Bellman-sufficient Information Complexity

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

Abstract

We introduce Bellman-sufficient information complexity for minimax analysis of sequential decision problems. A Bellman-sufficient state retains enough of the history to close the controlled recursion, while an index $Y = \chi(\Omega)$ specifies the decision-relevant information being charged. The upper bound is a log-penalized Bellman program; the lower bound is a Bellman–Fano comparison along an algorithm-dependent reference trajectory. If the two values match at a common localization scale and the stated admissibility, calibration, and growth conditions hold, they form an information-risk sandwich. UCB, E2D, and AMS/EBO control or relax the upper Bellman bracket in different ways.

For the main application, we give a negative answer to a widely studied form of the GP–UCB minimax-optimality question. For every $0 < \alpha < 1/4$, we construct one bounded continuous kernel whose minimax regret is $\Theta(T^{1-\alpha})$ along an infinite sequence of horizons, while two globally calibrated GP–UCB rules incur linear regret under one fixed truth. An epochwise finite-marginal action-index AIR Bellman policy, implemented through robust AIR/AMS/EBO control, attains the minimax order. The construction separates realized information from the cost of uniform optimism: many low-value directions inflate the exploration multiplier and change the trajectory. Through the canonical RKHS feature map, it also yields a finite-horizon polynomial minimax separation for the specified maximal-information-calibrated LinUCB rule. A reproducible experiment illustrates the mechanism.

Contents

1 Introduction

Information-theoretic minimax theory asks how much risk remains when an experiment can reveal only limited information. In classical noninteractive estimation the experiment is fixed before the data are observed, and sharp rates are obtained by matching an upper information-risk construction with a lower entropy calculation: KL information measures what the experiment can distinguish, while local prior mass or packing entropy measures how many alternatives remain statistically indistinguishable. This is the perspective behind the information-risk upper and lower bounds of Zhang (2006), the entropy characterization of minimax rates in Yang and Barron (1999), and the classical Fano–Assouad method (Yu, 1997). In interactive decision making and dynamic control, the learner's actions change the experiment. Because the policy is nonanticipating, the experiment adapts to the past. Decisions, observations, and information accumulation are therefore produced jointly by one controlled stochastic process.

Algorithmic Fano's method is one way to study this adaptivity. It compares the real history with an algorithm-dependent reference, or "ghost," history rather than with a fixed nonadaptive sampling scheme. This viewpoint appears in recent interactive Fano frameworks for interactive decision making (Chen et al., 2024) and in algorithmic upper and lower bounds for representation learning (Li and Xu, 2026; Xu, 2026). It is also closely related to the decision-estimation coefficient (DEC) approach of Foster et al. (2021), its constrained or localized refinements (Foster et al., 2023), and the earlier information-ratio approach (Russo and Van Roy, 2014; Lattimore and György, 2021). For the dynamic comparisons pursued here, it is useful to retain the T-round reference geometry until the necessary localization has been identified. This preserves the evolution of the posterior, confidence set, or reference state, complementing analyses based on one-step complexity measures.

Kernel bandits provide the main application. GP–UCB chooses an action by adding a confidence bonus to a Gaussian-process posterior mean. Whether the gap between its known guarantees and kernel-bandit minimax rates reflects the analysis or the acquisition rule has remained a prominent open question (Vakili et al., 2021; Whitehouse et al., 2023; Wang and Zhang, 2026). We show that maximal information gain can alter the learner's trajectory: many individually negligible directions enlarge the global confidence multiplier and draw GP–UCB into a low-value cloud.

The construction starts from the weighted-basis hub–cloud geometry of Xu (2026) and reverses its amplitude ordering, so that the cloud is negligible for minimax regret but still controls the exploration multiplier. Under the canonical RKHS feature map, GP posterior means and standard deviations coincide with ridge predictions and ellipsoidal widths. The finite blocks therefore give a polynomial separation for the specified maximal-information-calibrated LinUCB rule; gluing the blocks gives one fixed bounded continuous kernel on which the two GP–UCB schedules studied here incur linear regret along an infinite subsequence. A predetermined finite-marginal action-index AIR Bellman policy, implemented through robust AIR/AMS control, attains the minimax order. The main theorem states how this result relates to the original GP–UCB schedule, the Matérn results, and the remaining open cases.

We organize the analysis around an indexed Bellman-sufficient representation. The environment

remains a fixed frequentist truth. A state S_t compresses the history while retaining the feasible actions, predictive laws, losses, and future values needed for a Bellman recursion. An index $Y = \chi(\Omega)$ identifies the information worth learning—for example, an optimal action or policy, an active finite marginal, or the full environment. The state closes the dynamics; the index determines the information charge.

On the upper side, the coordinate potential $\gamma \log(1/q_t(y^\star))$ turns the fixed-truth AIR bracket into a Bellman telescope. The resulting dynamic program is the information-theoretic upper object. UCB, E2D, and AMS/EBO control or relax this object in different ways: through calibrated optimism, a one-step offset, and robust belief optimization, respectively. These connections are useful, but the algorithms and their optimization problems remain distinct.

On the lower side, Bellman–Fano compares the true trajectory with an algorithm-dependent reference trajectory and counts reference histories that are already good for a randomly sampled index. Subject to the stated admissibility, calibration, and growth conditions, the upper and lower values match when the coordinate code length and ghost-good entropy have the same order at a common localization scale. This is the sequential analogue of matching information and local entropy in a fixed experiment, except that the learner now controls the experiment and both quantities evolve through the same Bellman state.

A related benefit is that some of the framework's statements are general enough to include reinforcement learning across episodes and in continuing environments (see, e.g., Section 1.2 of Littman (1996)), although this aspect is treated only conceptually here.

1.1 Main contributions.

1. We define indexed Bellman-sufficient representations for general nonanticipating sequential decision problems. The state closes the controlled recursion, while the index specifies the decision-relevant information being charged.

2. We prove the indexed posterior-reference chain rule and exact fixed-truth AIR identity, then derive a logarithmically penalized Bellman upper program. AIR and MAIR follow by choosing the optimal decision and the environment, respectively, as the index.

3. We develop an algorithm-uniform Bellman–Fano lower bound and an information-risk sandwich. The comparison places the upper coordinate code length and the entropy of low-loss ghost histories on the same state/index representation.

4. For kernel bandits, we construct one fixed bounded continuous kernel on which two globally calibrated GP–UCB rules have linear regret along an infinite sequence of horizons, although the minimax regret is $\Theta(T^{1-\alpha})$ there and an epochwise finite-marginal AIR/AMS policy attains this order. The same geometry gives, for each sufficiently large T, $T+1$ linear-bandit actions in $\mathbb{R}^T$ on which the specified globally calibrated LinUCB rule is worse than the unrestricted minimax value by $\Omega(T^\alpha)$. Gaussian multi-armed bandits and hypercube linear bandits give further matching examples.

1.2 Organization.

Section 2 defines the sequential model, information indices, and Bellman-sufficient representations. Section 3 gives the indexed information telescope and the exact AIR/MAIR identity. Section 4 develops the log-penalized Bellman upper program, and Section 5 gives the quantile and Bellman–Fano lower bounds. Section 5.4 then compares the two sides at a common entropy scale. Section 6

turns to kernel bandits: it proves the finite-marginal AIR bound and the fixed-kernel GP–UCB separation, presents the numerical illustration, and states the finite-dimensional LinUCB consequence and the remaining GP–UCB questions. The appendices contain the formal entropy diagnostics, algorithm-family comparisons, full linear and Matérn statements, additional examples, and proofs.

2 Sequential decision making with Bellman-sufficient representations

2.1 Sequential decision making and Bellman state compression

The primitive object is a controlled process over histories: future actions, observations, and losses may depend on everything revealed so far. This unified framework for sequential decision making encompasses bandits, episodic and continuing reinforcement learning, and planning and reasoning from experience. (Barto et al., 1989; Littman, 1996; Sutton and Barto, 2018; Lattimore and Szepesvári, 2020; Abel et al., 2023; Silver and Sutton, 2025). The formalism below clarifies the history, controls, losses, and Bellman state underlying these problems.

A formal model of the sequential decision-making problem. There is an environment class Ω, and performance is required for every fixed truth $\omega^\star \in \Omega$. A prior μ is introduced only when we form a Bayes value, a Yao lower bound, a posterior-reference trajectory, or an algorithmic belief. The primitive history used in the paper is the full pre-action history, denoted H_{t-1}. To avoid separate notation for contexts, physical states, public randomization, and protocol variables, we regard observations as packets. There may be an initial packet O_0 before the first decision, and recursively

$$H_0 = O_0, \qquad H_t = (H_{t-1}, A_t, O_t).$$

The packet O_t contains whatever is publicly revealed after the round-t action and before the next decision: feedback, the next context or physical state, the next stage marker, public randomization, or other protocol information. The learner's decision rule is the policy being optimized and is not itself part of the history; its randomization is represented by the decision kernels below.

At time t, after $h \in \mathcal{H}_{t-1}$, the learner may choose an action in a measurable set $\mathcal{A}_t(h)$. An unrestricted nonanticipating algorithm is a sequence of kernels

$$\pi_t^{\mathsf{Alg}}(\cdot \mid h) \in \Delta(\mathcal{A}_t(h)), \qquad h \in \mathcal{H}_{t-1},$$

with deterministic algorithms included as degenerate kernels. Let $\mathfrak{A}_T^{\mathrm{na}}$ denote the class of all such algorithms, and write

$$p_t = \pi_t^{\mathsf{Alg}}(\cdot \mid H_{t-1}), \qquad A_t \sim p_t.$$

The most general one-step primitives allow the environment or an adaptive adversary to depend on the announced mixed action as a current control:

$$O_t \sim P_{\omega^\star,t}(\cdot \mid H_{t-1}, p_t, A_t), \qquad \ell_t(\omega^\star, H_{t-1}, p_t, A_t). \tag{1}$$

When the law and loss depend only on the realized action, as in ordinary stochastic bandits and Markov decision processes, we suppress the argument p_t and write $P_{\omega,t}(\cdot \mid H_{t-1}, A_t)$ and $\ell_t(\omega, H_{t-1}, A_t)$. The increment may be signed, provided the cumulative target used in a lower-bound theorem is integrable and has the stated lower bound, usually $L_\omega(H_T) \geq 0$. The Bellman clock t is abstract. In episodic reinforcement learning, it may represent either a macro-episode, with

the action given by a policy governing all within-episode stages, or a micro-time pair consisting of an outer episode index and an inner-stage index.

The frequentist objective studied in this paper is

$$\mathfrak{R}_T^\star(\Omega) := \inf_{\mathsf{Alg}\in\mathfrak{A}_T^{\mathrm{na}}} \sup_{\omega\in\Omega} \mathbb{E}_\omega^{\mathsf{Alg}} \sum_{t=1}^{T} \ell_t(\omega, H_{t-1}, p_t, A_t).$$

This explicit stepwise model is compatible with the interactive statistical decision making (ISDM) framework of Chen et al. (2024); its purpose here is to make the histories, controls, losses, and Bellman state used by the subsequent identities unambiguous.

Bellman state and Bellman sufficiency. A Bellman state is a measurable compression $S_t = \phi_t(H_{t-1})$ that determines the feasible actions, fixed-truth predictive law and loss, and the next-state update. It need not be the physical state or a posterior. The full history always qualifies; a useful representation is a smaller state on which the same controlled recursions close. When the environment reacts to the announced mixed action, that distribution remains a control argument rather than part of the past. Definition 2.2 states the precise requirements and the additional conditions needed for posterior-indexed information accounting.

Throughout, the history, action, observation, state, environment, and index spaces are standard Borel. All displayed kernels, losses, and state maps are measurable, and feasible-action correspondences have measurable graphs. Every displayed Bellman optimization is assumed to admit universally measurable ε-selectors; exact attainment is assumed only where it is invoked.

Reference dynamic program and minimax lower bound. Suppose the state is Bellman sufficient for the fixed-truth primitives and also makes the reference posterior b_s, posterior predictive law, and Bayes loss state-measurable. With update $S_{t+1} = \tau_t(S_t, p_t, A_t, O_t)$, or with p_t suppressed when irrelevant, the Bayes/reference dynamic program induced by a posterior b_s is

$$V_{T+1}(s) = 0, \qquad V_t(s) = \inf_{p\in\Delta(\mathcal{A}_t(s))} \left\{\ell(s,p) + \mathbb{E}_{a\sim p,\, o\sim P_{s,p,a}} V_{t+1}(\tau_t(s,p,a,o))\right\}, \tag{2}$$

where

$$\ell(s,p) = \mathbb{E}_{a\sim p} \int \ell_\omega(s,p,a) b_s(d\omega), \qquad P_{s,p,a} = \int P_{\omega,s,p,a} b_s(d\omega).$$

Equation (2) is not the definition of the frequentist problem; it is the Bellman recursion induced by a chosen reference representation. Minimax lower bounds are connected to such reference recursions through Yao's principle,

$$\mathfrak{R}_T^\star(\Omega) \geq \sup_{\mu\in\Delta(\Omega)} \inf_{\mathsf{Alg}\in\mathfrak{A}_T^{\mathrm{na}}} \mathbb{E}_{\Omega\sim\mu} \mathbb{E}_{\mathbb{P}_\Omega^{\mathsf{Alg}}} \sum_{t=1}^{T} \ell_t(\Omega, H_{t-1}, p_t, A_t).$$

Remark 2.1 (Adversarial and estimated losses)**.** *The same control convention accommodates a nonanticipating adversary that observes the announced mixed action, as well as state-measurable estimated-loss surrogates. The precise interpretation and the lower-bound qualification are given in Appendix B.*

2.2 Information index, conditional loss, and entropy accounting

The latent environment space is Ω, but the information charged by the upper and lower bounds may be a coarser coordinate. An information index is a measurable map

$$Y = \chi(\Omega) \in \mathcal{Y}.$$

The index is chosen by the analyst and should reflect the decision-relevant object: an optimal arm, an optimal policy, a value function, a finite active marginal, or, when no compression is justified, the full environment. The fixed-truth Bellman recursion still uses $\omega^\star$. The index enters through the logarithmic potential, the posterior/reference information accounting, and the lower ghost event.

The loss need not be a function of Y alone. For Bayes or Fano calculations we therefore use the conditional index loss, assuming the relevant regular conditional laws exist. In the general convention with a current mixed-action control,

$$\ell_t^\chi(y, h, p, a) := \mathbb{E}_\mu[\ell_t(\Omega, h, p, a) \mid \chi(\Omega) = y, H_{t-1} = h]\,. \tag{3}$$

When the one-step primitives do not depend on p, the argument is suppressed. When the retained state is an exact posterior-indexed lift in Definition 2.2, this conditional loss and the corresponding conditional predictive law are state-measurable and may be written as $\ell_{t,y}^\chi(s, p, a)$ and $P_{t,s,p,a}^y$. These conditional-index objects are not required for ordinary fixed-truth Bellman control; they are required for indexed posterior averaging and for the Bellman–Fano lower-bound method.

We use the cumulative and average losses, with the same suppression convention,

$$L_\omega(H_T) = \sum_{t=1}^T \ell_t(\omega, H_{t-1}, p_t, A_t), \qquad \bar{L}_\omega(H_T) = T^{-1} L_\omega(H_T),$$

and

$$L_\chi(y, H_T) = \sum_{t=1}^T \ell_t^\chi(y, H_{t-1}, p_t, A_t), \qquad \bar{L}_\chi(y, H_T) = T^{-1} L_\chi(y, H_T).$$

Under the Bayesian mixture generated by $\Omega \sim \mu$,

$$\mathbb{E}_{\Omega\sim\mu, H_T\sim\mathbb{P}_\Omega^{\mathsf{Alg}}} L_\Omega(H_T) = \mathbb{E}_{Y,H_T} L_\chi(Y, H_T). \tag{4}$$

For the upper identities, the increments may be any integrable costs for which the displayed expectations exist. For the quantile lower bounds, the object that must be nonnegative is the cumulative indexed loss, not necessarily each one-step increment. Throughout the clean lower-bound statements we assume $L_\chi(Y, H_T) \geq 0$ almost surely under the true mixture. If a problem has a known lower bound $L_\chi \geq -B$, the same statements apply to the shifted loss $L_\chi + B$, with the corresponding shift subtracted at the end.

The two logarithmic quantities compared later are the upper coordinate code length and the lower ghost entropy. If q_1 is the initial reference marginal on $\mathcal{Y}$, the fixed-truth upper telescope pays

$$L_0(\Omega_0; q_1, \chi) := \sup_{\omega\in\Omega_0} \log \frac{1}{q_1(\chi(\omega))}.$$

If the upper comparison is over a finite index set of size M and q_1 is uniform on that index set, then $L_0 = \log M$. At a regret radius r, a hard prior μ with $q_1 = \chi_\# \mu$ gives the algorithm-uniform Bellman–Fano ghost mass and entropy

$$p_{\mu,r}^\star = \sup_{\mathsf{Alg}\in\mathfrak{A}_T^{\mathrm{na}}} \mathbb{P}_{Y\sim q_1,\, H_T'\sim\mathbb{P}_\mu^{\mathsf{Alg}}}\{\bar{L}_\chi(Y, H_T') \leq r\}, \qquad E_{\mu,r}^\star := \log \frac{1}{p_{\mu,r}^\star}.$$

Thus the primitive entropy comparison is $L_0/E^\star_{\mu,r}$, not the size of the full model class. Section 5.4 gives the main synthesis; the formal comparison and its conversion into a regret gap are stated in Appendix A.

2.3 Bellman-sufficient representation

We now give the formal representation axioms used throughout the rest of the paper. The frequentist truth $\omega^\star$ remains fixed. Priors, posteriors, confidence objects, exponential-weights distributions, and algorithmic beliefs are reference objects: they may define algorithms or analytical bounds, but they are not substitutes for the fixed environment unless a theorem explicitly takes a Bayesian average.

Definition 2.2 (Indexed Bellman-sufficient representation)**.** *Fix an index map $\chi : \Omega \to \mathcal{Y}$. A retained state process*

$$S_t = \phi_t(H_{t-1})$$

is a fixed-truth Bellman-sufficient representation *if the following objects are determined by the current time-state pair (t, s).*

(i) Admissible actions. *There is a state-measurable action set $\mathcal{A}_t(s)$ such that $\mathcal{A}_t(h) = \mathcal{A}_t(s)$ whenever $S_t(h) = s$.*

(ii) Predictive sufficiency. *For every environment ω, mixed action $p \in \Delta(\mathcal{A}_t(s))$, and realized action $a \in \mathcal{A}_t(s)$, there is a kernel $P_{\omega,t,s,p,a}$ such that, for every history h with $S_t(h) = s$,*

$$P_{\omega,t}(\cdot \mid h, p, a) = P_{\omega,t,s,p,a}.$$

When the environment does not react to the announced mixed action, the argument p is suppressed.

(iii) Loss sufficiency. *There is an integrable state-measurable cost function $\ell_{\omega,t}(s, p, a) \in \mathbb{R}$ such that*

$$\ell_t(\omega, h, p, a) = \ell_{\omega,t}(s, p, a) \qquad \text{whenever } S_t(h) = s.$$

The lower-bound quantile theorem does not require every increment to be nonnegative; it requires the cumulative indexed loss to be lower bounded, and in the displayed lower bounds we assume $L_\chi(Y, H_T) \geq 0$ almost surely.

