategic measures)**.** *Suppose that C is a nonempty compact metric space. Then $\mathcal{P}_T$ is a nonempty convex weakly compact subset of $\mathcal{P}(\mathsf{H}_T)$.*

More precisely, a probability measure P on H_T belongs to $\mathcal{P}_T$ if and only if, for every $t \le T$, $a \in C_b((C \times \mathbb{R})^{t-1} \times C)$, and $b \in C_b(\mathbb{R})$,

$$\int a(h_{t-1}, x_t) b(y_t)\, dP = \left(\int b\, d\nu \right) \int a(h_{t-1}, x_t)\, dP \tag{94}$$

Proof. Every reference strategic measure satisfies (94) by iterated conditioning. Conversely, the bounded-continuous identities extend, by a measure-determining and monotone-class argument on Polish spaces, to bounded Borel a and b. They therefore imply

$$P(Y_t \in \cdot \mid H_{t-1}, X_t) = \nu(\cdot) \qquad P\text{-a.s.}$$

Because all spaces are standard Borel, P has Borel regular conditional distributions

$$\pi_t(\cdot \mid h_{t-1}) = P(X_t \in \cdot \mid H_{t-1} = h_{t-1}).$$

Successive disintegration, together with the displayed conditional law of Y_t, gives

$$P(dh_T) = \prod_{t=1}^{T} \pi_t(dx_t \mid h_{t-1})\, \nu(dy_t),$$

so $P \in \mathcal{P}_T$. This proves the characterization.

The characterization is linear in P, so $\mathcal{P}_T$ is convex. It also implies that the $Y_{1:T}$-marginal of every $P \in \mathcal{P}_T$ is $\nu^{\otimes T}$. Since C^T is compact, this common observation marginal makes $\mathcal{P}_T$ tight.

If $P_n \in \mathcal{P}_T$ and $P_n \Rightarrow P$, then both sides of (94) converge for every bounded continuous a, b. Hence P satisfies the same identities and, by the characterization, belongs to $\mathcal{P}_T$. Thus $\mathcal{P}_T$ is weakly closed. Prokhorov's theorem now gives weak compactness. □

For $f \in \mathcal{C}(C,[0,1])$, write

$$m(f) = \min_{x \in C} f(x), \qquad G_f(h_T) = \sum_{t=1}^{T} \{f(x_t) - m(f)\},$$

and

$$L_f(h_T) = \exp\left\{ \sum_{t=1}^{T} \left(y_t f(x_t) - \frac{1}{2} f(x_t)^2 \right) \right\},$$

and define

$$u(P,f) = \int L_f(h_T) G_f(h_T)\, P(dh_T).$$

Lemma E.3 (Continuity of the Gaussian regret payoff)**.** *The map*

$$(P,f) \longmapsto u(P,f)$$

is jointly continuous on $\mathcal{P}_T \times \mathcal{C}(C,[0,1])$, where the first space has the weak topology and the second has the uniform topology. It is affine in P and takes values in $[0,T]$.

If P is the reference strategic measure generated by a policy π, then

$$L_f = \frac{dP_f^\pi}{dP} \quad \text{on } \mathsf{H}_T, \qquad u(P,f) = \mathbb{E}_f^\pi R_T.$$

Proof. The likelihood-ratio identity follows by multiplying the one-step density ratios

$$\frac{\varphi(y_t - f(x_t))}{\varphi(y_t)} = \exp\left\{ y_t f(x_t) - \frac{1}{2} f(x_t)^2 \right\}.$$

The action kernels cancel because the same policy is used under the reference and shifted observation laws. Thus $L_f P$ is the actual strategic law, and

$$u(P,f) = \mathbb{E}_f^\pi R_T \in [0,T].$$

Fix f. The integrand $L_f G_f$ is continuous on H_T, but it is unbounded. Put

$$S(y_{1:T}) = \sum_{t=1}^{T} |y_t|.$$

Since $0 \le G_f \le T$,

$$0 \le L_f(h_T) G_f(h_T) \le T e^{S(y_{1:T})}. \tag{95}$$

The right-hand side is integrable under $\nu^{\otimes T}$, and every $P \in \mathcal{P}_T$ has this same observation marginal. Let $\chi_M \in C_c(\mathbb{R}^T)$ satisfy $0 \le \chi_M \le 1$ and $\chi_M(y) \to 1$ pointwise. Then $\chi_M(y_{1:T}) L_f(h_T) G_f(h_T)$ is bounded and continuous, while

$$\sup_{P \in \mathcal{P}_T} \int (1 - \chi_M) L_f G_f \, dP \le T \int (1 - \chi_M(y)) e^{S(y)}\, \nu^{\otimes T}(dy) \longrightarrow 0.$$

It follows that weak convergence $P_n \Rightarrow P$ implies $u(P_n, f) \to u(P, f)$.

It remains to control changes in f, uniformly over P. Let $\delta = \|f - f'\|_\infty$. Since

$$|m(f) - m(f')| \le \delta, \qquad |G_f - G_{f'}| \le 2T\delta,$$

and, upon writing $A_f = \log L_f$,

$$|A_f - A_{f'}| \le \delta\{S(y_{1:T}) + T\}, \qquad A_f, A_{f'} \le S(y_{1:T}),$$

the mean-value theorem gives

$$|L_f G_f - L_{f'} G_{f'}| \le T\delta e^{S(y_{1:T})}\{S(y_{1:T}) + T + 2\}.$$

Therefore

$$\sup_{P \in \mathcal{P}_T} |u(P, f) - u(P, f')| \le K_T \|f - f'\|_\infty,$$

where

$$K_T = T \int e^{S(y)}\{S(y) + T + 2\}\, \nu^{\otimes T}(dy) < \infty.$$

Combining this uniform estimate with continuity in P for fixed f proves joint continuity. Affinity in P is immediate. □

Proof of (34). Fix $C \in \mathcal{C}_d$ and abbreviate $\mathfrak{F} = \mathfrak{F}_d(C)$. Let $\mathcal{V}_{\mathrm{fs}}$ be the real vector space of finitely supported signed measures on $\mathfrak{F}$, equipped with the locally convex topology generated by the seminorms

$$q \longmapsto \left| \int \phi(f)\, q(df) \right|, \qquad \phi \in C_b(\mathfrak{F}).$$

Let $\mathcal{Q}_0 \subset \mathcal{V}_{\mathrm{fs}}$ be the convex set of finitely supported probability measures, and define

$$U(P, Q) = \int_{\mathfrak{F}} u(P, f)\, Q(df).$$

By Lemma E.3, $f \mapsto u(P, f)$ is bounded and continuous. Hence $Q \mapsto U(P, Q)$ is continuous and affine on $\mathcal{Q}_0$. For each $Q \in \mathcal{Q}_0$, $P \mapsto U(P, Q)$ is a finite convex combination of continuous affine functions and is therefore continuous and affine. Lemma E.2 makes $\mathcal{P}_T$ compact and convex. Sion's theorem (Sion, 1958) gives

$$\min_{P \in \mathcal{P}_T} \sup_{Q \in \mathcal{Q}_0} U(P, Q) = \sup_{Q \in \mathcal{Q}_0} \min_{P \in \mathcal{P}_T} U(P, Q).$$

Because every Dirac measure belongs to $\mathcal{Q}_0$,

$$\sup_{Q \in \mathcal{Q}_0} U(P, Q) = \sup_{f \in \mathfrak{F}} u(P, f).$$

The preceding lemmas identify every $P \in \mathcal{P}_T$ with a Borel randomized nonanticipating policy and identify $u(P, f)$ with that policy's regret under f. Consequently,

$$\inf_{\pi} \sup_{f \in \mathfrak{F}} \mathbb{E}_f^\pi R_T = \sup_{Q \in \mathcal{Q}_0} \inf_{\pi} \mathbb{E}_Q^\pi R_T. \tag{96}$$