(iv) Update sufficiency. *There is an update map or controlled kernel τ_t such that the next retained state is generated from the retained state and the current interaction. In deterministic-update notation,*

$$S_{t+1} = \tau_t(S_t, p_t, A_t, O_t),$$

with p_t suppressed when irrelevant. Random next contexts, physical next states, stage markers, and public randomization are part of the observation packet O_t.

When these clauses hold, Bellman recursion is closed under each fixed truth. A compressed state is therefore a dynamic sufficient statistic for the Bellman primitives: feasible actions, prediction, loss evaluation, and future-value evaluation.

If, in addition, a prior or reference law μ is fixed and one wants exact state-based posterior-reference information accounting, the full-history posterior objects must descend to the state. On the full history define

$$b_{t,h} := \mathcal{L}_\mu(\Omega \mid H_{t-1} = h), \qquad q_{t,h} := \chi_\# b_{t,h},$$

and, for $q_{t,h}$-almost every y,

$$b^y_{t,h} := \mathcal{L}_\mu(\Omega \mid H_{t-1} = h, \chi(\Omega) = y).$$

The full-history index-conditional predictive law and loss are

$$P^y_{t,h,p,a} := \int P_{\omega,t}(\cdot \mid h, p, a)\, b^y_{t,h}(d\omega),$$

$$\ell^\chi_{t,h,p,a}(y) := \int \ell_t(\omega, h, p, a)\, b^y_{t,h}(d\omega).$$

An exact posterior-indexed lift on the state *is the additional fiber-invariance requirement that, whenever $S_t(h) = s$, there exist state-measurable objects satisfying*

$$q_{t,h} = q_{t,s}, \tag{5}$$

and, for $q_{t,s}$-almost every y,

$$P^y_{t,h,p,a} = P^y_{t,s,p,a}, \qquad \ell^\chi_{t,h,p,a}(y) = \ell^\chi_{t,y}(s, p, a), \tag{6}$$

together with a state-measurable reference update for $q_{t,s}$. These posterior and index-conditional clauses are not needed for the fixed-truth Bellman recursion itself. They are needed when the posterior-averaged chain rule, the Bayes information-potential program, or the Bellman–Fano lower bound is run on the compressed state.

The notation will usually suppress the explicit time subscript when the stage is part of the state, and it will suppress the mixed-action argument when the environment does not react to the announced distribution. Thus $P_{\omega,s,a}$, $\ell_\omega(s,a)$, and $\mathcal{A}(s)$ mean the appropriate special cases of $P_{\omega,t,s,p,a}$, $\ell_{\omega,t}(s,p,a)$, and $\mathcal{A}_t(s)$ at the current Bellman time. This convention is harmless only after the state and current control include the variables that make actions, losses, observation laws, and updates state-measurable.

Bellman sufficiency versus classical sufficiency and Bellman rank. Three different notions should be distinguished. Classical model sufficiency retains the full posterior of Ω and therefore every fixed index marginal. Index-marginal sufficiency retains only the posterior of $Y = \chi(\Omega)$. Bellman sufficiency instead requires the controlled actions, predictive laws, losses, and updates to close on the state. Bellman closure alone need not preserve a posterior, while an index-sufficient state need not determine the dynamics. The exact posterior-indexed lift in (5)–(6) combines the pieces needed by the indexed recursion without requiring sufficiency for the full environment.

The formalism is aligned with the representation-level philosophy behind low Bellman rank (Jiang et al., 2017), but it is deliberately stricter. Bellman rank is a factorization of expected Bellman errors relative to a function class, roll-in distribution, and Bellman-error evaluation family. It is not, by itself, a state: it does not determine the realized next observation, the fixed-truth loss, the posterior or reference update, or the conditional laws for an index. It becomes an indexed Bellman-sufficient representation only after those missing objects are supplied and shown to close the recursions above.

Appendix B collects basic examples involving sufficient statistics, posterior states, and finite marginals, and shows how the general framework encompasses the standard Decision Making with Structured Observations (DMSO) setting (Foster et al., 2021, 2023).

2.4 Why sufficient states and compression matter

The full history is always a valid Bellman state, so compression is representational rather than necessary for existence. Its value is to expose a smaller state on which both the controlled recursion and the chosen information accounting remain valid. Such a state need not be unique or minimal. Proposition 2.3 records the two basic facts: a valid compression preserves the Bellman operator and cannot increase information about a fixed index.

Proposition 2.3 (State compression and information monotonicity)**.** *Let $S_t = \phi_t(H_{t-1})$ be a fixed-truth Bellman-sufficient representation. Suppose a one-step Bellman cost c_ω and a continuation value F_{t+1} are state-measurable, so that*

$$c_\omega(h,p) = c_\omega(s,p), \qquad F_{t+1}(H_t) = f_{t+1}(S_{t+1}), \qquad s = S_t(h).$$

Then the full-history Bellman operator factors through S_t: for every h with $S_t(h) = s$,

$$\inf_{p\in\Delta(\mathcal{A}_t(h))} \sup_{\omega\in\mathcal{C}_t(h)} \{c_\omega(h,p) + \mathbb{E}_{\omega,h,p} F_{t+1}(H_t)\} = \inf_{p\in\Delta(\mathcal{A}_t(s))} \sup_{\omega\in\mathcal{C}_t(s)} \{c_\omega(s,p) + \mathbb{E}_{\omega,s,p} f_{t+1}(S_{t+1})\},$$

whenever the comparison set is also state-measurable. Moreover, for any reference law and any fixed index $Y = \chi(\Omega)$,

$$I(Y; S_t) \le I(Y; H_{t-1}).$$

When $I(Y; H_{t-1}) < \infty$, equality in the information display holds if and only if index-marginal sufficiency holds, $Y \perp H_{t-1} \mid S_t$.

Proof. See Appendix H.1.1.

Proposition 2.3 shows why compression is information-theoretically fundamental: it preserves the Bellman operator while never increasing the information cost associated with a fixed index. A strict reduction is useful only if both the upper and lower recursions remain valid on the compressed state.

3 Indexed information and exact AIR/MAIR identities

3.1 Posterior-reference histories and index information

The mutual-information identities in this subsection use the exact posterior-reference version of Definition 2.2. The fixed-truth log-gain regret identity in the next subsection should not be substituted into the posterior chain rule unless the two objects are induced by a coherent reference law.

For a prior μ and algorithm Alg, define the Bayesian posterior-reference trajectory law

$$\bar{\mathbb{P}}_\mu^{\mathsf{Alg}} := \int \mathbb{P}_\omega^{\mathsf{Alg}} \, \mu(d\omega).$$

A reference history is denoted

$$H'_T \sim \bar{\mathbb{P}}_\mu^{\mathsf{Alg}}, \qquad S'_t = \phi_t(H'_{t-1}), \qquad b'_t = \mathcal{L}_\mu(\Omega \mid H'_{t-1}),$$

and the algorithm's reference decision kernel is

$$p'_t = p_t(\cdot \mid H'_{t-1}).$$

Definition 3.1 (Indexed information gain)**.** *At a current time-state (t, s), let q_s be the posterior law of $Y = \chi(\Omega)$ and let $P^y_{s,p,a}$ and $P_{s,p,a}$ be the conditional and marginal predictive laws from Definition 2.2; we suppress t in these laws when the stage is part of s. For $p \in \Delta(\mathcal{A}_t(s))$, define*

$$\mathcal{I}_\chi(s, p) := \mathbb{E}_{a\sim p} \int D_{\mathrm{KL}}(P^y_{s,p,a} \| P_{s,p,a}) q_s(dy). \tag{7}$$

When the environment does not respond to the announced mixed action, the argument p is suppressed. When $Y = \Omega$, this is model-index information. When $Y = A^\star(\Omega)$, this is action-index information. In the action-index case, the belief is still a belief over environments; only the information target is the optimal action.

Define the cumulative and average indexed information of an algorithm by

$$C^{\mathsf{Alg}}_\chi(\mu) := \mathbb{E}_{H'_T\sim\bar{\mathbb{P}}^{\mathsf{Alg}}_\mu} \sum_{t=1}^{T} \mathcal{I}_\chi(S'_t, p'_t), \qquad \bar{C}^{\mathsf{Alg}}_\chi(\mu) := \frac{1}{T} C^{\mathsf{Alg}}_\chi(\mu). \tag{8}$$

Proposition 3.2 (Reference-history indexed chain rule)**.** *Assume that the retained state is an exact posterior-indexed lift in the sense of Definition 2.2, that the displayed KL terms are well defined and integrable, and that the initial packet O_0 is deterministic or independent of Y. For every prior μ, index map χ, and algorithm Alg,*

$$I_\mu(Y; H_T) = \mathbb{E}_{H'_T\sim\bar{\mathbb{P}}^{\mathsf{Alg}}_\mu} \sum_{t=1}^{T} \mathcal{I}_\chi(S'_t, p'_t) = T\bar{C}^{\mathsf{Alg}}_\chi(\mu). \tag{9}$$

If O_0 may carry information about Y, the right-hand side equals $I_\mu(Y; H_T \mid O_0)$ and the unconditional identity has the additional term $I_\mu(Y; O_0)$.

Proof. See Appendix H.2.1.

3.2 Fixed-truth indexed AIR bracket

The posterior-averaged information gain in (7) is the correct object for Bayesian chain rules and the Bellman–Fano lower-bound method. The upper Bellman program uses the corresponding fixed-truth coordinate identity. The coordinate identity is most cleanly stated in AIR form: a current reference marginal on the index is scored by the logarithmic posterior of the index after the current action and observation. The point-coordinate formulas below are literal for finite or countable indices with positive masses. For a dominated non-atomic index, $q(y)$ and $q^+_\nu(y \mid \cdot)$ denote Radon–Nikodym densities with respect to a fixed dominating measure, and the identities require positivity and integrability of the displayed log-density ratios. The later step that discards the terminal coordinate log loss is restricted to a discrete index.

Coefficient convention. Throughout the upper-bound identities and Bellman programs, the information multiplier is denoted by $\gamma > 0$: a one-step log gain G is charged as γG, and an initial coordinate code length is charged as $\gamma \log(1/q)$. Some AIR/MAIR/DEC papers use a temperature parameter η with multiplier $1/\eta$ (Xu and Zeevi, 2025; Liu et al., 2025, 2026); in the present notation this is simply $\gamma = 1/\eta$.

Fix a state s, an action distribution $p \in \Delta(\mathcal{A}_t(s))$, and a pair belief

$$\nu \in \Delta(\Omega \times \mathcal{Y})$$

over environments and index values. Let q_ν be the $\mathcal{Y}$-marginal of ν, and let q be a current reference marginal that is positive on every coordinate evaluated by the logarithmic score. The pair belief induces, for each realized action a, the predictive mixture

$$P_{\nu,s,p,a}(\cdot) := \int P_{\omega,t}(\cdot \mid s,p,a)\,\nu(d\omega, dy),$$

and the posterior index marginal

$$q_\nu^+(B \mid s,p,a,o) := \nu(Y \in B \mid s,p,a,O=o), \qquad B \subseteq \mathcal{Y}.$$

When the observation laws are dominated by a common measure and have densities $p_{\omega,t,s,p,a}(o)$, this update is explicitly

$$q_\nu^+(B \mid s,p,a,o) = \frac{\int_{\Omega\times B} p_{\omega,t,s,p,a}(o)\,\nu(d\omega,dy)}{\int_{\Omega\times\mathcal{Y}} p_{\omega,t,s,p,a}(o)\,\nu(d\omega,dy)}. \tag{10}$$

In ordinary stochastic bandits and MDPs the dependence on p is suppressed.

Let $\ell_{\omega,y}(s,p)$ denote the comparison loss associated with pair (ω,y); for a valid fixed truth one takes $y=\chi(\omega)$ and $\ell_{\omega,\chi(\omega)}=\ell_\omega$. Define the general indexed AIR functional

$$\mathfrak{A}_{q,\gamma}(s,p,\nu) := \mathbb{E}_{(\Omega,Y)\sim\nu}\ell_{\Omega,Y}(s,p) - \gamma\mathbb{E}_{(\Omega,Y)\sim\nu,\,a\sim p,\,O\sim P_{\Omega,t}(\cdot|s,p,a)} \log\frac{q_\nu^+(Y \mid s,p,a,O)}{q(Y)}. \tag{11}$$

Equivalently,

$$\mathfrak{A}_{q,\gamma}(s,p,\nu) = \mathbb{E}_\nu\ell_{\Omega,Y}(s,p) - \gamma\mathcal{I}_\nu(Y;O \mid s,p) - \gamma D_{\mathrm{KL}}(q_\nu\|q), \tag{12}$$

where

$$\mathcal{I}_\nu(Y;O \mid s,p) := \mathbb{E}_{a\sim p}\int D_{\mathrm{KL}}(P^y_{\nu,s,p,a}\|P_{\nu,s,p,a})q_\nu(dy)$$

is the one-step posterior-averaged information computed under the candidate pair belief. Thus q is the current reference index marginal, while q_ν^+ is constructed from the algorithmic pair belief ν and the current likelihood model. In an exact Bayesian specialization, ν is the current posterior over $(\Omega,\chi(\Omega))$ and $q=q_\nu$. In AIR/AMS/EBO, ν may be an optimized or robust candidate pair belief, and the KL term in (12) accounts for moving its index marginal away from the current reference q.

Lemma 3.3 (AIR gradient bracket)**.** *Assume a finite or countable index with positive reference masses and finite or dominated observation laws, or use the dominated-index density convention above. Assume also that the displayed Gateaux derivatives and log ratios exist and are integrable. Up to an additive constant on the probability simplex,*

$$\frac{\partial}{\partial\nu(\omega,y)}\mathfrak{A}_{q,\gamma}(s,p,\nu) = \ell_{\omega,y}(s,p) - \gamma\mathbb{E}_{a\sim p,\,O\sim P_{\omega,t}(\cdot|s,p,a)}\log\frac{q_\nu^+(y \mid s,p,a,O)}{q(y)}. \tag{13}$$

Consequently, for every fixed pair (ω,y),

$$\begin{aligned} B^{\mathrm{AIR}}_{q,\gamma}(s,p,\nu;\omega,y) &:= \mathfrak{A}_{q,\gamma}(s,p,\nu) + \left\langle \nabla_\nu\mathfrak{A}_{q,\gamma}(s,p,\nu), \delta_{(\omega,y)} - \nu\right\rangle \\ &= \ell_{\omega,y}(s,p) - \gamma\mathbb{E}_{a\sim p,\,O\sim P_{\omega,t}(\cdot|s,p,a)}\log\frac{q_\nu^+(y \mid s,p,a,O)}{q(y)}. \end{aligned} \tag{14}$$

Proof. See Appendix H.2.2.

3.3 Exact AIR identity and model-index specialization

We state simplified versions and alternative proofs of several AIR/MAIR identities. These are regret identities rather than only upper bounds. Xu and Zeevi (2025) developed them for arbitrary history-dependent model and decision kernels inside the conditional expectations; see especially the latter part of Section 5.2 and the concluding discussion. This history dependence allows complexity to accumulate adaptively over time and motivates the Bellman formulation used here.

Each identity requires the same compatibility condition: the next index marginal must be generated by the pair belief ν_t in the AIR functional. Under this update, Lemma 3.3 gives the one-step AIR bracket and the coordinate logarithm telescopes. Other reference updates may give useful log-potential bounds, but they require separate calibration or relaxation arguments unless they admit the same pair-belief representation.

Theorem 3.4 (Exact indexed AIR regret identity)**.** *Let an algorithm generate, at each reached state S_t, a current reference index marginal q_t that is positive on the coordinate being evaluated, a pair belief $\nu_t \in \Delta(\Omega \times \mathcal{Y})$, and a decision law $p_t \in \Delta(\mathcal{A}_t(S_t))$. After sampling $A_t \sim p_t$ and observing O_t, update the reference marginal by*

$$q_{t+1}(\cdot) = q^{+}_{\nu_t}(\cdot \mid S_t, p_t, A_t, O_t). \tag{15}$$

Fix a pair $(\omega^\star, y^\star)$ such that $q_t(y^\star) > 0$ and $q_{t+1}(y^\star) > 0$ almost surely along the analyzed trajectory, and assume that the displayed losses and log ratios are integrable. Then, for every $\gamma > 0$,

$$\mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \sum_{t=1}^{T} \ell_{\omega^\star, y^\star}(S_t, p_t) = \gamma \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \log \frac{q_{T+1}(y^\star)}{q_1(y^\star)} + \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \sum_{t=1}^{T} B^{\mathrm{AIR}}_{q_t,\gamma}(S_t, p_t, \nu_t; \omega^\star, y^\star). \tag{16}$$

For the original fixed-truth regret, take $y^\star = \chi(\omega^\star)$ and $\ell_{\omega^\star, y^\star} = \ell_{\omega^\star}$. The expectation is over action randomization, observations, and any exogenous algorithmic randomness.

Proof. See Appendix H.2.3.

The variational interpretation through posterior scoring and Danskin's theorem is given in Appendix C.1.

Identity versus lower bound. The identity is tightest when the final coordinate potential is retained. Rearranging (16) gives

$$\mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \sum_{t=1}^{T} \ell_{\omega^\star, y^\star}(S_t, p_t) = \gamma \log \frac{1}{q_1(y^\star)} - \gamma \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \log \frac{1}{q_{T+1}(y^\star)} + \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \sum_{t=1}^{T} B^{\mathrm{AIR}}_{q_t,\gamma}(S_t, p_t, \nu_t; \omega^\star, y^\star).$$

For a finite or countable index, dropping the nonpositive terminal term $-\gamma \mathbb{E} \log(1/q_{T+1}(y^\star))$ gives the usual upper value and discards the remaining coordinate code length. For a non-atomic index represented by densities, that term need not be nonpositive and must be retained or controlled separately. This observation does not itself give a minimax lower bound; the lower side still requires a Bellman–Fano or ghost-quantile value. It explains why the same logarithmic coordinate is the right quantity on both sides: the upper proof pays the initial code length only after discarding the final unresolved code length, while the lower proof bounds how much indexed entropy must remain under ghost histories.

Stationary-posterior specialization. In the exact posterior-indexed lift of Definition 2.2, the pair belief is the current posterior law of $(\Omega, \chi(\Omega))$ and q_t is its index marginal. Then $q^+_{\nu_t}$ is the exact posterior index update and the one-step fixed-truth log gain can also be written in predictive-law form. For $y = \chi(\omega)$,

$$J_\chi(s, p; \omega) := \mathbb{E}_{a\sim p,\, O\sim P_{\omega,t}(\cdot|s,p,a)} \log \frac{q^+_{\nu_s}(y \mid s, p, a, O)}{q_s(y)}. \tag{17}$$

Equivalently, when the relevant laws are dominated,

$$J_\chi(s, p; \omega) = \mathbb{E}_{a\sim p} \left[D_{\mathrm{KL}}(P_{\omega,s,p,a} \| P_{s,p,a}) - D_{\mathrm{KL}}(P_{\omega,s,p,a} \| P^{\chi(\omega)}_{s,p,a}) \right]. \tag{18}$$

Posterior averaging of J_χ over ω conditional on $Y = y$, and then over $y \sim q_s$, recovers $\mathcal{I}_\chi(s, p)$. Thus J_χ is a fixed-truth coordinate log gain, whereas $\mathcal{I}_\chi$ is its posterior-averaged information counterpart.