For an arbitrary Borel prior Q on $\mathfrak{F}$, boundedness and Borel measurability of $u(P, \cdot)$ give

$$\min_{P \in \mathcal{P}_T} \int u(P, f)\, Q(df) \le \min_{P \in \mathcal{P}_T} \sup_{f \in \mathfrak{F}} u(P, f).$$

Taking the supremum over all Borel priors cannot exceed the minimax value, whereas, by (96), the supremum over finitely supported priors already equals the minimax value. Hence

$$\inf_{\pi} \sup_{f \in \mathfrak{F}} \mathbb{E}_f^\pi R_T = \sup_{Q \text{ on } \mathfrak{F}} \inf_{\pi} \mathbb{E}_Q^\pi R_T,$$

and the supremum on the right may be restricted to finitely supported priors. Taking the outer supremum over $C \in \mathcal{C}_d$ proves (34). □

E.2 Measurable selection of stopped experiments

Lemma E.4 (Capped Gaussian selection)**.** *Fix a query cap $L < \infty$. The capped stopped value*

$$v_L(Q, b) = \inf_{\mathcal{E}: N \le L} \frac{b\,\mathbb{E}_Q R_N}{\mathbb{P}_Q(\mathsf{M})}$$

is universally measurable. For every $\epsilon > 0$, there is a universally measurable state-to-experiment rule with value at most $v_L(Q, b) + \epsilon$ at every active state.

Proof. The belief space $\mathcal{P}(\mathfrak{F}_d(C))$ is standard Borel. For a belief μ, define

$$\ell(\mu, x) = \int \{f(x) - \min_C f\}\, \mu(df), \qquad r(\mu) = \min_{x \in C} \ell(\mu, x).$$

The first map is Borel and continuous in x; compactness of C makes r Borel and supplies a Borel Bayes-action selector. With φ the standard Gaussian density, the posterior update is the Borel map

$$\Psi(\mu, x, o)(df) = \frac{\varphi(o - f(x))\, \mu(df)}{\int \varphi(o - f'(x))\, \mu(df')}.$$

For a rational price $q > 0$, the capped penalized problem is exactly the recursion (10), with $\mu' = \Psi(\mu, x, o)$. Its one-step costs are bounded below and all primitive kernels are Borel. Standard integration, partial minimization, and universally measurable selection for lower-semianalytic dynamic programs therefore apply (Bertsekas and Shreve, 1978, Propositions 7.47–7.50). In particular, its root value $J_L^q(\mu, b)$ is lower semianalytic and has universally measurable near-minimizers.

Directly from the definition,

$$v_L(\mu, b) < q \quad \Longleftrightarrow \quad J_L^q(\mu, b) < 0.$$

The rational strict sublevel sets make v_L universally measurable. On $\{v_L < +\infty\}$, enumerate the positive rationals and measurably choose the first q with

$$v_L + \epsilon/4 < q < v_L + \epsilon/2.$$

Then $J_L^q < 0$. Partition this set further by the first integer n such that $J_L^q \le -1/n$, and use a Bellman selector whose approximation error is less than $1/(2n)$. The resulting penalized value remains negative, so its success probability is positive and its ratio is below $q < v_L + \epsilon/2$. On $\{v_L = +\infty\}$, any measurable one-query rule satisfies the extended-value approximation. Countable patching gives one universally measurable ϵ-optimal rule. □

Lemma E.5 (Removing the query cap)**.** *The unrestricted value $\mathsf{SC}_b(Q)$ is universally measurable and has a universally measurable ϵ-optimal selector for every $\epsilon > 0$. The selected experiment may have a finite deterministic cap $L(Q, b)$ depending on the state.*