For $\gamma > 0$, the exact-posterior fixed-truth indexed bracket is

$$B_{\chi,\gamma}(s, p; \omega) := \ell_\omega(s, p) - \gamma J_\chi(s, p; \omega). \tag{19}$$

It is the specialization of (14) obtained by taking the pair belief to be the current posterior over $(\Omega, \chi(\Omega))$ and the reference marginal to be q_s.

Environment-index specialization. If $Y = \Omega$, the pair belief is concentrated on the graph $y = \omega$. Write the current environment belief as μ and the reference model marginal as ρ. Point masses below are interpreted literally on a finite or countable model class and as densities relative to a fixed dominating measure otherwise; all displayed density ratios must be positive at the evaluated truth and integrable. Then (11) becomes

$$\mathrm{MAIR}_{\rho,\gamma}(s, p, \mu) := \mathbb{E}_{\omega\sim\mu} \ell_\omega(s, p) - \gamma \mathbb{E}_{a\sim p} I_\mu(\Omega; O \mid s, p, a) - \gamma D_{\mathrm{KL}}(\mu \| \rho). \tag{20}$$

The fixed-truth bracket is

$$B^{\mathrm{MAIR}}_{\rho,\gamma}(s, p, \mu; \omega^\star) = \ell_{\omega^\star}(s, p) - \gamma \mathbb{E}_{a\sim p} D_{\mathrm{KL}}(P_{\omega^\star,s,p,a} \| P_{\mu,s,p,a}) - \gamma \log \frac{\mu(\omega^\star)}{\rho(\omega^\star)}. \tag{21}$$

If ρ_{t+1} is the Bayes posterior update of μ_t after (S_t, p_t, A_t, O_t), the required point masses or densities remain positive at $\omega^\star$, and the losses and log ratios are integrable, then

$$\mathbb{E}^{\mathsf{Alg}}_{\omega^\star} L_{\omega^\star}(H_T) = \gamma \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \log \frac{\rho_{T+1}(\omega^\star)}{\rho_1(\omega^\star)} + \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \sum_{t=1}^{T} B^{\mathrm{MAIR}}_{\rho_t,\gamma}(S_t, p_t, \mu_t; \omega^\star). \tag{22}$$

The stationary or current-posterior choice $\mu_t = \rho_t$ removes the explicit density-ratio correction in (21) and gives the KL-DEC offset. In this specialization the potential is the model-coordinate log posterior. In action-index AIR it is instead the log posterior of the optimal action or policy. When the decision problem requires only this index, the action-index formulation can charge the smaller decision-relevant information target.

Differentiating the MAIR objective yields an elegant derivation of three familiar offsets: (i) the KL–DEC offset when the comparison and reference beliefs coincide, (ii) robust AMS/EBO control at an interior maximizing belief, and (iii) a Hellinger offset under the square-root posterior construction. The precise statement appears in Lemma C.1 of Appendix C.1.

4 Upper bounds from logarithmic Bellman potentials

We first give the fixed-truth log-potential telescope and then state when it is an AIR/MAIR identity. The potential is the logarithmic score of the maintained index marginal. For an AIR/MAIR identity, the next marginal must be the posterior index marginal produced by the same pair belief ν over (Ω, Y).

Fix a truth $\omega^\star$ and write $y^\star = \chi(\omega^\star)$. At state s, the log potential uses the current reference marginal q_s on the index. A local algorithmic control u specifies an action law $p_u \in \Delta(\mathcal{A}_t(s))$ and a state-measurable reference-index update

$$q_u^+(\cdot \mid s, a, o) \in \Delta(\mathcal{Y}).$$

In the AIR/MAIR specialization, this update is generated by a pair belief $\nu_u \in \Delta(\Omega \times \mathcal{Y})$, namely

$$q_u^+(\cdot \mid s, a, o) = q_{\nu_u}^+(\cdot \mid s, p_u, a, o),$$

with q_ν^+ defined in (10). Exact Bayes and stationary-posterior MAIR are ν-generated updates. Confidence-to-belief and exponential-weights procedures may also define logarithmic reference updates, but then the theorem uses the corresponding state-measurable update directly; such a bound should not be called the AIR gradient identity unless an admissible pair-belief representation has been verified.

The fixed-truth reference log gain used by the Bellman program is

$$\mathsf{G}_\chi(s, u; \omega^\star) := \mathbb{E}_{a \sim p_u,\, O \sim P_{\omega^\star,t}(\cdot \mid s, p_u, a)} \log \frac{q_u^+(y^\star \mid s, a, O)}{q_s(y^\star)}. \tag{23}$$

When $u = (p, \nu)$ is represented by an AIR pair belief, we also write $\mathsf{G}_\chi(s, p, \nu; \omega^\star)$. In that case $q_u^+ = q_\nu^+$, and G_χ is the fixed-truth AIR posterior-coordinate log gain. When ν is the exact posterior pair law, this is $J_\chi(s, p; \omega^\star)$ from (17). Thus the log potential is always the coordinate log score of a specified reference index marginal, but the AIR/MAIR regret identity additionally requires that the next marginal be the ν-posterior coordinate associated with the same pair belief used in the AIR functional.

For any predictable continuation potential V_t on states and any coefficient $\gamma > 0$, define the specialized fixed-truth logarithmic potential

$$\Phi_t^{V,\gamma}(s; y^\star) := V_t(s) + \gamma \log \frac{1}{q_s(y^\star)}. \tag{24}$$

The corresponding log-potential Bellman bracket is

$$\mathfrak{B}_\chi^{V,\gamma}(s, u; \omega^\star) := \ell_{\omega^\star}(s, p_u) + \mathbb{E}_{\omega^\star}[V_{t+1}(S^+) \mid s, u] - V_t(s) - \gamma \mathsf{G}_\chi(s, u; \omega^\star). \tag{25}$$

Here $A \sim p_u$, $O \sim P_{\omega^\star,t}(\cdot \mid s, p_u, A)$, the reference part of the next state is updated by $q_u^+(\cdot \mid s, A, O)$, and any other state variables are updated by the rule specified by u. In the AIR specialization $u = (p, \nu)$, this bracket is written $\mathfrak{B}_\chi^{V,\gamma}(s, p, \nu; \omega^\star)$. The identity behind the upper section is the telescope

$$\begin{aligned} \mathbb{E}_{\omega^\star} L_{\omega^\star}(H_T) = V_1(s_1) - \mathbb{E}_{\omega^\star} V_{T+1}(S_{T+1}) + \gamma \log \frac{1}{q_1(y^\star)} - \gamma \mathbb{E}_{\omega^\star} \log \frac{1}{q_{T+1}(y^\star)} \\ + \mathbb{E}_{\omega^\star} \sum_{t=1}^{T} \mathfrak{B}_\chi^{V,\gamma}(S_t, u_t; \omega^\star). \end{aligned} \tag{26}$$

Retaining the terminal term gives the sharp identity for the analyzed algorithm. The discrete-versus-density sign convention, and the exact loss from dropping the terminal coordinate code length, are discussed after Theorem 3.4. This remains an upper identity; a Bellman–Fano minimax lower bound is a separate argument.

Thus an upper bound is a Bellman supersolution for this log-potential bracket. The exact indexed AIR Bellman program optimizes dynamically over the state, action law, and any algorithmically chosen reference-update rule. If that update rule is represented by a pair belief, the pair belief is an algorithmic prediction variable. By contrast, AMS/EBO places candidate beliefs under a robust maximization and uses a saddle or first-order condition to control the same fixed-truth AIR bracket. UCB controls the bracket by calibration and optimism, while E2D drops or bounds the continuation to obtain a one-step offset.

Theorem 4.1 (Generic fixed-truth indexed AIR/MAIR upper bound)**.** *Fix $\omega^\star$ and $y^\star = \chi(\omega^\star)$. Assume $q_1(y^\star) > 0$ and that the reference update remains positive on $y^\star$ along the analyzed trajectory. Assume also that the point-coordinate or dominated-density convention applies and that the losses, log ratios, continuation potentials, and displayed error sum are integrable. If an algorithm chooses a decision law p_t and a state-measurable prediction belief ν_t so that the reference update is $q_{t+1} = q^+_{\nu_t}(\cdot \mid S_t, p_t, A_t, O_t)$ and, for every reached state,*

$$\mathfrak{B}^{V,\gamma}_\chi(S_t, p_t, \nu_t; \omega^\star) \le \varepsilon_t,$$

then

$$\mathbb{E}^{\mathsf{Alg}}_{\omega^\star} L_{\omega^\star}(H_T) \le V_1(s_1) + \gamma \log \frac{1}{q_1(y^\star)} - \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} V_{T+1}(S_{T+1}) - \gamma \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \log \frac{1}{q_{T+1}(y^\star)} + \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \sum_{t=1}^{T} \varepsilon_t. \quad (27)$$

Proof. See Appendix H.2.5.

It is useful to keep one concrete state specialization in mind. A fixed estimation procedure maps histories to a reference update. In AIR form the update is represented by a pair belief ν_t over (Ω, Y); in an exact model-posterior specialization it is a belief over environments with $Y = \chi(\Omega)$. The retained state includes the current reference index marginal

$$q_t \in \Delta(\mathcal{Y}), \quad (28)$$

and, when the update is posterior based, the next marginal is $q_{t+1} = q^+_{\nu_t}(\cdot \mid S_t, p_t, A_t, O_t)$. In a well-specified Bayes analysis, ν_t is the exact posterior over $(\Omega, \chi(\Omega))$. In a frequentist analysis, the state may instead contain a calibrated algorithmic belief, confidence object, or exponential-weights density. Such an object is valid in the AIR identity only when it induces an admissible pair belief and posterior-coordinate update; otherwise it gives a separate log-score bound whose validity must be proved through calibration or domination. Posterior averaging of the coordinate potential gives

$$\mathbb{E}_{Y \sim \nu_\chi} \gamma \log \frac{1}{q_s(Y)} = \gamma H(\nu_\chi) + \gamma D_{\mathrm{KL}}(\nu_\chi \| q_s), \quad (29)$$

and for a finite index set the same logarithmic scoring rule has the KL-dual log-partition form

$$\Psi_\gamma(q, z) := \gamma \log \sum_{y \in \mathcal{Y}} q(y) e^{z_y/\gamma} = \sup_{r \in \Delta(\mathcal{Y})} \{\langle r, z\rangle - \gamma D_{\mathrm{KL}}(r \| q)\}. \quad (30)$$

The coordinate log loss, its posterior average, and the dual log partition are three representations of the same KL/logarithmic scoring rule. They are not interchangeable in a proof: fixed-truth

guarantees use (24), posterior-averaged Bayes identities use (29), and AMS/EBO relaxations use (30).

In the model-index specialization, write the posterior/reference belief as μ_t. If the maintained reference update is exact Bayes, then $\mathsf{G}_\chi = J_\chi$ and the fixed-truth MAIR information increment is

$$J_t^{\mathrm{MAIR}}(p;\omega^\star) := \mathbb{E}_{a\sim p} D_{\mathrm{KL}}(P_{\omega^\star,t}(\cdot \mid S_t,a) \,\|\, P_{\mu_t,t}(\cdot \mid S_t,a)) \,. \tag{31}$$

The action-index AIR version is identical with the index $Y = A^\star$, the reference marginal q_t, and the AIR posterior coordinate $q^+(a^\star \mid a,o)$ in place of the model posterior.

4.1 Information-potential Bellman programming

The frequentist upper object used here is an indexed AIR log-penalized Bellman program on calibrated state/index representations. It is an exact dynamic program for the specified state, comparison sets, information potential, and control family, but it is not asserted to equal the unrestricted minimax value. Let $\mathfrak{C}_t(s)$ be the set of environments that the algorithmic state regards as admissible at state s. Let $\mathfrak{U}_t(s)$ be the local set of admissible algorithmic controls. A control $u \in \mathfrak{U}_t(s)$ specifies an action law $p_u \in \Delta(\mathcal{A}_t(s))$ and a reference-update rule. When this update is represented by a pair belief, write that belief as ν_u and set $q^+ = q^+_{\nu_u}$. If the pair belief is fixed by the state, then u is simply the action law together with this fixed update. The robust AMS/EBO belief variable is not this algorithmic control; it is introduced later under a supremum to control the bracket. For $\gamma > 0$, define

$$W^\gamma_{T+1}(s) = 0, \qquad W^\gamma_t(s) := \inf_{u\in\mathfrak{U}_t(s)} \sup_{\omega\in\mathfrak{C}_t(s)} \left\{ \ell_\omega(s,p_u) + \mathbb{E}_\omega[W^\gamma_{t+1}(S^+) \mid s,u] - \gamma \mathsf{G}_\chi(s,u;\omega) \right\}. \tag{32}$$

A measurable selector attaining the infimum in (32) is the information-penalized Bellman dynamic-programming algorithm. In the common AIR specialization, $u = (p,\nu)$ and

$$\mathsf{G}_\chi(s,u;\omega) = \mathsf{G}_\chi(s,p,\nu;\omega).$$

In exact Bayes, ν is fixed by the posterior state; in a non-AIR log-score bound, u specifies the state-measurable reference update directly. This construction builds on the principle of admissible relaxations (Rakhlin et al., 2012), which organize learning through dynamic programming over loss states, and is particularly close to the partial-information relaxation viewpoint of Rakhlin and Sridharan (2016), where estimator-like state variables and potential terms quantify uncertainty. It can also be viewed as a sequential analogue of classical minimum-complexity, information-risk, and Gibbs-type estimation in fixed experiments (Barron and Cover, 1991; Yang and Barron, 1999; Zhang, 2006). The key new feature is that a generic logarithmic mass penalty is propagated through a controlled Bellman recursion, rather than applied only once to a terminal estimator, and the recursion is carried by a Bellman-sufficient state together with an explicit information index.

Theorem 4.2 (Frequentist log-penalized Bellman upper bound)**.** *Fix $\Omega_0 \subseteq \Omega$ and $\gamma > 0$. Suppose the reference update and the comparison sets $\mathfrak{C}_t$ are surely calibrated for Ω_0: for every $\omega^\star \in \Omega_0$, one has $\omega^\star \in \mathfrak{C}_t(S_t)$ almost surely whenever round t is reached. Suppose the algorithm chooses a local control $u_t \in \mathfrak{U}_t(S_t)$, writes $p_t = p_{u_t}$ for its action law, updates the index marginal by the reference-update rule contained in u_t, and satisfies*

$$\sup_{\omega\in\mathfrak{C}_t(S_t)} \left\{ \ell_\omega(S_t,p_{u_t}) + \mathbb{E}_\omega[W^\gamma_{t+1}(S^+) \mid S_t,u_t] - \gamma \mathsf{G}_\chi(S_t,u_t;\omega) \right\} \le W^\gamma_t(S_t) + \varepsilon_t.$$

Then, for every fixed truth $\omega^\star \in \Omega_0$ with $y^\star = \chi(\omega^\star)$ and $q_1(y^\star) > 0$, provided the index is finite or countable, the reference update remains positive on $y^\star$ along the trajectory, and the losses, coordinate log ratios, continuation values, and displayed error sum are integrable,

$$\mathbb{E}^{\mathsf{Alg}}_{\omega^\star} L_{\omega^\star}(H_T) \le W_1^\gamma(s_1) + \gamma \log \frac{1}{q_1(y^\star)} + \mathbb{E}^{\mathsf{Alg}}_{\omega^\star} \sum_{t=1}^{T} \varepsilon_t. \tag{33}$$

Proof. See Appendix H.2.6.

Remark 4.3 (High-probability calibration)**.** *The preceding theorem deliberately assumes sure state-wise calibration. A future-dependent event $\mathcal{E}$ cannot simply be inserted into the conditional Bellman brackets and charged afterward by $\mathbb{E}[L^+\mathbf{1}\{\mathcal{E}^c\}]$: conditioning a continuation value and then restricting to $\mathcal{E}$ does not preserve the displayed Bellman inequality. A high-probability version requires a separate stopped-process or supermartingale argument, or explicit predictable failure brackets. In the kernel-UCB result below we therefore prove the deterministic regret inequality on its confidence event directly and only then take expectations.*

The posterior-averaged Bayesian recursion is a specialization rather than the primitive fixed-truth statement. Write $H_\chi(s) := H(q_s)$ for the entropy of the posterior index marginal. If the robust comparison set is replaced by the current posterior law and one averages the coordinate identity over $Y \sim q_s$, then

$$\mathbb{E}[H_\chi(S_{t+1}) \mid s, p] - H_\chi(s) = -\mathcal{I}_\chi(s, p).$$

For a fixed multiplier $\gamma > 0$, define the Bayes information-regularized value

$$U^\gamma_{T+1} \equiv 0, \qquad U^\gamma_t(s) := \inf_{p \in \Delta(\mathcal{A}_t(s))} \left\{ \ell(s,p) + \mathbb{E}[U^\gamma_{t+1}(S_{t+1}) \mid s, p] - \gamma \mathcal{I}_\chi(s,p) \right\}. \tag{34}$$

Equivalently,

$$\Phi^\gamma_t(s) := U^\gamma_t(s) + \gamma H_\chi(s) \tag{35}$$

solves the admissible-relaxation recursion

$$\Phi^\gamma_t(s) = \inf_{p \in \Delta(\mathcal{A}_t(s))} \left\{ \ell(s,p) + \mathbb{E}[\Phi^\gamma_{t+1}(S_{t+1}) \mid s, p] \right\}, \qquad \Phi^\gamma_{T+1}(s) = \gamma H_\chi(s). \tag{36}$$

Theorem 4.4 (Posterior-averaged information-potential upper bound)**.** *Assume the indexed Bellman representation is exact under the Bayesian mixture law, Y is finite or countable, and $H_\mu(Y) < \infty$. If a policy chooses p_t satisfying*

$$\ell(S_t, p_t) + \mathbb{E}[U^\gamma_{t+1}(S_{t+1}) \mid S_t, p_t] - \gamma \mathcal{I}_\chi(S_t, p_t) \le U^\gamma_t(S_t) + \varepsilon_t,$$

then

$$\mathbb{E} L_\Omega(H_T) \le U^\gamma_1(s_1) + \gamma I_\mu(Y; H_T) + \mathbb{E} \sum_{t=1}^{T} \varepsilon_t \le U^\gamma_1(s_1) + \gamma H_\mu(Y) + \mathbb{E} \sum_{t=1}^{T} \varepsilon_t. \tag{37}$$

This is the posterior average of the fixed-truth coordinate telescope, not a pointwise replacement for Theorem 4.2.

Proof. See Appendix H.3.1.

Section 5.4 summarizes the minimax information-risk sandwich, whose formal statement appears in Appendix A. The fixed-truth coordinate telescope and the exact log-penalized Bellman program above form its upper side. The next subsection gives a short map of the main tractable controls; Appendix D gives the details.

4.2 UCB, E2D, and AMS/EBO as tractable controls

The log-penalized Bellman program is an information-theoretic object; its statewise optimization need not be computationally easy. Three familiar algorithmic families control it in different ways. UCB combines a calibrated confidence width with optimism and then sums the realized, unmultiplied uncertainty by an elliptical-potential argument. E2D replaces the continuation term by a localized one-step separation penalty. AMS optimizes a robust candidate belief and action distribution, while EBO gives the corresponding KL-dual optimization principle. These are complementary controls of the same logarithmic Bellman bracket, not identical algorithms or a single universal coefficient.