Proof. Truncate an admissible experiment after L queries. Its stopped regret and hit probability increase to those of the complete experiment:

$$\mathbb{E}R_{N \wedge L} \uparrow \mathbb{E}R_N, \qquad \mathbb{P}(\mathsf{M}_L) \uparrow \mathbb{P}(\mathsf{M}).$$

Hence its ratio converges, including the zero-success case under the $+\infty$ convention. Indeed, if $\mathbb{P}(\mathsf{M}) > 0$, componentwise convergence gives convergence of the ratios; if $\mathbb{P}(\mathsf{M}) = 0$, then every truncated hit probability is zero and all ratios are $+\infty$. Since capped experiments are a subset of

all experiments, $\mathsf{SC}_b(Q) \le v_L(Q,b)$ for every L. Conversely, truncating any admissible experiment and passing to the limit gives the reverse inequality. Therefore

$$\mathsf{SC}_b(Q) = \inf_{L\ge 1} v_L(Q,b).$$

This is universally measurable by Lemma E.4. On $\{\mathsf{SC}_b < +\infty\}$, choose measurably the first L for which $v_L < \mathsf{SC}_b + \epsilon/2$, then use the capped $\epsilon/2$-selector. On $\{\mathsf{SC}_b = +\infty\}$, use any measurable one-query rule; the approximation statement is vacuous there. Countable patching proves the claim. □

Proof of (35). Lemma E.5 selects, simultaneously at every active state, an experiment within ϵ of its pointwise infimum. Taking the worst state and then letting $\epsilon \downarrow 0$ gives $\mathsf{SC}^{\mathrm{unif}} \le \mathsf{SC}^{\mathrm{pt}}$. The reverse inequality is immediate from the definitions. For each fixed prior and finite global horizon, Lemma E.1 replaces the universal selector by a Borel policy with the same strategic law. □

Acknowledgements

The author used ChatGPT and Codex as research and editorial aids. The author proposed the core ideas and assumes full responsibility for the statements and proofs in the paper.

References

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 373–382. PMLR, 2021. URL https://proceedings.mlr.press/v139/arumugam21a.html.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017. URL https://proceedings.mlr.press/v70/azar17a.html.

Alireza Bakhtiari, Tor Lattimore, and Csaba Szepesvári. Thompson sampling for bandit convex optimisation. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 231–263. PMLR, 2025. URL https://proceedings.mlr.press/v291/bakhtiari25a.html.

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, New York, 1978.

Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991. doi: 10.1287/moor.16.3.580. URL https://web.mit.edu/dimitrib/www/Stochasticsp.pdf.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Sébastien Bubeck and Mark Sellke. First-order Bayesian regret analysis of Thompson sampling. *IEEE Transactions on Information Theory*, 69(3):1795–1823, 2023. URL https://ieeexplore.ieee.org/document/9915622/.

Alexandra Carpentier. A simple and improved algorithm for noisy, convex, zeroth-order optimisation. *Mathematical Statistics and Learning*, 8:165–192, 2025. doi: 10.4171/MSL/54. URL https://ems.press/content/serial-article-files/51581.

Fan Chen, Dylan J. Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. Assouad, Fano, and Le Cam with interaction: A unifying lower bound framework and characterization for bandit learnability. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2407. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8a23a95e26d016711c0d70f79ade3c95-Abstract-Conference.html.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30:755–770, 1959. doi: 10.1214/aoms/1177706205. URL https://doi.org/10.1214/aoms/1177706205.

Morris H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962. doi: 10.1214/aoms/1177704567. URL https://doi.org/10.1214/aoms/1177704567.

Shi Dong and Benjamin Van Roy. An information-theoretic analysis for Thompson sampling with many actions. In *Advances in Neural Information Processing Systems 31*, pages 4161–4169, 2018. URL https://arxiv.org/abs/1805.11845.