The precise bracket inequalities, the distinction between algorithmic update beliefs and robust comparison beliefs, and the single-round DEC specializations are collected in Appendices D and D.4. The kernel application below uses the localized robust AIR/AMS control only after a finite marginal has been fixed; it does not assume that any of these methods discovers the localization scale automatically.

5 Lower bounds: reference histories and quantile indices

5.1 Ghost probability and true good probability

Fix an average-regret threshold $r \geq 0$. The posterior-reference ghost-good probability is

$$p_r^{\mathsf{Alg}}(\mu, \chi) := \mathbb{P}_{Y \sim \mu_\chi,\, H'_T \sim \bar{\mathbb{P}}_\mu^{\mathsf{Alg}}} \left(\bar{L}_\chi(Y, H'_T) \leq r \right), \tag{38}$$

where Y and H'_T are independent under the ghost law and μ_χ is the marginal law of Y. The true good probability is

$$q_r^{\mathsf{Alg}}(\mu, \chi) := \mathbb{P}_{Y, H_T} \left(\bar{L}_\chi(Y, H_T) \leq r \right), \tag{39}$$

where (Y, H_T) are generated by $\Omega \sim \mu$ and $H_T \sim \mathbb{P}_\Omega^{\mathsf{Alg}}$.

The ghost probability is algorithm-dependent. That dependence is intentional: it preserves the reference trajectory geometry instead of replacing it with a worst-case static small ball.

5.2 Reference-history quantile theorem

Let

$$\mathrm{kl}(q, p) = q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p}$$

be binary KL divergence. For $p \in [0, 1]$ and $c \geq 0$, define

$$\psi(p, c) := \sup\{q \in [0, 1] : \mathrm{kl}(q, p) \leq c\}.$$

Theorem 5.1 (Reference-history quantile indexed-information lower bound)**.** *For every prior μ, index map χ, threshold $r \geq 0$, and algorithm Alg, assume that the posterior-reference process satisfies the exact posterior-indexed lift in the sense of Definition 2.2, that O_0 is deterministic or independent of Y, and that the cumulative indexed loss satisfies $L_\chi(Y, H_T) \geq 0$ almost surely under the true mixture law. Then*

$$\mathrm{kl}\big(q_r^{\mathsf{Alg}}(\mu, \chi), p_r^{\mathsf{Alg}}(\mu, \chi)\big) \leq C_\chi^{\mathsf{Alg}}(\mu) = T \bar{C}_\chi^{\mathsf{Alg}}(\mu). \tag{40}$$

Consequently,

$$\mathbb{E}_{\Omega \sim \mu} \mathbb{E}_{\mathbb{P}_\Omega^{\mathsf{Alg}}} L_\Omega(H_T) \geq Tr \left[1 - \psi\big(p_r^{\mathsf{Alg}}(\mu, \chi), T \bar{C}_\chi^{\mathsf{Alg}}(\mu)\big) \right]. \tag{41}$$

In particular, if $p_r^{\mathsf{Alg}}(\mu,\chi) \le 1/2$ *and*

$$T\bar{C}_\chi^{\mathsf{Alg}}(\mu) \le \mathrm{kl}\left(\frac{1}{2}, p_r^{\mathsf{Alg}}(\mu,\chi)\right), \tag{42}$$

then

$$\mathbb{E}_{\Omega\sim\mu}\mathbb{E}_{\mathbb{P}_\Omega^{\mathsf{Alg}}}\bar{L}_\Omega(H_T) \ge \frac{r}{2}, \qquad \mathbb{E}_{\Omega\sim\mu}\mathbb{E}_{\mathbb{P}_\Omega^{\mathsf{Alg}}}L_\Omega(H_T) \ge \frac{Tr}{2}. \tag{43}$$

A convenient sufficient condition is

$$T\bar{C}_\chi^{\mathsf{Alg}}(\mu) + \log 2 \le \frac{1}{2}\log\frac{1}{p_r^{\mathsf{Alg}}(\mu,\chi)}. \tag{44}$$

If O_0 *may be informative about* Y*, the same conclusions hold after replacing every occurrence of* $T\bar{C}_\chi^{\mathsf{Alg}}(\mu)$ *in the information budget by*

$$I_\mu(Y;O_0) + T\bar{C}_\chi^{\mathsf{Alg}}(\mu).$$

Proof. See Appendix H.3.4.

Appendix C.2 details the adaption of interactive Fano and the posterior-reference chain rule.

5.3 Bellman–Fano lower-bound method

Theorem 5.1 is algorithm-specific. We now give an algorithm-uniform lower-bound method with two components: the first bounds the information of any algorithm, and the second bounds the ghost probability of success of any algorithm.

Exact information Bellman value and supersolutions. The intrinsic algorithm-uniform information quantity is an exact Bellman value. For a function F on states, define

$$(\mathsf{C}_t F)(s) := \sup_{p\in\Delta(\mathcal{A}_t(s))} \left\{\mathcal{I}_\chi(s,p) + \mathbb{E}_{a\sim p, o\sim P_{s,p,a}} F(\tau(s,p,a,o))\right\}.$$

Set

$$C_{T+1}^\star(s) = 0, \qquad C_t^\star(s) = (\mathsf{C}_t C_{t+1}^\star)(s).$$

Then $C_1^\star(s_1)$ is the exact posterior-reference indexed-information capacity on the retained state. A sequence $\overline{C}_t$ is a valid information supersolution if

$$\overline{C}_{T+1} \ge 0, \qquad \overline{C}_t(s) \ge (\mathsf{C}_t\overline{C}_{t+1})(s) \quad \forall t, s.$$

For every algorithm starting from s_1,

$$C_\chi^{\mathsf{Alg}}(\mu) \le C_1^\star(s_1) \le \overline{C}_1(s_1). \tag{45}$$

Thus the exact Bellman recursion is the tight intrinsic object; supersolutions are used only as computable or analytically convenient upper bounds.

Exact ghost-good Bellman value and supersolutions. Let $m \in \Delta(\mathcal{Y})$ be the target-index distribution used to evaluate success, let s be the posterior-reference state, and let $g : \mathcal{Y} \to \mathbb{R}$ be an accumulated average-loss profile. For threshold r, define the exact ghost-good Bellman value by

$$\Gamma^{r,\star}_{T+1}(m, s, g) = m\{y : g(y) \le r\},$$

$$\Gamma^{r,\star}_t(m, s, g) = \sup_{p \in \Delta(\mathcal{A}_t(s))} \mathbb{E}_{a\sim p, o\sim P_{s,p,a}} \Gamma^{r,\star}_{t+1}\left(m, \tau(s, p, a, o), g + \frac{1}{T}\ell^{\chi}_{\cdot}(s, p, a)\right),$$

where $\ell^{\chi}_{\cdot}(s, p, a)$ is the function $y \mapsto \ell^{\chi}_y(s, p, a)$. A function sequence $\overline{\Gamma}^r_t$ is a valid ghost-mass supersolution if it dominates the terminal condition and the same Bellman right-hand side with $\overline{\Gamma}^r_{t+1}$ in place of $\Gamma^{r,\star}_{t+1}$. We henceforth take ghost-mass supersolutions to be $[0, 1]$-valued. Any nonnegative supersolution may be clipped at one, since the ghost Bellman operator is monotone and maps $[0, 1]$-valued functions into $[0, 1]$; this clipping preserves the supersolution inequality. Then, for every algorithm,

$$p^{\mathsf{Alg}}_r(\mu, \chi) \le \Gamma^{r,\star}_1(\mu_\chi, s_1, 0) \le \overline{\Gamma}^r_1(\mu_\chi, s_1, 0). \tag{46}$$

The exact ghost entropy and its computable lower bound are therefore

$$\mathcal{E}^r_{\star,1}(\mu, \chi) := -\log \Gamma^{r,\star}_1(\mu_\chi, s_1, 0), \qquad \underline{\mathcal{E}}^r_1(\mu, \chi) := -\log \overline{\Gamma}^r_1(\mu_\chi, s_1, 0). \tag{47}$$

The entropy lower bound is conservative: $\underline{\mathcal{E}}^r_1 \le \mathcal{E}^r_{\star,1}$.

Theorem 5.2 (Bellman–Fano quantile-index lower bound)**.** *Assume that the retained state is an exact posterior-indexed lift under μ, that the indexed information terms are well defined and integrable, and that a fixed jointly measurable version of $\ell^{\chi}_y(s, p, a)$ is used in the ghost product experiment. The theorem is stated conditionally on a fixed initial packet and resulting state s_1; an unconditional result for a random initial state follows by applying the full conditional bound and then averaging. Assume the cumulative indexed loss is nonnegative as in Theorem 5.1. Suppose $\overline{C}_t$ is an information supersolution and $\overline{\Gamma}^r_t$ is a $[0, 1]$-valued ghost-good-mass supersolution; the exact choices $\overline{C} = C^\star$ and $\overline{\Gamma} = \Gamma^{r,\star}$ give the tight intrinsic Bellman–Fano values. The statement is for the unrestricted nonanticipating class; for a Bellman-compatible restricted class, replace each local supremum in the exact recursions and in the supersolution recursions by the corresponding local feasible set. Then*

$$\inf_{\mathsf{Alg}\in\mathfrak{A}^{\mathrm{na}}_T} \mathbb{E}_{\Omega\sim\mu}\mathbb{E}_{\mathbb{P}^{\mathsf{Alg}}_\Omega} L_\Omega(H_T) \ge Tr\left[1 - \psi\big(e^{-\underline{\mathcal{E}}^r_1(\mu,\chi)}, \overline{C}_1(s_1)\big)\right]. \tag{48}$$

In particular, if

$$\overline{C}_1(s_1) + \log 2 \le \frac{1}{2}\underline{\mathcal{E}}^r_1(\mu, \chi), \tag{49}$$

then

$$\inf_{\mathsf{Alg}\in\mathfrak{A}^{\mathrm{na}}_T} \mathbb{E}_{\Omega\sim\mu}\mathbb{E}_{\mathbb{P}^{\mathsf{Alg}}_\Omega} L_\Omega(H_T) \ge \frac{Tr}{2}. \tag{50}$$

Consequently,

$$\mathfrak{R}^\star_T \ge \sup_{\mu,\chi,r}\left\{\frac{Tr}{2} : \overline{C}_1(s_1) + \log 2 \le \frac{1}{2}\underline{\mathcal{E}}^r_1(\mu, \chi)\right\}.$$

Proof. See Appendix H.3.5.

The critical average-regret radius is

$$r_\star(\mu, \chi) := \sup\left\{r : \overline{C}_1(s_1) + \log 2 \le \frac{1}{2}\underline{\mathcal{E}}^r_1(\mu, \chi)\right\}. \tag{51}$$

The lower bound is order $Tr_\star$. A two-point Le Cam proof corresponds to the special case in which $\underline{\mathcal{E}} = O(1)$. A Fano or local-prior-mass lower bound has $\underline{\mathcal{E}}$ of order dimension or effective dimension.

Relation to DEC. The Bellman formulation keeps hard-prior admissibility, evaluation of the upper recursion at the lower entropy scale, and growth of the upper value separate, while DEC derives upper and lower guarantees from a single-round coefficient. Appendix D.4 gives the formal comparison and discusses when a single-round reduction preserves the dynamic rate.

5.4 The information-complexity sandwich

The two constructions can now be compared in common units. For a comparison class Ω_0, let

$$L_0 = \sup_{\omega \in \Omega_0} \log \frac{1}{q_1(\chi(\omega))}$$

be the coordinate code length paid by the upper Bellman telescope. For a hard prior μ and average-loss radius r, let

$$E^\star_{\mu,r} = \log \frac{1}{p^\star_{\mu,r}}$$

be the algorithm-uniform entropy of reference histories that are already good for the sampled index. The upper construction gives $\mathfrak{R}^\star_T(\Omega_0) \le \mathsf{U}_T(L_0)$, whereas the Bellman–Fano construction gives $\mathfrak{R}^\star_T(\Omega_0) \ge Tr/2$ whenever the hard prior is admissible at radius r.

The rates therefore meet when the upper value closes at the same entropy scale, $\mathsf{U}_T(E^\star_{\mu,r}) \lesssim Tr$, and has regular growth between $E^\star_{\mu,r}$ and L_0. In that case the remaining gap is controlled by $L_0/E^\star_{\mu,r}$. Thus coordinate code length replaces the complexity paid by an upper algorithm, while ghost entropy replaces the packing or local prior mass in a fixed experiment. The exact admissibility condition, growth hypothesis, finite-index diagnostics, and regret-ratio statement are collected in Appendix A; see Theorem A.8.

6 Applications: kernel bandits and information-complexity sandwiches

This section contains complementary kinds of applications. Kernel bandits are the major algorithmic application: they illustrate the Bellman decomposition and its consequences for algorithm design through a full-RKHS separation between finite-marginal AIR Bellman programming and globally calibrated maximal-information GP–UCB, including the unit-ball exploration schedule of the original RKHS GP–UCB algorithm. We also show that the construction implies the suboptimality of LinUCB on a linear bandit with a heterogeneous action set, and compare this separation with results for spatially homogeneous Matérn kernels.

The Gaussian multi-armed bandit and hypercube linear bandit examples instantiate the lower information-risk construction of Section 5.4 and yield minimax rates matched by standard algorithms. The detailed analyses are deferred to Appendices F and G.

6.1 Kernel bandits: Bellman decompositions and algorithms

This subsection specializes the indexed Bellman language to kernel bandits on a finite active action set. The full RKHS truth may be infinite-dimensional, but the bandit experiment restricted to a finite active set is equivalent to a finite-dimensional linear experiment in the span of the corresponding kernel features. This is an equivalence of the observable reward model, not an assumption that the unknown function has finitely many parameters on the full domain. This marginal is enough to define the GP posterior update, the model-index MAIR information, and the action-index AIR law

of the optimal active action. We present the finite-marginal indexed Bellman program, prove its action-index AIR bound, and then compare it with GP–UCB. Additional kernel details, including the AMS/EBO specialization and the Matérn comparison, are given in Appendix E.

Problem and finite active marginal. Let $\mathcal{H}_k$ be a reproducing kernel Hilbert space (RKHS) on a possibly infinite domain $\mathcal{X}_{\text{all}}$, with kernel k satisfying $k(x,x) \le \kappa^2$. The learner is evaluated on a finite active set

$$\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathcal{X}_{\text{all}}.$$

The unknown reward function $f^\star \in \mathcal{H}_k$ satisfies $\|f^\star\|_{\mathcal{H}_k} \le B$, and at round t the learner chooses $X_t \in \mathcal{X}$ and observes

$$Y_t = f^\star(X_t) + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t \mid H_{t-1}, X_t] = 0, \tag{52}$$

where the noise is conditionally R-sub-Gaussian. For the Gaussian calculations below we use the algorithmic reference model with variance parameter $\lambda > 0$; the frequentist confidence event is obtained by the usual self-normalized argument.

For an action set $\mathcal{X}$ and reward function f, define cumulative regret by

$$\mathrm{Reg}_T(f; \mathsf{Alg}) := \sum_{t=1}^{T} \left\{ \max_{x \in \mathcal{X}} f(x) - f(X_t) \right\}. \tag{53}$$

We write $\mathrm{Reg}_T(f)$ when the algorithm is clear and $\mathrm{Reg}_T(\mathsf{Alg})$ when the environment is carried by the expectation.

The posterior and AIR/MAIR chain-rule identities in this subsection use the well-specified Gaussian observation model $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0,\lambda)$. The pathwise GP–UCB confidence argument also holds for conditionally sub-Gaussian noise, but under non-Gaussian noise a Gaussian pseudo-posterior is only an algorithmic state and does not satisfy the Bayesian chain rule without a separate transfer argument.

Let $F = (f(x_1), \dots, f(x_n)) \in \mathbb{R}^n$ be the active reward vector and let $K_{\mathcal{X}}$ be the $n \times n$ kernel matrix. The GP reference prior is

$$F \sim N(0, K_{\mathcal{X}}), \qquad Y_t \mid F, X_t = x_i \sim N(F_i, \lambda).$$

After history H_{t-1}, the algorithmic posterior on F is Gaussian,

$$F \mid H_{t-1} \sim N(m_t, \Sigma_t).$$

For $x_i \in \mathcal{X}$, write $m_t(x_i) = e_i^\top m_t$ and $s_t^2(x_i) = e_i^\top \Sigma_t e_i$. If $X_t = x_i$, the exact posterior update is

$$m_{t+1} = m_t + \frac{\Sigma_t e_i}{\lambda + s_t^2(x_i)} \{Y_t - m_t(x_i)\}, \tag{54}$$

$$\Sigma_{t+1} = \Sigma_t - \frac{\Sigma_t e_i e_i^\top \Sigma_t}{\lambda + s_t^2(x_i)}. \tag{55}$$

This is a finite marginal posterior update. It should not be read as an assumption that $f^\star$ has n parameters. The RKHS function may be infinite-dimensional; only the observations and decisions in this finite experiment depend on the vector F.

Model-index information and action-index information. For the model-index specialization, take the index to be the active reward vector $Y_{\mathrm{MAIR}} = F$. The one-step GP/MAIR information at state H_{t-1} is

$$\mathcal{I}_t^{\mathrm{GP}}(p) := \mathbb{E}_{x\sim p}\frac{1}{2}\log\left(1+\frac{s_t^2(x)}{\lambda}\right). \tag{56}$$

Along a realized action sequence,

$$C_T^{\mathrm{GP}} := \sum_{t=1}^{T}\mathcal{I}_t^{\mathrm{GP}}(\delta_{X_t}) = \frac{1}{2}\log\det\bigl(I+\lambda^{-1}K_{X_{1:T},X_{1:T}}\bigr), \tag{57}$$

where repeated actions are included in the Gram matrix. This is the realized information gain of the algorithmic trajectory. It can be much smaller than the maximal information gain that maximizes over all length-T designs.

For the action-index specialization, assume ties are broken by a fixed rule and set

$$A^\star(F) \in \arg\max_{x\in\mathcal{X}} F_x, \qquad q_t(a) := \mathbb{P}(A^\star = a \mid H_{t-1}).$$

Let ν_t^a be the exact conditional law of F given H_{t-1} and $A^\star = a$. It is generally a truncated Gaussian on the cone where a is optimal. Define the conditional predictive law

$$P_{t,a,x}(\cdot) := \int N(F_x,\lambda)(\cdot)\,\nu_t^a(dF),$$

for the observation at action x given the event $A^\star = a$. The marginal predictive is

$$P_{t,x} = \sum_{a\in\mathcal{X}} q_t(a)P_{t,a,x} = N(m_t(x), \lambda + s_t^2(x)).$$

The AIR information increment is

$$\mathcal{I}_t^{\mathrm{AIR}}(p) := \mathbb{E}_{x\sim p}\sum_{a\in\mathcal{X}} q_t(a)D_{\mathrm{KL}}(P_{t,a,x}\|P_{t,x}). \tag{58}$$

The conditional AIR regret is

$$\Delta_t^{\mathrm{AIR}}(p) := \sum_{a\in\mathcal{X}} q_t(a)\mathbb{E}_{F\sim\nu_t^a}\mathbb{E}_{x\sim p}[F_a - F_x]. \tag{59}$$

The exact indexed chain rule gives

$$\sum_{t=1}^{T}\mathbb{E}\mathcal{I}_t^{\mathrm{AIR}}(p_t) = I(A^\star; H_T) \le H(A^\star) \le \log n.$$

Thus AIR places the entropy telescope on the decision index rather than on the ambient RKHS. The price is that the coefficient converting action-index information into regret must be controlled by an algorithmic bracket.

Algorithm 1: finite-marginal indexed Bellman DP. On a finite active set, (m_t, Σ_t, q_t) is a Bellman state, where q_t is the maintained law of the optimal active action. Applying the action-index Bellman program to this state gives the finite-marginal policy analyzed below. The full recursion, its robust pair-belief interpretation, and the distinction between finite action and continuous model indices are given in Appendix E. In addition, Appendix E.5 gives a kernel-specific specialization of AMS/EBO as a supersolution to the exact Bellman program.

The next theorem gives a uniform frequentist rate for the action-index branch. For a finite retained action set $\mathcal{R}$, write

$$\mathrm{Reg}_T^{\mathrm{all}}(f) := \sum_{t=1}^{T} \left\{ \sup_{x \in \mathcal{X}_{\mathrm{all}}} f(x) - f(X_t) \right\}$$

for regret against the full decision domain, and define

$$\varepsilon_{\mathcal{R}} := \sup_{f \in \mathcal{F}} \left\{ \sup_{x \in \mathcal{X}_{\mathrm{all}}} f(x) - \max_{x \in \mathcal{R}} f(x) \right\}.$$

Theorem 6.1 (Finite-marginal action-index AIR Bellman bound)**.** *Let $|\mathcal{R}| = K \geq 2$, suppose the retained reward-vector class*

$$\Theta_{\mathcal{R}} := \{(f(x))_{x \in \mathcal{R}} : f \in \mathcal{F}\}$$

is compact and contained in $[-L, L]^K$, and let the observations have independent $N(0, \lambda)$ noise with $\lambda > 0$. For every $\gamma > 0$, the finite-marginal action-index AIR Bellman program admits a policy, confined to $\mathcal{R}$ and initialized with the uniform law on the attainable optimal-action indices, such that

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \, \mathrm{Reg}_T^{\mathrm{all}} \leq T\varepsilon_{\mathcal{R}} + \frac{(\lambda + L^2)KT}{2\gamma} + \gamma \log K. \tag{60}$$

Consequently, the choice

$$\gamma = \sqrt{\frac{(\lambda + L^2)KT}{2 \log K}}$$

gives

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \, \mathrm{Reg}_T^{\mathrm{all}} \leq T\varepsilon_{\mathcal{R}} + \sqrt{2(\lambda + L^2)KT \log K}. \tag{61}$$

The robust AIR/AMS saddle supplies the control for an explicit supersolution of the exact Bellman program and itself achieves the same bound. Under the minimax and KL-duality identification stated above, the corresponding EBO formulation has the same guarantee. More precisely, the robust action-index control is the finite-marginal specialization of the AIR/AMS saddle of Xu and Zeevi (2025); its EBO interpretation follows the optimization principle of Lattimore and György (2021), rather than identifying the two algorithms line by line.

Proof. See Appendix H.4.2.

The theorem is deliberately stated for the finite-valued optimal-action index. Since $A^\star$ is a function of F, data processing gives the analogous posterior-averaged one-step information estimate with the model index F. A uniform frequentist MAIR statement, however, must retain the terminal log-density term for this continuous index or replace point coordinates by neighborhoods, a finite cover, or a PAC–Bayes comparison. Finite marginality alone does not turn the model coordinate into a discrete code length.

Algorithm 2: GP–UCB. GP–UCB uses the Gaussian marginal posterior (m_t, Σ_t) on the finite active marginal. In this paragraph, e_x denotes the coordinate vector associated with $x \in \mathcal{X}$, and $s_t(x) = (e_x^\top \Sigma_t e_x)^{1/2}$ denotes the unmultiplied posterior standard deviation, i.e., the generic uncertainty scale $\sigma_t(x)$ from Lemma D.1. The confidence half-width is

$$w_t(x) := \beta_t s_t(x),$$

where β_t is the confidence multiplier. The algorithm chooses

$$X_t \in \arg\max_{x \in \mathcal{X}} \{m_t(x) + \beta_t s_t(x)\}. \tag{62}$$

The following event is the frequentist calibration:

$$\mathcal{E}_{\mathrm{GP}} := \{|f^\star(x) - m_t(x)| \le w_t(x) = \beta_t s_t(x) \quad \text{for all } t \le T, x \in \mathcal{X}\}. \tag{63}$$

For finite $\mathcal{X}$, standard RKHS self-normalized concentration gives $\mathbb{P}(\mathcal{E}_{\mathrm{GP}}) \ge 1 - \delta$ for the usual choice of β_t. For example, up to universal constants one may take

$$\beta_t \asymp B + \frac{R}{\sqrt{\lambda}} \sqrt{\Gamma^{\mathrm{GP}}_{t-1} + \log(1/\delta)}, \qquad \Gamma^{\mathrm{GP}}_m := \sup_{x_{1:m} \in \mathcal{X}^m} \frac{1}{2} \log\det\left(I + \lambda^{-1} K_{x_{1:m}, x_{1:m}}\right),$$

with the precise form depending on the normalization of k and the ridge parameter. Here $K_{x_{1:m}, x_{1:m}}$ is the Gram matrix of the possibly repeated design sequence $(x_1, \dots, x_m)$. The quantity Γ^{GP}_m is a worst-case calibration device for the confidence event; it is different from the realized information gain C^{GP}_T paid in the regret bound. The next theorem isolates the deterministic part of the GP–UCB proof.

Theorem 6.2 (GP–UCB realized-information bound)**.** *Assume that $(\beta_t)_{t \le T}$ is a deterministic nondecreasing sequence. On the event $\mathcal{E}_{\mathrm{GP}}$, the GP–UCB rule (62) satisfies*

$$\mathrm{Reg}_T(f^\star) \le 2\beta_T \sum_{t=1}^{T} s_t(X_t) \le 2\beta_T \sqrt{T\, c_{\kappa,\lambda}\, C^{\mathrm{GP}}_T}, \tag{64}$$

where

$$c_{\kappa,\lambda} := \frac{2\kappa^2}{\log(1 + \kappa^2/\lambda)}$$

and C^{GP}_T is the realized information gain in (57). Consequently, if rewards are bounded in $[0,1]$ and $\mathbb{P}(\mathcal{E}_{\mathrm{GP}}) \ge 1 - \delta$, then

$$\mathbb{E}\,\mathrm{Reg}_T(f^\star) \le 2\beta_T \sqrt{T\, c_{\kappa,\lambda}\, \mathbb{E} C^{\mathrm{GP}}_T} + T\delta.$$

Proof. See Appendix H.4.1.

Calibration and realized information. The theorem is Lemma D.1 with the generic uncertainty scale $\sigma_t(X_t) = s_t(X_t)$ and the realized posterior-reference log-determinant telescope

$$\sum_t \frac{1}{2} \log\left(1 + s_t^2(X_t)/\lambda\right) = C^{\mathrm{GP}}_T.$$

The confidence half-width is $w_t(X_t) = \beta_t s_t(X_t)$, but the multiplier β_t is not part of the information telescope; it is the price of calibration and optimism. The proof therefore uses s_t and w_t for different

purposes: the log-determinant argument sums s_t^2, while w_t bounds instantaneous regret. Under posterior averaging, the log determinant is mutual information. At a fixed truth, however, the MAIR log gain is a KL divergence from the true predictive law to the posterior predictive law and is not equal term by term to $\frac{1}{2}\log(1+s_t^2/\lambda)$. The theorem combines calibration with an elliptical-potential bound; it does not identify these two quantities. GP–UCB pays the realized C_T^{GP} in its regret analysis, even when the multiplier used to select actions is calibrated by the worst-case Γ_T^{GP}.

Kernel and linear formulations. Under the canonical feature map $\varphi(x) = k(x,\cdot)$,

$$f(x) = \langle f, \varphi(x)\rangle_{\mathcal{H}_k}.$$

The zero-mean GP posterior mean and variance coincide with the Hilbert-space ridge estimate and ellipsoidal width. Hence GP–UCB and the corresponding ridge LinUCB rule generate identical actions, observations, and regret on every sample path. For finite $\mathcal{X}$, the Euclidean dimension is $\mathrm{rank}(K_{\mathcal{X}})$. This is the rank of the full action geometry, not the rank of the design observed during T rounds: unsampled directions remain in the UCB maximization. Proposition E.1 gives the precise operator identity and its infinite-dimensional qualification.

6.2 A full-RKHS separation for maximal-information GP–UCB

The finite-scale construction underlying the fixed-kernel result is given in Appendix E.2. Building on the hub–cloud construction of Xu (2026), our construction uses a small-amplitude cloud that an unrestricted minimax policy can afford to ignore, even though the cloud's multiplicity inflates the maximal-information exploration multiplier. The weighted-basis Gaussian geometry of that work is already a linear feature geometry: if an isonormal Gaussian process is written as $G_x = W(\psi(x))$, then

$$k(x,x') = \mathbb{E}[G_x G_{x'}] = \langle \psi(x), \psi(x')\rangle.$$

The same vectors $\psi(x)$ are therefore both kernel features and linear-bandit actions. No infinite-dimensional intermediate embedding is needed for the finite construction. The new ingredients here are the reversal of the hub–cloud amplitude ordering and the sequential comparison with the minimax value over the full unit parameter ball. Infinite dimension is used only in the gluing argument that combines the horizon-dependent finite blocks into one fixed kernel. The result is complementary to the minimax adaptivity gap of Maiti et al. (2026): their theorem compares adaptive and nonadaptive procedures for fixed-confidence best-arm identification, whereas ours compares a particular adaptive optimistic rule with the unrestricted cumulative-regret minimax benchmark.

For a bounded kernel k and unit Gaussian observation noise, define the maximal information gain

$$\overline{\Gamma}_s(k) := \sup_{x_{1:s}\in\mathcal{X}^s} \frac{1}{2}\log\det(I + K_{x_{1:s},x_{1:s}}), \qquad \overline{\Gamma}_0(k) := 0. \tag{65}$$

Starting from the zero-mean GP with covariance k and unit ridge parameter, let m_t, s_t be the posterior mean and standard deviation. The single anytime rule studied below is

$$\overline{\beta}_t := 1 + \sqrt{2\left\{\overline{\Gamma}_{t-1}(k) + 2\log(t\vee 2)\right\}}, \qquad X_t \in \arg\max_{x\in\mathcal{X}} \left\{m_t(x) + \overline{\beta}_t s_t(x)\right\}, \tag{66}$$

with a fixed tie-breaking rule.

The multiplier in (66) has the same maximal-information form as the modern anytime calibration of Chowdhury and Gopalan (2017). The same construction also covers the exploration

schedule in the original RKHS GP–UCB theorem of Srinivas et al. (2010). To state this precisely, fix any $\delta \in (0,1)$ and define its unit-ball specialization, in the present multiplier notation, by

$$\Gamma_t^{\mathrm{SKKS}}(k) := \sup_{\substack{A \subset \mathcal{X} \\ |A|=t}} \frac{1}{2} \log\det(I + K_A),$$

and

$$b_t^{\mathrm{SKKS}} := \left[2 + 300\, \Gamma_t^{\mathrm{SKKS}}(k) \log^3\left(\frac{t}{\delta}\right)\right]^{1/2}, \qquad X_t \in \arg\max_{x \in \mathcal{X}} \{m_t(x) + b_t^{\mathrm{SKKS}} s_t(x)\}, \tag{67}$$

with the same fixed tie-breaking convention. Srinivas et al. write the acquisition bonus as $\sqrt{\beta_t}\, s_t$, assume $\|f\|_k^2 \le B$, and set $\beta_t = 2B + 300\gamma_t \log^3(t/\delta)$; equation (67) is obtained by setting $B = 1$, $\gamma_t = \Gamma_t^{\mathrm{SKKS}}(k)$, and unit noise variance. Their frequentist confidence theorem assumes bounded martingale noise. Here we analyze the same acquisition rule under independent Gaussian noise, so the lower bound uses the original exploration schedule but does not invoke that confidence theorem.

The two fixed-kernel rules use the same zero-mean GP posterior and the same UCB acquisition map $m_t(x) + b_t s_t(x)$; they differ in their global confidence calibration, information-gain convention, and logarithmic factors. The finite-horizon rule (99) in Appendix E.2 is a transparent prototype of the anytime maximal-information form, not either fixed-kernel rule verbatim. Appendix H.5.1 verifies the anytime and original schedules separately.

Theorem 6.3 (Fixed-kernel separation: GP–UCB versus AIR and minimax)**.** *For every $0 < \alpha < 1/4$, there exist a compact countable action space $\mathcal{X}$, a bounded continuous kernel k on $\mathcal{X}$, its fixed unit RKHS ball $\mathcal{F}$, a sequence $T_m \to \infty$, and a single epochwise finite-marginal action-index AIR Bellman policy such that, under observations $Y_t = f(X_t) + \xi_t$ with $\xi_t \overset{\text{i.i.d.}}{\sim} N(0,1)$,*

$$\inf_{\mathsf{Alg}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \operatorname{Reg}_{T_m}(\mathsf{Alg}) = \Theta(T_m^{1-\alpha}), \tag{68}$$

and the AIR Bellman policy, fixed in advance and independent of the evaluation horizon, satisfies

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \operatorname{Reg}_{T_m}(\text{AIR–Bellman}) = O(T_m^{1-\alpha}). \tag{69}$$

In contrast, both the single anytime GP–UCB rule (66) and, for every fixed $\delta \in (0,1)$, the original-schedule rule (67) have

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \operatorname{Reg}_{T_m}(\mathsf{R}) = \Omega(T_m), \qquad \mathsf{R} \in \{\text{GP–UCB}_{\text{any}}, \text{GP–UCB}_{\text{SKKS}}\}. \tag{70}$$

For each of these GP–UCB rules, both the minimax ratio and the worst-case regret ratio relative to the AIR Bellman policy are $\Omega(T_m^{\alpha})$ along an infinite subsequence.

Proof. See Appendix H.5.1.

6.3 Finite-scale illustration of the hub–cloud mechanism

The preceding results are nonasymptotic; this experiment illustrates only their finite-scale mechanism. For each displayed horizon, take the small-cloud instance of Appendix E.2 with $\alpha = 1/5$, $a = T^{-1/5}$, $N = 4T$, and hub gap $\Delta = 1/4$.

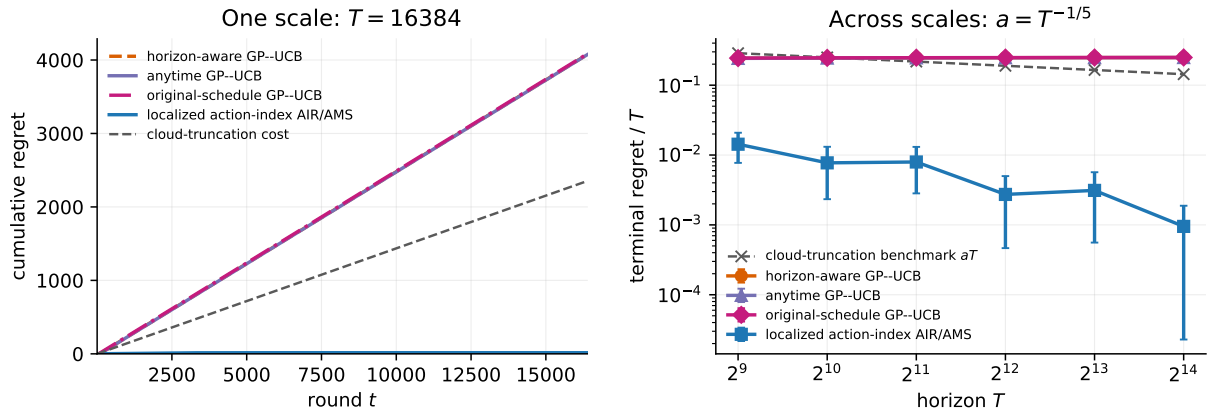

Figure 1: Finite-scale hub–cloud experiment. Algorithmic curves show means and two standard errors over 200 runs under the common positive hub truth; the truncation curve is a deterministic counterfactual. Left: cumulative pseudoregret at $T = 16384$ for three GP–UCB calibrations and the localized AIR/AMS control; the dashed curve is the deterministic one-hot-cloud truncation cost. Right: terminal regret divided by T, with normalized truncation benchmark $a = (aT)/T$. The finite kernel and $a = T^{-1/5}$ change with T, so the experiment illustrates the block mechanism rather than the glued fixed-kernel theorem.

On each finite block we evaluate three GP–UCB calibrations: the horizon-aware rule (99), the finite-block specialization of the anytime rule (66), and the unit-ball specialization of the original schedule (67), with $\delta = 0.1$. We compare them under the same positive hub truth with a numerical implementation of the robust action-index AIR/AMS control on $\{x_0, x_h\}$. This control is a direct two-model Gaussian specialization of the AIR/AMS saddle of Xu and Zeevi (2025); its relation to Lattimore and György (2021) is through the EBO optimization principle and the corresponding minimax/KL duality, not through a literal implementation of the functional-estimator mirror-descent algorithm.

Proposition E.1 shows that no new simulation is required for the linear-bandit interpretation: after replacing each action label by its feature vector, the ridge estimate, confidence width, selected actions, observations, and regret are identical on every sample path.

At $T = 16384$, the 200-run mean regrets under the common positive hub truth are 4074.8, 4072.5, and 4083.8 for the horizon-aware, anytime, and original-schedule GP–UCB rules, respectively, compared with 15.6 for AIR/AMS. Thus the three GP–UCB calibrations incur approximately 261–262 times as much mean regret. The right panel shows that this finite-block mechanism persists across the displayed scales. The experiment does not establish the fixed-kernel theorem: the finite kernel changes with T, whereas Theorem 6.3 concerns one fixed kernel and an infinite subsequence. The truncation calculation, negative retained truth, and Bellman-bracket checks are given in Appendix E.6.

6.4 Finite-dimensional consequence

The finite small-cloud block is also a genuine stochastic linear bandit. For every sufficiently large T, Corollary E.3 constructs $T+1$ actions in $\mathbb{R}^T$ for which the specified maximal-information-calibrated LinUCB rule has worst-case regret $\Omega(T)$, while the unrestricted minimax regret is $\Theta(T^{1-\alpha})$. The ratio is therefore $\Omega(T^{\alpha})$. The action set depends on T and has heterogeneous norms; obtaining a comparable separation for $d = o(T)$, fixed d, or a more homogeneous geometry remains open. Ap-

pendix E gives the full statement, proof, and comparison with layered contextual methods, asymptotic optimism lower bounds, OFUL, and the adaptive-versus-nonadaptive best-arm-identification result of Maiti et al. (2026).