Hidde Fokkema, Dirk van der Hoeven, Tor Lattimore, and Jack J. Mayo. Online Newton method for bandit convex optimisation. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1713–1714. PMLR, 2024. URL https://proceedings.mlr.press/v247/fokkema24a.html. Full version: arXiv:2406.06506.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making, 2021. URL https://arxiv.org/abs/2112.13487.

Dylan J. Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. In *Advances in Neural Information Processing Systems*, volume 35, pages 35404–35417, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e5daf532498ebba4fe6544d49c811678-Abstract-Conference.html.

Dylan J. Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3969–4043. PMLR, 2023. URL https://proceedings.mlr.press/v195/foster23b.html.

Amaury Gouverneur, Borja Rodríguez-Gálvez, Tobias J. Oechtering, and Mikael Skoglund. Chained information-theoretic bounds and tight regret rate for linear bandit problems, 2024. URL https://arxiv.org/abs/2403.03361.

Tor Lattimore and András György. Mirror descent and the information ratio. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2965–2992. PMLR, 2021. URL https://proceedings.mlr.press/v134/lattimore21b.html.

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*, pages 2067–2094. PMLR, 2023. URL https://proceedings.mlr.press/v195/lattimore23a.html.

Shaojie Li and Yunbei Xu. Pointwise generalization in deep neural networks, 2026. URL https://arxiv.org/abs/2605.18598.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069. URL https://doi.org/10.1214/aoms/1177728069.

Arnab Maiti, Yunbei Xu, and Kevin Jamieson. On the power of adaptivity for ε-best arm identification in linear bandits. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 4911–4968. PMLR, 2026. URL https://proceedings.mlr.press/v336/maiti26a.html.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013. doi: 10.1214/13-AOS1144. URL https://doi.org/10.1214/13-AOS1144.

Gergely Neu, Matteo Papini, and Ludovic Schwartz. Optimistic information directed sampling. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 3970–4006. PMLR, 2024. URL https://proceedings.mlr.press/v247/neu24a.html.

Nived Rajaraman. The price of hidden curvature: An $\widetilde{\Omega}(d^{5/4}\sqrt{T})$ lower bound for bandit convex optimization, 2026. URL https://arxiv.org/abs/2607.18652.

Idan Rejwan and Yishay Mansour. Top-k combinatorial bandits with full-bandit feedback. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 752–776. PMLR, 2020. URL https://proceedings.mlr.press/v117/rejwan20a.html.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016. URL https://jmlr.org/papers/v17/14-087.html.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018. doi: 10.1287/opre.2017.1663. URL https://pubsonline.informs.org/doi/10.1287/opre.2017.1663.

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 47(4):2815–2839, 2022. doi: 10.1287/moor.2021.1229. URL https://doi.org/10.1287/moor.2021.1229.

Rajat Sen, Alexander Rakhlin, Lexing Ying, Rahul Kidambi, Dean Foster, Daniel N. Hill, and Inderjit S. Dhillon. Top-k extreme contextual bandits with arm hierarchy. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9422–9433. PMLR, 2021. URL https://proceedings.mlr.press/v139/sen21a.html.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171. URL https://doi.org/10.2140/pjm.1958.8.171.

Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d9d3837ee7981e8c064774da6cdd98bf-Abstract.html.

Yunbei Xu. Bellman-sufficient information complexity, 2026a. URL https://arxiv.org/abs/2606.11171.

Yunbei Xu. The dimension of nonterminating resampling computations. *arXiv preprint arXiv:2607.17469*, 2026b. doi: 10.48550/arXiv.2607.17469. URL https://arxiv.org/abs/2607.17469.

Yunbei Xu. Pointwise complexity for Gaussian fields: Upper envelopes, algorithmic lower bounds, and separation, 2026c. URL https://arxiv.org/abs/2606.07931.

Yunbei Xu and Assaf Zeevi. Bayesian design principles for frequentist sequential learning. *Journal of the ACM*, 72(5):37:1–37:65, October 2025. doi: 10.1145/3766898. URL https://dl.acm.org/doi/10.1145/3766898.