Remark 6.4 (Bellman and hub–cloud interpretation)**.** *The weighted-basis hub–cloud construction of Xu (2026) has cloud scale larger than hub scale. When that geometry is reinterpreted as a kernel action set, it can draw GP–UCB toward the cloud, but the same large cloud scale also makes the full RKHS ball minimax-hard. Theorem E.2 reverses the amplitude comparison while retaining the multiplicity comparison. The pointwise RKHS envelope $|f(c_j)| \le a$ makes the entire cloud worth at most aT to an unrestricted algorithm. Nevertheless, its maximal information gain is of order Ta^2, so the global confidence multiplier is of order $a\sqrt{T}$ and the acquisition width of each unqueried cloud point is of order $a^2\sqrt{T}$. When $a = T^{-\alpha}$ and $\alpha < 1/4$, that width diverges even though the decision value of the cloud vanishes.*

In Bellman terms, the upper analysis compares retaining the cloud in the remaining-horizon problem with truncating it at a cost of at most a per remaining round. The policy used in the theorem implements the resulting finite-marginal truncation on a predetermined epoch schedule and pays only for the optimal-action index on the retained set. The GP–UCB relaxation replaces that decision-aligned comparison by a global uniform-optimism coefficient. Maximal information gain affects the trajectory, not only the final bound: through $\overline{\beta}_t$, it draws the algorithm into the low-value cloud. The relevant localization is the elementary pointwise envelope in this example, and in more structured problems it may be a posterior region, a norm shell, or an optimal-action index.

6.5 Resolved and open questions

There is no single conjecture covering every kernel and every rule called GP–UCB. The recurring question is whether the gap between minimax rates and known guarantees for the acquisition rule reflects loose analysis or the rule itself (Vakili et al., 2021; Whitehouse et al., 2023; Wang and Zhang, 2026). Theorem 6.3 answers this for two specified global calibrations: both have linear regret along an infinite subsequence under one fixed hub truth, although the minimax regret on the same RKHS ball is sublinear. It does not cover arbitrary localized, predictable, or data-dependent tuning.

For standard Matérn kernels, Wang and Zhang (2026) prove a complementary horizonwise lower bound for prespecified polynomial–logarithmic effective-optimism schedules, subject to their monotonicity, exponent, and simultaneous uniform-confidence assumptions. Their hard function may depend on the horizon. Our theorem instead treats two specified schedules on one fixed, spatially inhomogeneous kernel and one fixed truth. Neither result contains the other. In particular, their result rules out polynomially growing effective optimism under its assumptions. Neither result settles whether merely polylogarithmic effective optimism is minimax optimal up to polylogarithmic factors, or whether localized or data-dependent tuning can attain the minimax rate. A fixed-truth, linear-regret Matérn separation for either schedule in Theorem 6.3 also remains open. Appendix E.4 states the Wang–Zhang hypotheses and rate in detail.

6.6 Further matching examples

Gaussian multi-armed bandits give a finite-index Bellman–Fano entropy calculation and a matching minimax regret rate $\Theta(\sqrt{KT})$. Hypercube linear bandits give a high-dimensional indexed lower bound of order $d\sqrt{T}$, matched by standard upper bounds up to logarithmic factors. These examples do not claim that MOSS or OFUL is itself the optimizer of the exact log-potential Bellman program. The complete calculations are in Appendices F and G.

7 Conclusion

The paper develops a general information-theoretic minimax framework for sequential decision-making problems. The Bellman-sufficient state retains what is needed for the controlled experiment; the index identifies what is worth learning. The fixed-truth logarithmic potential and the Bellman–Fano ghost entropy then provide upper and lower information accounts on the same representation. Under the conditions stated in Theorem A.8, matching them is the sequential counterpart of matching information and local entropy in a fixed experiment.

The kernel construction illustrates both the algorithmic and statistical power of this framework. Along the constructed subsequence (T_m), the fixed RKHS ball has minimax regret $\Theta(T_m^{1-\alpha})$, attained by a predetermined finite-marginal AIR/AMS policy, while two global GP–UCB calibrations incur $\Omega(T_m)$ regret because low-value cloud directions inflate their exploration bonuses. The same feature geometry gives a horizon-dependent linear-bandit instance with $T+1$ actions in $\mathbb{R}^T$ and a polynomial minimax gap for the specified maximal-information-calibrated LinUCB rule. The glued theorem is likewise a separation on one fixed infinite-dimensional Hilbert-space linear-bandit action set, whereas the finite-dimensional action set depends on T. Whether such a finite-dimensional separation persists for $d = o(T)$, especially fixed d, and whether localized or data-dependent GP–UCB can recover the minimax rate remain open. The resulting complexity belongs not to the raw model class alone, but to the Bellman-sufficient representation on which the upper and lower bounds meet.

A Technical form of the information-complexity sandwich

This appendix states the admissibility, entropy-comparison, growth, and calibration conditions underlying the information-risk sandwich summarized in Section 5.4.

A.1 Code length, hard priors, and ghost entropy

Let $\Omega_0 \subseteq \Omega$ be the problem class being lower and upper bounded, let $\chi : \Omega \to \mathcal{Y}$ be the index, and let q_1 be an initial reference marginal on $\mathcal{Y}$. The fixed-truth upper telescope pays the worst-case initial coordinate code length

$$L_0(\Omega_0; q_1, \chi) := \sup_{\omega \in \Omega_0} \log \frac{1}{q_1(\chi(\omega))}, \tag{71}$$

with the convention that the value is infinite if $q_1(\chi(\omega)) = 0$ for some $\omega \in \Omega_0$.

All Bellman values in this subsection are conditional on a fixed initial packet and resulting state s_1. If S_1 is random, one conditions on O_0 and averages the entire conditional bound. In particular, the conditional ghost masses must be averaged before applying $-\log$; the conditional ghost entropies cannot generally be averaged instead.

For a hard prior μ supported on Ω_0, write $q_1 = \chi_{\#}\mu$. At an average-loss radius r, the ghost-good probability of an algorithm Alg is

$$p^{\mathsf{Alg}}_{\mu,r} := p^{\mathsf{Alg}}_r(\mu, \chi) = \mathbb{P}_{Y \sim q_1,\, H'_T \sim \bar{\mathbb{P}}^{\mathsf{Alg}}_\mu}\{\bar{L}_\chi(Y, H'_T) \le r\}, \tag{72}$$

where Y and the algorithm-induced reference history H'_T are independent under the ghost law. The algorithm-uniform ghost mass and its entropy are

$$p^\star_{\mu,r} := \sup_{\mathsf{Alg} \in \mathfrak{A}^{\mathrm{na}}_T} p^{\mathsf{Alg}}_{\mu,r} = \Gamma^{r,\star}_1(q_1, s_1, 0), \qquad E^\star_{\mu,r} := \log \frac{1}{p^\star_{\mu,r}}. \tag{73}$$

A ghost-mass supersolution gives $p^\star_{\mu,r} \le \overline{\Gamma}^r_1(q_1, s_1, 0)$ and hence the computable entropy lower bound $-\log \overline{\Gamma}^r_1(q_1, s_1, 0) \le E^\star_{\mu,r}$.

Definition A.1 (Bellman–Fano admissible hard prior)**.** *Fix $r > 0$. A prior μ supported on Ω_0 is* Bellman–Fano admissible at radius r *if the lower Bellman recursions, or valid supersolutions of them, provide a nonnegative indexed-information upper bound C_μ and a valid algorithm-uniform ghost-entropy lower bound $\underline{E}_{\mu,r} \le E^\star_{\mu,r} = \log(1/p^\star_{\mu,r})$ such that*

$$C_\mu + \log 2 \le \frac{1}{2}\underline{E}_{\mu,r}. \tag{74}$$

The intrinsic choice is $C_\mu = C^\star_1(s_1)$ and $\underline{E}_{\mu,r} = E^\star_{\mu,r} = -\log \Gamma^{r,\star}_1(q_1, s_1, 0)$, where $C^\star$ and $\Gamma^{r,\star}$ are the exact information-capacity and ghost-good Bellman values from Section 5.3. Computable proofs may use the conservative choices $C_\mu = \overline{C}_1(s_1)$ and $\underline{E}_{\mu,r} = -\log \overline{\Gamma}^r_1(q_1, s_1, 0)$; since $\overline{\Gamma}^r_1$ upper-bounds the algorithm-uniform ghost mass, this is indeed no larger than $E^\star_{\mu,r}$.

The condition $p^\star_{\mu,r} \le \varrho < 1$ sometimes appears as a convenient nondegeneracy diagnostic: it prevents the lower entropy denominator from vanishing. It plays the same mathematical role as the nondegeneracy requirements in offset or constrained DEC comparisons (Foster et al., 2021, 2023): without positive decision separation at the chosen information scale, the denominator of the rate comparison is zero and no lower bound can be extracted. In the Bellman–Fano formulation this condition need not be imposed separately once the hard prior is defined through admissibility.

Lemma A.2 (Admissibility implies nondegenerate ghost entropy)**.** *If μ is Bellman–Fano admissible at radius r, then*

$$E^\star_{\mu,r} \geq \underline{E}_{\mu,r} \geq 2\log 2, \qquad p^\star_{\mu,r} \leq \frac{1}{4}.$$

Consequently a separate assumption $p^\star_{\mu,r} \leq \varrho < 1$ is unnecessary for any admissible hard prior. Conversely, if one starts from a proposed prior and only knows $p^\star_{\mu,r} \leq \varrho < 1$, then the entropy denominator is at least $\log(1/\varrho)$, but this by itself does not prove Bellman–Fano admissibility because the information upper bound C_μ must also satisfy (74).

Proof. See Appendix H.1.2.

Proposition A.3 (Finite-index entropy comparison with an explicit hard prior)**.** *Let $\mathcal{Y}_h \subseteq \chi(\Omega_0)$ be finite with $|\mathcal{Y}_h| = M \geq 2$. Choose one representative environment $\omega^y \in \Omega_0$ for each $y \in \mathcal{Y}_h$ such that $\chi(\omega^y) = y$, and let μ_h be the uniform prior on $\{\omega^y : y \in \mathcal{Y}_h\}$. Then q_1 is uniform on $\mathcal{Y}_h$, and*

$$L_0(\{\omega^y : y \in \mathcal{Y}_h\}; q_1, \chi) = \log M.$$

If μ_h is Bellman–Fano admissible at radius r, then the entropy mismatch between the upper coordinate code length and the lower ghost entropy satisfies

$$\frac{L_0}{E^\star_{\mu_h,r}} \leq \frac{\log M}{2\log 2}. \tag{75}$$

Thus, without pursuing constant matching, admissibility alone gives an at-most-logarithmic finite-index entropy gap. More sharply, if for every reference history h',

$$G_r(h') := \{y \in \mathcal{Y}_h : \bar{L}_\chi(y, h') \leq r\}$$

has cardinality at most m_r, for some integer $1 \leq m_r < M$, then

$$E^\star_{\mu_h,r} \geq \log \frac{M}{m_r}, \qquad \frac{L_0}{E^\star_{\mu_h,r}} \leq \frac{\log M}{\log(M/m_r)}. \tag{76}$$

In particular, the one-sided constant-factor entropy comparison used here follows from $E^\star_{\mu_h,r} \geq c\log M$, for example through $m_r \leq M^{1-\alpha}$ for some constant $\alpha > 0$. A two-sided entropy equivalence requires the stronger statement $E^\star_{\mu_h,r} \asymp \log M$. A mere nondegeneracy bound $p^\star_{\mu_h,r} \leq \varrho < 1$ gives only $L_0/E^\star_{\mu_h,r} \leq \log M/\log(1/\varrho)$.

Proof. See Appendix H.1.3.

A.2 From entropy comparison to regret comparison

An entropy comparison alone is not a regret comparison. The missing ingredient is a regularity property of the upper Bellman value as its information budget changes. This is the same structural reason that constrained DEC bounds are stated through a coefficient, localization, or fixed-point condition rather than through a bare logarithmic-model-cardinality entropy term (Foster et al., 2023): cardinality controls the amount of information charged, but a risk modulus is still needed to turn that information budget into a regret radius.

For a fixed state/index representation and the log-penalized Bellman program in Section 4.1, write

$$\mathsf{U}_T(L) := \inf_{\gamma > 0} \left\{ W_1^\gamma(s_1) + \gamma L + \Delta_T^\gamma \right\}, \tag{77}$$

where W^γ is the log-penalized Bellman value and $\Delta_T^\gamma \geq 0$ collects explicit statewise bracket, approximation, and optimization errors for that value as in Theorem 4.2. For the exact surely calibrated Bellman dynamic program, $\Delta_T^\gamma = 0$. By construction, U_T is nondecreasing in L.

Assumption A.4 (Bellman upper-growth condition)**.** *There exist constants $C_{\mathrm{gr}} \geq 1$ and $\beta \in [0,1]$ such that, for all $L > 0$ and $a \geq 1$,*

$$\mathsf{U}_T(aL) \leq C_{\mathrm{gr}} a^{\beta} \mathsf{U}_T(L).$$

This is a regularity condition on the chosen upper value. It is not a consequence of finite index cardinality alone; it is the Bellman analogue of the sub-root or growth condition used to convert entropy or localized information radii into risk radii in fixed experiments and in constrained DEC analyses.

Lemma A.5 (Why a growth condition is needed)**.** *No regret-ratio comparison follows from an entropy-ratio comparison for an arbitrary nondecreasing upper value U_T. More precisely, fix $0 < E < L$ and any $B > 0$. There is a nondecreasing function U such that $\mathsf{U}(L)/\mathsf{U}(E) \geq B$.*

Proof. See Appendix H.1.4.

Definition A.6 (Upper-at-lower-entropy calibration)**.** *Let μ be Bellman–Fano admissible at radius r with algorithm-uniform ghost entropy $E^{\star}_{\mu,r}$. The upper value is* calibrated at the lower entropy scale *if*

$$\mathsf{U}_T(E^{\star}_{\mu,r}) \leq C_{\mathrm{base}}\, Tr \tag{78}$$

for a universal or problem-controlled constant C_{base}. Equivalently, the phrase "the upper and lower values close at the same radius" means (78); it is not an additional philosophical assumption. It can be verified directly, by a localized or constrained information-gain estimate, or by choosing r as a fixed point of the rate equation $\mathsf{U}_T(E^{\star}_{\mu,r}) \asymp Tr$.

Proposition A.7 (Three scale-closing diagnostics)**.** *For a Bellman–Fano admissible pair (μ, r), each of the following conditions implies upper-at-lower-entropy calibration (78).*

(i) Direct budget control: $\mathsf{U}_T(E^{\star}_{\mu,r}) \leq C_{\mathrm{base}} Tr$.

(ii) Coefficient control: there exists $\gamma > 0$ such that

$$W_1^{\gamma}(s_1) + \gamma E^{\star}_{\mu,r} + \Delta_T^{\gamma} \leq C_{\mathrm{base}} Tr.$$

(iii) Fixed-point control: r is chosen so that $\mathsf{U}_T(E^{\star}_{\mu,r})/T \leq C_{\mathrm{base}} r$.

Condition (ii) is the form closest to constrained DEC and localized information-gain analyses: the same localized comparison that controls the information budget also bounds the decision term at radius r. Condition (iii) packages the same requirement as a rate fixed point.

Proof. See Appendix H.1.5.

Theorem A.8 (Frequentist Bellman information-risk sandwich)**.** *Fix $\Omega_0 \subseteq \Omega$, a horizon T, an index χ, a Bellman-sufficient state/reference update, and an initial index law q_1 with finite $L_0 = L_0(\Omega_0; q_1, \chi)$. Suppose the upper side admits a calibrated log-penalized Bellman bound whose budget value is U_T in (77). Then*

$$\mathfrak{R}^{\star}_T(\Omega_0) \leq \mathsf{U}_T(L_0). \tag{79}$$

Conversely, if a prior μ supported on Ω_0 is Bellman–Fano admissible at radius r, then

$$\mathfrak{R}^{\star}_T(\Omega_0) \geq \frac{Tr}{2}. \tag{80}$$

If, in addition, Assumption A.4 holds and the upper value is calibrated at the lower entropy scale in the sense of Definition A.6, then

$$\mathfrak{R}_T^\star(\Omega_0) \le C_{\mathrm{gr}} C_{\mathrm{base}} \left(\max\left\{1, \frac{L_0}{E^\star_{\mu,r}}\right\} \right)^\beta Tr, \tag{81}$$

and hence the ratio between the displayed upper value and the Bellman–Fano lower bound $Tr/2$ is at most

$$2C_{\mathrm{gr}} C_{\mathrm{base}} \left(\max\left\{1, \frac{L_0}{E^\star_{\mu,r}}\right\} \right)^\beta.$$

If, in addition, $\mu = \mu_h$ is the uniform representative prior of Proposition A.3, the upper reference law is the same marginal $q_1 = \chi_\# \mu_h$, and $\chi(\Omega_0)$ consists of those M indices, then $L_0 = \log M$, and admissibility alone gives the crude logarithmic comparison

$$\frac{L_0}{E^\star_{\mu,r}} \le \frac{\log M}{2 \log 2},$$

while the multiplicity/localization condition $|G_r(h')| \le m_r$ gives

$$\frac{L_0}{E^\star_{\mu,r}} \le \frac{\log M}{\log(M/m_r)}.$$

The same statement applies after replacing Ω_0 throughout by the representative subfamily used by a uniform hard prior. Thus logarithmic-gap matching on a common finite comparison class requires only Bellman–Fano admissibility, upper growth, and calibration at the lower entropy scale. Constant-factor matching follows from the stronger one-sided localization $E^\star_{\mu,r} \ge c \log M$, or from an equivalent constraint or fixed-point argument; a two-sided entropy characterization additionally requires $E^\star_{\mu,r} = O(\log M)$.

Proof. See Appendix H.1.6.

B Examples of Bellman-sufficient state representations

The following examples verify Definition 2.2 directly. They also fix the interpretation of the state used later: the fixed truth remains the environment, while sufficient statistics, posteriors, and finite marginals are retained Bellman states or reference objects.

Adversarial and estimated losses. The framework accommodates adversarial losses beyond stochastic regret (Cesa-Bianchi and Lugosi, 2006; Abernethy et al., 2011). A nonstochastic or adaptive adversary may be treated as part of the fixed environment ω when its rule is nonanticipating. If the adversary observes the history and the learner's announced mixed action, then that mixed action is the current control, not an additional component of the past, and the primitives are

$$\ell_t(\omega, H_{t-1}, p, a), \qquad P_{\omega,t}(\cdot \mid H_{t-1}, p, a).$$

With the full history as state and p as a control argument, this representation is automatic for any fixed nonanticipating adversary rule. Estimated-loss upper bounds may replace the true loss in the Bellman bracket by a state-measurable surrogate or domination bound while carrying the approximation error. A lower bound still requires the cumulative indexed loss used in the ghost event to be lower bounded, usually nonnegative after a known shift.

B.1 Basic examples of Bellman-sufficient representations

Guided by the classical intuition of sufficient statistics, the Bellman-sufficiency framework is particularly transparent in structured bandit problems.

Example B.1 (Classical sufficient statistics as one-step Bellman states)**.** Consider a nonadaptive dominated experiment with observations $X_1, \dots, X_n$ from P_θ, and no control. Suppose that for every prefix t, $T_t = T_t(X_1, \dots, X_t)$ is sufficient for the prefix experiment and admits a recursive update $T_t = u_t(T_{t-1}, X_t)$. If the loss and index depend on θ only through quantities whose posterior distribution is determined by T_t, then the process $S_t = (t, T_{t-1})$ is Bellman sufficient. Predictive sufficiency follows from the prefix factorization; posterior predictive and index sufficiency follow because the posterior or conditional reference law depends on the history only through T_t. In a regular exponential family,

$$p_\theta(x) = h(x) \exp\{\vartheta(\theta)^\top T(x) - A(\theta)\},$$

with conjugate or otherwise statistic-measurable reference law, the cumulative statistic $\sum_{i<t} T(X_i)$ gives the corresponding state. This is the static prototype for all examples below.

Example B.2 (Finite stochastic multi-armed bandits)**.** Let $\Omega \subseteq \Theta^K$ and let arm a produce an observation from P_{θ_a}. For unit-variance Gaussian arms, $P_{\theta_a} = N(\theta_a, 1)$. A reference Gaussian prior with independent coordinates gives the posterior state

$$S_t = \big(t, (n_{t,a}, \bar{X}_{t,a}, v_{t,a})_{a=1}^K\big),$$

where $n_{t,a}$ is the number of previous pulls of arm a, $\bar{X}_{t,a}$ is the empirical mean when $n_{t,a} > 0$, and $v_{t,a}$ is the posterior variance. Equivalently one may store the posterior hyperparameters. For a fixed truth $\omega = \theta$,

$$P_{\omega,s,a} = N(\theta_a, 1), \qquad \ell_\omega(s,a) = \max_b \theta_b - \theta_a,$$

and the update of $(n, \bar{X}, v)$ after observing arm a is a function of (s, a, o). The posterior predictive law is $N(m_{t,a}, 1 + v_{t,a})$, and for the action index $Y = A^\star(\theta)$ the conditional predictive law $P^y_{s,a}$ and conditional loss $\ell^\chi_y(s,a)$ are obtained by integrating the same Gaussian posterior over the event $A^\star(\theta) = y$. Hence Definition 2.2 is satisfied. The empirical mean alone is not sufficient: the count or posterior variance is needed for prediction, confidence, and information gain.

Example B.3 (Finite-dimensional Gaussian linear bandits)**.** Let $\Omega \subseteq \mathbb{R}^d$, actions satisfy $\|a\|_2 \le 1$, and

$$O_t = \langle \theta, A_t \rangle + \xi_t, \qquad \xi_t \sim N(0,1).$$

Under a Gaussian reference prior, the exact Bayesian state is $S_t = (t, m_t, \Sigma_t)$, where

$$\Sigma_{t+1}^{-1} = \Sigma_t^{-1} + A_t A_t^\top, \qquad m_{t+1} = \Sigma_{t+1}\{\Sigma_t^{-1} m_t + A_t O_t\}.$$

For fixed θ, $P_{\theta,s,a} = N(\langle \theta, a \rangle, 1)$ and the regret loss is $\sup_{\|b\|_2 \le 1} \langle \theta, b \rangle - \langle \theta, a \rangle$. The posterior predictive law is $N(\langle m_t, a \rangle, 1 + a^\top \Sigma_t a)$. For any of the indices $\chi(\theta) = \theta$, $\chi(\theta) = \theta/\|\theta\|_2$ (with an arbitrary fixed value at $\theta = 0$), or $\chi(\theta) = A^\star(\theta)$, conditioning the Gaussian posterior on $\chi(\theta) = y$ determines $P^y_{s,a}$ and $\ell^\chi_y(s,a)$. Therefore the Gaussian mean–covariance pair is an exact Bellman-sufficient state. A frequentist UCB state $(\hat{\theta}_t, V_t, \beta_t)$ is different: it can control a fixed-truth bracket on the calibration event $\|\hat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t$, but it is not an exact posterior-reference state unless a reference belief or confidence-to-belief map is added.

Example B.4 (Finite active marginals in kernel bandits)**.** Let $f^\star$ belong to an RKHS on a possibly infinite domain, but suppose the learner is evaluated on a finite active set $\mathcal{X} = \{x_1, \dots, x_n\}$. Under the GP reference law on the active reward vector $F = (f(x_1), \dots, f(x_n))$, the finite marginal posterior $(m_t, \Sigma_t) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$ is Bellman sufficient for the finite experiment. For fixed $f^\star$, pulling x_i has law $N(f^\star(x_i), \lambda)$ in the Gaussian reference calculation; the posterior predictive law is $N(m_{t,i}, \lambda + \Sigma_{t,ii})$; and the usual rank-one Gaussian update is a function of (m_t, Σ_t, i, O_t). If the index is the optimal active action $Y = A^\star(F)$, then index sufficiency follows by conditioning the finite Gaussian posterior on $A^\star(F) = y$, using the fixed tie-breaking rule. The infinite-dimensional RKHS function is still the frequentist truth; the sufficient state is finite only because the decisions and observations use the active marginal.

B.2 The full environment posterior as a fallback sufficient state

For sequential decision making, a natural Bellman-sufficient representation is obtained by retaining the full environment posterior. For suitably defined Markovian model classes over the state space, which include standard episodic model-based reinforcement learning settings, this posterior serves as a fallback sufficient state.

Example B.5 (Full environment posterior as a fallback sufficient state)**.** *Consider a realizable model class in which each environment $\omega \in \Omega$ specifies, for every retained physical state or context z, action a, and stage t, an observation kernel $P_{\omega,t}(\cdot \mid z, a)$, a loss $\ell_{\omega,t}(z, a)$, and the physical-state transition rule. Let Z_t denote the physical state, context, stage, or episode coordinate needed to make these kernels Markov. Fix a reference prior μ and set*

$$\Pi_t = \mathcal{L}_\mu(\Omega \mid H_{t-1}), \qquad S_t = (Z_t, \Pi_t).$$

Assume that the observation laws are dominated on the relevant support and that the Bayesian update denominator is positive along the reference experiment. Fixed-truth coordinate identities additionally require the evaluated truth's predictive laws to be absolutely continuous with respect to the corresponding reference predictive laws. Then S_t gives an exact Bellman-sufficient state. Indeed, for each fixed truth ω,

$$P_{\omega,S_t,a} = P_{\omega,t}(\cdot \mid Z_t, a), \qquad \ell_\omega(S_t, a) = \ell_{\omega,t}(Z_t, a),$$

and the next retained state is obtained by updating the physical coordinate and the posterior

$$\Pi_{t+1}(d\omega) = \frac{p_{\omega,t}(O_t \mid Z_t, A_t)\Pi_t(d\omega)}{\int p_{\omega',t}(O_t \mid Z_t, A_t)\Pi_t(d\omega')}.$$

For an index $\chi : \Omega \to \mathcal{Y}$, the indexed marginal and conditional predictive components are

$$q_t = \chi_\# \Pi_t, \qquad \Pi_t^y = \mathcal{L}_{\Pi_t}(\Omega \mid \chi(\Omega) = y),$$

$$P^y_{S_t,a} = \int P_{\omega,t}(\cdot \mid Z_t, a)\, \Pi_t^y(d\omega), \qquad \ell^\chi_y(S_t, a) = \int \ell_{\omega,t}(Z_t, a)\, \Pi_t^y(d\omega).$$

Thus prediction, loss evaluation, posterior/reference updating, and indexed information accounting all close on S_t. This verifies Definition 2.2.

Example B.6 (DMSO as a posterior-state specialization)**.** *In independent-episode DMSO (Foster et al., 2021), fix a reference prior and likelihood, take the Bellman clock to be the episode index, and let $\Pi_k = \mathcal{L}(M \mid H_{k-1})$. Retaining Π_k together with any public context and decision constraints makes the predictive law, posterior update, and index marginal $\chi_\#\Pi_k$ functions of the state. More compressed representations require a separate sufficiency argument.*

C AIR/MAIR variational calculation and interactive Fano adaption

C.1 Posterior scoring and AIR/MAIR variational calculation

Danskin, posterior scoring, and why the AIR bracket is not accidental The short algebraic proofs above are the fixed-truth version of the variational argument behind AIR. In the Bellman-sufficient notation, fix a state s, an action law p, and the current reference marginal $q_s \in \Delta(\mathcal{Y})$ for the index. Let

$$\nu \in \Delta(\Omega \times \mathcal{Y})$$

be an admissible pair belief, usually with $Y = \chi(\Omega)$ and Y-marginal q_s. Given a prediction rule

$$Q : (s, p, a, o) \mapsto \Delta(\mathcal{Y}),$$

consider the posterior-scoring objective

$$\mathcal{J}_\gamma(s, p, \nu, Q; q_s) := \mathbb{E}_{\substack{(\Omega,Y)\sim\nu \\ A\sim p \\ O\sim P_{\Omega,s,p,A}}} \left[\ell_{\Omega,Y}(s, p) - \gamma \log \frac{Q(s, p, A, O)(Y)}{q_s(Y)} \right].$$

Here q_s is held fixed as part of the current state. If the analysis also optimizes over the index marginal itself, then that marginal is part of the Bellman control or enlarged state; the local Danskin statement is applied after fixing its current value.

For fixed (s, p, ν, q_s), the logarithmic score is convex in Q and is minimized, pointwise in (s, p, a, o), by the posterior index marginal generated by the same pair belief ν:

$$Q_\nu(s, p, a, o) = q^+_\nu(\cdot \mid s, p, a, o) := \nu(Y \in \cdot \mid s, p, a, O = o)\,,$$

with the usual regular-conditional interpretation, or equivalently the Bayes formula under a dominated observation model. The optimized value

$$\mathcal{A}_\gamma(s, p, \nu; q_s) := \inf_Q \mathcal{J}_\gamma(s, p, \nu, Q; q_s)$$

is the indexed AIR one-step objective: it is immediate indexed loss minus γ times the posterior log gain of the index.

For fixed Q and fixed q_s, the unoptimized objective is affine in the pair belief ν. Hence the optimized value is concave in ν, being the infimum of affine functions. Under the standard compactness, support, and differentiability conditions needed to apply Danskin's theorem, the directional derivative of $\mathcal{A}_\gamma$ is obtained by freezing the optimizer $Q_\nu = q^+_\nu$. Therefore the pointwise gradient in the direction of a fixed truth $(\omega^\star, y^\star)$ is

$$\ell_{\omega^\star,y^\star}(s, p) - \gamma\, \mathbb{E}_{\substack{A\sim p \\ O\sim P_{\omega^\star,s,p,A}}} \log \frac{q^+_\nu(y^\star \mid s, p, A, O)}{q_s(y^\star)},$$

which is the AIR gradient bracket used above. The bracket follows from three steps: posterior scoring selects the Bayes index predictor, Danskin's theorem differentiates through that predictor, and the coordinate log score telescopes under the maintained reference update.

The notation in Xu and Zeevi (2025, Section 5) uses the action index $Y = A^\star(\Omega)$. Taking $Y = \Omega$ gives the MAIR or model-index form. In the stationary-posterior case, ν is the current posterior over (Ω, Y) and $q_s = \nu_Y$. In robust AIR, AMS, or EBO, ν may be a selected or optimized

admissible belief, and validity comes from the corresponding Bellman bracket or calibration bound. The same concavity–convexity structure also explains the Nash-equilibrium statement in Xu and Zeevi (2025, Lemma 5.2), although the formal proof still has to check compactness, support, and boundary conditions.

Lemma C.1 (MAIR gradient bracket and three canonical specializations)**.** *Assume the observation laws are dominated and the displayed derivatives exist. For fixed s, p, ρ, μ, up to an additive constant on the probability simplex,*

$$\frac{\partial}{\partial \mu(\omega)} \mathrm{MAIR}_{\rho,\gamma}(s,p,\mu) = \ell_\omega(s,p) - \gamma \log \frac{\mu(\omega)}{\rho(\omega)} - \gamma \mathbb{E}_{a\sim p} D_{\mathrm{KL}}(P_{\omega,s,p,a} \| P_{\mu,s,p,a}).$$

Consequently, for every $\omega^\star$,

$$\begin{aligned} &\mathrm{MAIR}_{\rho,\gamma}(s,p,\mu) + \langle \nabla_\mu \mathrm{MAIR}_{\rho,\gamma}(s,p,\mu), \delta_{\omega^\star} - \mu \rangle \\ &\quad = \ell_{\omega^\star}(s,p) - \gamma \mathbb{E}_{a\sim p} D_{\mathrm{KL}}(P_{\omega^\star,s,p,a} \| P_{\mu,s,p,a}) - \gamma \log \frac{\mu(\omega^\star)}{\rho(\omega^\star)}. \end{aligned} \tag{82}$$

This bracket has three common specializations.

1. *If $\mu = \rho$, the logarithmic correction vanishes and the bracket becomes the KL-DEC offset*

$$\ell_{\omega^\star}(s,p) - \gamma \mathbb{E}_{a\sim p} D_{\mathrm{KL}}(P_{\omega^\star,s,p,a} \| P_{\mu,s,p,a}). \tag{83}$$

2. *If $\bar{\mu} \in \arg\max_{\mu \in \Delta(\Omega)} \mathrm{MAIR}_{\rho,\gamma}(s,p,\mu)$ is an interior maximizer, then*

$$\mathrm{MAIR}_{\rho,\gamma}(s,p,\bar{\mu}) + \langle \nabla_\mu \mathrm{MAIR}_{\rho,\gamma}(s,p,\bar{\mu}), \delta_{\omega^\star} - \bar{\mu} \rangle \le \mathrm{MAIR}_{\rho,\gamma}(s,p,\bar{\mu}). \tag{84}$$

   This is the model-index AMS/EBO bracket control.

3. *Let $p_{\omega,a} = dP_{\omega,s,p,a}/d\nu_{s,p,a}$ and define, for an auxiliary pair $(\bar{a}, \bar{o})$,*

$$\frac{d\mu^{1/2}_{\bar{a},\bar{o}}}{d\rho}(\omega) := \frac{\sqrt{p_{\omega,\bar{a}}(\bar{o})}}{\int \sqrt{p_{\omega',\bar{a}}(\bar{o})}\, \rho(d\omega')}. \tag{85}$$

   Let $h^2(P,Q) := 1 - \int \sqrt{dP\, dQ}$. If $\bar{a} \sim p$ and $\bar{o} \sim P_{\omega^\star,s,p,\bar{a}}$, then

$$\begin{aligned} &\mathbb{E}_{\bar{a},\bar{o}}\Big[\mathrm{MAIR}_{\rho,\gamma}(s,p,\mu^{1/2}_{\bar{a},\bar{o}}) + \langle \nabla_\mu \mathrm{MAIR}_{\rho,\gamma}(s,p,\mu^{1/2}_{\bar{a},\bar{o}}), \delta_{\omega^\star} - \mu^{1/2}_{\bar{a},\bar{o}} \rangle\Big] \\ &\quad \le \ell_{\omega^\star}(s,p) - \gamma \mathbb{E}_{a\sim p} \mathbb{E}_{\omega\sim\rho} h^2(P_{\omega^\star,s,p,a}, P_{\omega,s,p,a}). \end{aligned} \tag{86}$$

   Thus the square-root posterior turns the MAIR bracket into a Hellinger DEC offset.

Proof. See Appendix H.2.4.

C.2 Adaption of interactive Fano

Theorem 5.1 adapts the interactive Fano method of Chen et al. (2024): compare the true experiment with a ghost experiment, apply data processing to the low-loss event, and compare trajectory information with the logarithmic inverse ghost-good probability. Here the ghost experiment is the Bayesian posterior-predictive trajectory and the charged coordinate is $Y = \chi(\Omega)$. If the initial

packet is deterministic or independent of Y, the same reference history supplies both the ghost event and the exact Bellman-state chain rule

$$I_\mu(Y; H_T) = \mathbb{E}_{H'_T \sim \bar{\mathbb{P}}_\mu^{\mathsf{Alg}}} \sum_{t=1}^{T} \mathcal{I}_\chi(S'_t, p'_t).$$

If O_0 is informative, this is the conditional information acquired after O_0, and the unconditional budget includes $I_\mu(Y; O_0)$. Taking $Y = \Omega$ gives the model-index form, while $Y = A^\star(\Omega)$ gives the action-index form.

D Algorithmic controls of the Bellman bracket

D.1 UCB families: calibration plus optimism

An upper-confidence-bound (UCB) algorithm maintains a prediction and an uncertainty scale, builds a confidence band from them, and chooses an action that is optimistic within that band (Auer et al., 2002; Abbasi-Yadkori et al., 2011; Srinivas et al., 2010). We use the following notation throughout. The unmultiplied uncertainty scale is denoted by $\sigma_t(a)$ in the generic discussion and by $s_t(x)$ in the GP specialization below. The confidence half-width is

$$w_t(a) = \beta_t \sigma_t(a),$$

where β_t is the confidence multiplier. Calibration says that the fixed truth is contained in the band of radius $w_t(a)$; optimism says that the selected action maximizes the upper envelope. Together, in reward-maximization problems, these two facts give

$$r_{\omega^\star}(s, a^\star) \le m_t(a^\star) + w_t(a^\star) \le m_t(a_t) + w_t(a_t) \le r_{\omega^\star}(s, a_t) + 2w_t(a_t),$$

and hence the instantaneous regret is at most $2\beta_t\sigma_t(a_t)$. An elliptical-potential/log-determinant lemma then controls the realized sum of $\sigma_t^2(a_t)$. Thus UCB does not solve the AIR/MAIR bracket optimization directly; it controls the relevant fixed-truth bracket after the optimistic action has been chosen.

Lemma D.1 (Calibration and optimism imply bracket control)**.** *Fix a truth $\omega^\star$ and condition on an event $\mathcal{E}$ on which the algorithm's state is calibrated for that truth. Suppose that, for every reached round t, the action a_t chosen by the algorithm and a predictable uncertainty scale $\sigma_t(a) \ge 0$ satisfy*

$$\ell_{\omega^\star}(S_t, a_t) \le c_{\mathrm{opt}}\, \beta_t \sigma_t(a_t), \tag{87}$$

where c_{opt} is a numerical optimism constant; in the usual two-sided reward-confidence proof above, $c_{\mathrm{opt}} = 2$. Then for every $\gamma > 0$,

$$\ell_{\omega^\star}(S_t, a_t) - \gamma\sigma_t^2(a_t) \le \frac{c_{\mathrm{opt}}^2 \beta_t^2}{4\gamma}. \tag{88}$$

Consequently, for the deterministic-action fixed-truth bracket in (19),

$$B_{\chi,\gamma}(S_t, \delta_{a_t}; \omega^\star) \le \frac{c_{\mathrm{opt}}^2 \beta_t^2}{4\gamma} + \gamma\left(\sigma_t^2(a_t) - J_\chi(S_t, \delta_{a_t}; \omega^\star)\right), \tag{89}$$

where the fixed-truth index log gain is defined in (17). If, in addition, the log-determinant/elliptical-potential argument gives a realized bound

$$\sum_{t=1}^{T} \sigma_t^2(a_t) \le c_{\mathrm{sn}} G_T,$$

for an information telescope G_T, then summing (89) reduces UCB regret control to the comparison between G_T and the fixed-truth log-gain telescope.

Proof. See Appendix H.3.2.

Here $\sigma_t^2 - J_\chi$ is only the difference between the tractable self-normalized variance proxy and the fixed-truth log gain. The elliptical potential controls the former, while β_t separately converts calibrated uncertainty into regret through optimism.

D.2 E2D: robust one-step offset optimization

The estimation-to-decision (E2D) algorithms (Foster et al., 2021) choose p by solving a one-round robust DEC offset at the current time-state pair. In model-index notation the schematic KL form is

$$\inf_{p\in\Delta(\mathcal{A}_t(s))} \sup_{\omega\in\mathfrak{C}_t(s)} \left\{\ell_\omega(s,p) - \gamma \mathbb{E}_{a\sim p} D_{\mathrm{KL}}(P_{\omega,s,p,a} \| P_{\rho,s,p,a})\right\}, \tag{90}$$

where ρ is a reference belief or reference model and $\mathfrak{C}_t(s)$ is a localized comparison set that contains the fixed truth on the calibration event. The notation is intentionally aligned with the log-penalized Bellman program. As Lemma C.1 illustrates, the E2D objective can be recovered from the one-step MAIR bracket: it uses the current feasible action set and localized comparison set, but replaces the dynamic continuation value by a one-step separation penalty. The same one-round minimax principle applies to the original Hellinger formulation and to constrained variants (Foster et al., 2023).

E2D packages immediate regret and statistical separation into a statewise one-step optimization. It controls the full fixed-truth bracket only after its localized comparison set, reference law, and separation term have been shown to dominate the chosen continuation relaxation. Under posterior averaging the KL term may become mutual information; a frequentist argument instead needs calibration and a compatible fixed-truth sequential KL bound.

D.3 AMS/EBO: robust convex belief optimization

The Exploration by Optimization (EBO) algorithm (Lattimore and György, 2021; Foster et al., 2022) and Adaptive Minimax Sampling (AMS) (Xu and Zeevi, 2025; Liu et al., 2025, 2026) lead to closely related robust belief optimizations of the log-penalized Bellman bracket. Their robust comparison belief is a different variable from the algorithmic update control in (32); KL duality relates the objectives. Work with an action or policy index $Y = \chi(\Omega)$. For clarity, this subsection states the coordinate argument for finite $\mathcal{Y}$; a dominated extension requires explicit positivity, absolute-continuity, and differentiability hypotheses. At state s, let $q \in \Delta(\mathcal{Y})$ be positive on every coordinate evaluated by the logarithmic score and let $\mathfrak{B}_t(s)$ be a calibrated convex set of candidate pair beliefs $\tilde{\nu}$ on (ω, y), supported on $y = \chi(\omega)$. The candidate belief is controlled by the inner maximization; after the maximizer $\bar{\nu}_t$ is selected, its posterior index update controls the fixed-truth AIR bracket. The one-step robust AIR/EBO objective is the indexed AIR functional from (11), equivalently

$$\mathfrak{A}_{q,\gamma}(s,p,\nu) = \ell_\nu(s,p) - \gamma \mathcal{I}_\chi(s,p;\nu) - \gamma D_{\mathrm{KL}}(\nu_\chi \| q), \tag{91}$$

where ν_χ is the index marginal, $\ell_\nu(s,p) = \mathbb{E}_{(\omega,y)\sim\nu}\ell_\omega(s,p)$, and $\mathcal{I}_\chi(s,p;\nu) = \mathbb{E}_{a\sim p} I_\nu(Y;O \mid s,p,a)$. The robust maximization over beliefs is essential:

$$p_t \in \operatorname*{arg\,min}_{p\in\Delta(\mathcal{A}_t(s))} \max_{\tilde\nu\in\mathfrak{B}_t(s)} \mathfrak{A}_{q_t,\gamma}(s,p,\tilde\nu), \qquad \bar\nu_t \in \operatorname*{arg\,max}_{\tilde\nu\in\mathfrak{B}_t(s)} \mathfrak{A}_{q_t,\gamma}(s,p_t,\tilde\nu). \tag{92}$$

Without the inner $\max_{\tilde\nu}$, the display is only a posterior AIR calculation for a chosen belief, not a robust AMS/EBO rule.

The convex-analytic interpretation is the following. The fixed-truth proof uses the coordinate log loss $\gamma\log(1/q_t(y^\star))$. Averaging this coordinate under a candidate belief gives $\gamma H(\nu_\chi)+\gamma D_{\mathrm{KL}}(\nu_\chi\|q_t)$. When the next reference marginal is the posterior coordinate induced by ν, the expected change of this averaged potential is $-\gamma\mathcal{I}_\chi(s,p;\nu) - \gamma D_{\mathrm{KL}}(\nu_\chi\|q_t)$. Equivalently, optimizing over possible next index laws yields the log-partition dual (30). Hence the EBO potential is the KL-dual log partition, while the fixed-truth AIR argument uses the coordinate posterior log loss; they coincide only through KL duality and the chosen update, not as identical scalar functions.

For the AIR functional in (91), concavity in $\tilde\nu$ follows from the posterior-scoring representation in Appendix C.1: it is the infimum over prediction rules Q of an objective affine in $\tilde\nu$, provided the prediction class and support convention are fixed across beliefs. Thus the concavity hypothesis in Lemma D.2 is automatic in this setting. For modified belief relaxations or restricted prediction classes, concavity must be verified directly or supplied by a valid concave upper bound.

Lemma D.2 (AMS/EBO saddle control of the AIR fixed-truth bracket)**.** *Assume the finite or dominated differentiability setting of Lemma 3.3. Fix $(\omega^\star,y^\star)$ with $y^\star = \chi(\omega^\star)$. Suppose $\tilde\nu \mapsto \mathfrak{A}_{q_t,\gamma}(s,p_t,\tilde\nu)$ is concave on the convex set $\mathfrak{B}_t(s)$, $\bar\nu_t$ maximizes it as in (92), and the comparison direction $\delta_{(\omega^\star,y^\star)} - \bar\nu_t$ is admissible for $\mathfrak{B}_t(s)$. If*

$$\max_{\tilde\nu\in\mathfrak{B}_t(s)} \mathfrak{A}_{q_t,\gamma}(s,p_t,\tilde\nu) \le \varepsilon_t,$$

then the fixed-truth AIR bracket with coefficient γ satisfies

$$\ell_{\omega^\star}(s,p_t) - \gamma\mathbb{E}_{a\sim p_t, O\sim P_{\omega^\star,t}(\cdot|s,p_t,a)} \log \frac{\bar\nu_t(Y=y^\star \mid s,p_t,a,O)}{q_t(y^\star)} \le \varepsilon_t. \tag{93}$$

If the reference update is $q_{t+1}(\cdot) = \bar\nu_t(Y\in\cdot \mid s,p_t,A_t,O_t)$ and remains positive at $y^\star$, and if all preceding hypotheses hold almost surely along the trajectory, then Theorem 4.1 with $V_t\equiv 0$ gives

$$\mathbb{E}_{\omega^\star} L_{\omega^\star}(H_T) \le \gamma\log\frac{1}{q_1(y^\star)} + \mathbb{E}_{\omega^\star}\sum_{t=1}^{T}\varepsilon_t \tag{94}$$

for finite or countable index sets. If the comparison direction is only approximately admissible, its effect must be bounded explicitly in the one-step bracket and added to ε_t; a future-dependent containment event cannot be appended after the telescope. Averaging this coordinate bound under a Bayesian prior replaces $\log(1/q_1(Y))$ by the corresponding cross-entropy or by $H_\mu(Y)$ when $q_1=\mu_\chi$.

Proof. See Appendix H.3.3.

D.4 DEC and single-round information complexity

Offset DEC gives a powerful single-round specialization of the fixed-truth log-potential bracket (25), as illustrated in Lemma C.1. It retains the immediate decision loss and represents statistical

separation by a one-step penalty relative to a reference predictive law, in place of the dynamic continuation term.

In the model-index case $\chi(\omega) = \omega$, let $\mathfrak{C}_t(s)$ be the local comparison set at the current time-state pair and let $\bar{\nu}$ be a reference belief at state s. We write

$$P_{\bar{\nu},s,a} := \int P_{\omega,s,a}\, \bar{\nu}(d\omega)$$

for its predictive law. The corresponding pointwise KL-DEC offset, evaluated at each state and belief, is

$$\mathrm{dec}^{\mathrm{KL}}_{\gamma}(s;\bar{\nu}) := \inf_{p\in\Delta(\mathcal{A}_t(s))} \sup_{\omega\in\mathfrak{C}_t(s)} \left\{\mathbb{E}_{a\sim p}\ell_\omega(s,a) - \gamma\mathbb{E}_{a\sim p} D_{\mathrm{KL}}(P_{\omega,s,a}\,\|\,P_{\bar{\nu},s,a})\right\}. \tag{one-step model-index offset}$$

Thus the usual offset DEC may be viewed as a one-step specialization of the Bellman program with a fixed reference law: the learner chooses the one-step decision distribution p, and the local adversary then chooses the hardest model relative to the fixed reference predictive law.

For this comparison, a single-round DEC coefficient is obtained by fixing a reference model or predictive law and decomposing trajectory divergence into epochwise terms. The quantile PAC–DEC of Chen et al. (2024, Section 3.2.2) gives the corresponding lower-bound reduction. DEC is sharp when the localized reference geometry is captured by the one-round game; the Bellman formulation retains the evolving reference trajectory when that localization changes with the state. They agree when the single-round coefficient controls the same comparison class and information budget as the dynamic recursion.

Localization in RKHS classes. For an infinite RKHS ball, a useful single-round DEC comparison is ordinarily localized rather than taken against every orthogonal direction at once. Constrained and localized DEC theory provides several principled ways to do this, including spectral truncation, finite active support, posterior localization, and calibrated confidence restrictions (Foster et al., 2023). These mechanisms can turn the ambient RKHS into an effective local comparison class on which the one-round game is sharp. The present paper pursues a complementary dynamic route: it keeps the evolving reference state explicit and, in the kernel construction, fixes a finite marginal before applying the robust action-index AIR/AMS saddle.

E Kernel-bandit details

Proposition E.1 (Exact Hilbert-space linear representation)**.** *Let k be a positive-semidefinite kernel on $\mathcal{X}$, let $\mathcal{H}_k$ be its RKHS, and define the canonical feature map*

$$\varphi(x) := k(x,\cdot) \in \mathcal{H}_k.$$

Then

$$f(x) = \langle f, \varphi(x)\rangle_{\mathcal{H}_k}, \qquad \|\varphi(x)\|^2_{\mathcal{H}_k} = k(x,x).$$

Consequently, the RKHS bandit with action set $\mathcal{X}$, parameter class $\{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le B\}$, and observations

$$Y_t = f(X_t) + \varepsilon_t$$

is equivalent to the Hilbert-space stochastic linear bandit with action set $\mathcal{A}_k := \{\varphi(x) : x \in \mathcal{X}\} \subset \mathcal{H}_k$, parameter $\theta = f$, and observations

$$Y_t = \langle \theta, \varphi(X_t)\rangle_{\mathcal{H}_k} + \varepsilon_t.$$

After actions with identical feature vectors are identified and tie-breaking is chosen compatibly, the two descriptions generate identical histories and regret, path by path. If $\operatorname{span}\varphi(\mathcal{X})$ *has finite dimension* d*, any isometry from this span to* $\mathbb{R}^d$ *gives an ordinary* d*-dimensional stochastic linear bandit. More explicitly, one replaces* f *by its orthogonal projection onto this span, which leaves every reward unchanged and cannot increase its norm; conversely, every vector of norm at most* B *in the span is itself an RKHS element. Hence the induced parameter class is the full radius-*B *ball in the feature span. For a finite action set* $\mathcal{X}$*, this dimension is* $d = \operatorname{rank}(K_{\mathcal{X}}) \le |\mathcal{X}|$*.*

The algorithmic correspondence also holds pathwise. For ridge parameter and Gaussian noise variance $\lambda > 0$*, let*

$$V_t := \lambda I_{\mathcal{H}_k} + \sum_{r<t} \varphi(X_r) \otimes \varphi(X_r), \qquad \widehat{\theta}_t := V_t^{-1} \sum_{r<t} Y_r \varphi(X_r).$$

The zero-mean kernel-ridge/GP posterior quantities satisfy

$$m_t(x) = \langle \widehat{\theta}_t, \varphi(x) \rangle_{\mathcal{H}_k}, \qquad s_t^2(x) = \lambda \langle \varphi(x), V_t^{-1} \varphi(x) \rangle_{\mathcal{H}_k}.$$

Thus GP–UCB with score $m_t(x) + \beta_t s_t(x)$ *chooses the same action as ridge LinUCB with score*

$$\langle \widehat{\theta}_t, u \rangle_{\mathcal{H}_k} + \sqrt{\lambda}\, \beta_t \|u\|_{V_t^{-1}}, \qquad u = \varphi(x).$$

At the unit ridge and noise normalization used in the separation, the two acquisition rules have the same multiplier.

Proof. The first identities are the reproducing property. For the posterior identities, let $\Phi_t : \mathcal{H}_k \to \mathbb{R}^{t-1}$ be the sampling operator $(\Phi_t h)_r = \langle h, \varphi(X_r) \rangle_{\mathcal{H}_k}$. Then $K_t = \Phi_t \Phi_t^*$ and $V_t = \lambda I + \Phi_t^* \Phi_t$. The push-through and Woodbury identities give

$$V_t^{-1} \Phi_t^* = \Phi_t^* (K_t + \lambda I)^{-1}$$

and

$$I - \Phi_t^* (K_t + \lambda I)^{-1} \Phi_t = \lambda V_t^{-1}.$$

Substitution into the standard GP posterior mean and covariance formulas proves the claim. □

The infinite-dimensional statement is an operator identity between kernel-ridge and Hilbert-space ridge formulas. It does not identify the GP prior with an isotropic Gaussian random element of $\mathcal{H}_k$: such a random element generally does not exist in infinite dimension, and GP sample paths generally need not belong to the RKHS. Nor does a T-round experiment automatically reduce the full action set to dimension at most T. The realized design update has rank at most T, but GP–UCB/LinUCB still maximizes over unsampled feature directions; those unsampled directions drive the small-cloud lower bound.

E.1 Finite-marginal indexed Bellman recursion

On the finite active marginal, the frequentist log-potential upper benchmark is the Bellman program of Theorem 4.2. Its action-index branch is AIR, while its model-index branch is MAIR. The program is relative to the displayed state, comparison sets, and reference update, and need not equal the unrestricted minimax value. The state contains (m_t, Σ_t, q_t), where (m_t, Σ_t) is the algorithmic GP marginal on F and q_t is the maintained reference law of the index, either $A^\star(F)$ for AIR or F for MAIR. Let $\mathfrak{C}_t(m, \Sigma, q)$ be a surely calibrated local set of candidate active reward vectors containing

the fixed truth $F^\star$ at every reached state. A local control $u \in \mathfrak{U}_t(m, \Sigma, q)$ specifies an action law p_u and a reference update. In AIR form the update is generated by a pair belief ν_u on active reward vectors and their optimal-action indices, so that $q^+ = q^+_{\nu_u}$. For the robust saddle calculation, candidate pair beliefs may be supported on the closed optimal relation $\{(f, y) : y \in \arg\max_x f(x)\}$; the true index still uses the prescribed tie-breaking rule $A^\star(f)$. In the action-index version define

$$y(f) = A^\star(f), \qquad \mathsf{G}_t^{\mathrm{AIR}}(u; f) := \mathbb{E}_{x\sim p_u,\, O\sim N(f_x,\lambda)} \log \frac{q^+(y(f) \mid m, \Sigma, q, x, O)}{q(y(f))},$$

where q^+ is the chosen reference update for the optimal-action marginal. For a candidate reward vector f, write

$$\Delta_f(p) := \max_{x\in\mathcal{X}} f(x) - \mathbb{E}_{x\sim p} f(x).$$

The robust finite-marginal Bellman recursion is

$$W_t^\gamma(m, \Sigma, q) = \inf_{u\in\mathfrak{U}_t(m,\Sigma,q)} \sup_{f\in\mathfrak{C}_t(m,\Sigma,q)} \left\{\Delta_f(p_u) + \mathbb{E}_f W_{t+1}^\gamma(m^+, \Sigma^+, q^+) - \gamma \mathsf{G}_t^{\mathrm{AIR}}(u; f)\right\}.$$

The model-index MAIR version replaces $y(f)$ by the finite reward vector f and uses the corresponding model-coordinate posterior log gain. For the finite action index, Theorem 4.2 implies that an approximate minimizer of this recursion has regret at most

$$W_1^\gamma(m_1, \Sigma_1, q_1) + \gamma \log \frac{1}{q_1(A^\star(F^\star))} + \text{explicit bracket/optimization errors.}$$

For the continuous model index $F \in \mathbb{R}^n$, one cannot drop the terminal log-density term by treating it as a discrete code length: a posterior density may exceed one. A valid continuous-index statement must retain that terminal term or replace point coordinates by neighborhoods, a finite cover, or a PAC–Bayes comparison. The finite action index $A^\star$ avoids this issue and is the clean frequentist specialization used in the main text.

If the maintained belief is the exact Bayesian posterior and one averages the coordinate recursion over F, the robust supremum is replaced by posterior expectation and the dynamic program reduces to

$$U_t^\gamma(m, \Sigma, q) = \inf_{p\in\Delta(\mathcal{X})} \left\{\Delta_t^{\mathrm{AIR}}(p) + \mathbb{E} U_{t+1}^\gamma(m^+, \Sigma^+, q^+) - \gamma \mathcal{I}_t^{\mathrm{AIR}}(p)\right\}.$$

Thus the frequentist recursion uses the coordinate log potential and calibrated candidate truths, while the Bayesian recursion is its posterior average. Appendix E.5 gives further specialization of thr robust action-index AIR/AMS/EBO control.

E.2 Finite-scale small-cloud construction

We will show that maximal information gain can be more than a loose final step in the analysis: when it is used to calibrate the acquisition rule, it can create a self-fulfilling exploration cost. The example is a sequential, small-amplitude version of the hub–cloud construction of Xu (2026). Unlike the direct large-cloud embedding of that construction, the cloud below does not make the full class linearly hard. Its height is chosen so that an unrestricted minimax algorithm may ignore it at cost aT, while its multiplicity still inflates the maximal-information confidence multiplier.

Fix a horizon T, an integer $N \geq 1$, and $a \in (0, 1]$. Let

$$\mathcal{X}_T = \{x_0, x_h, c_1, \ldots, c_N\}$$

and define the diagonal kernel

$$k_T(x_0, x) = 0, \qquad k_T(x_h, x_h) = 1, \qquad k_T(c_i, c_j) = a^2 \mathbf{1}\{i = j\}, \tag{95}$$

with all remaining covariances zero. Thus $k_T(x, x) \le 1$. Its unit RKHS ball is

$$\mathcal{F}_T = \left\{ f : f(x_0) = 0, \quad f(x_h)^2 + a^{-2} \sum_{j=1}^{N} f(c_j)^2 \le 1 \right\}. \tag{96}$$

Observations are $Y_t = f(X_t) + \xi_t$ with $\xi_t \overset{\text{i.i.d.}}{\sim} N(0, 1)$.

This kernel has the explicit feature map

$$\varphi(x_0) = 0, \qquad \varphi(x_h) = e_0, \qquad \varphi(c_j) = a e_j, \quad j \in [N], \tag{97}$$

in $\mathbb{R}^{N+1}$. Hence $k_T(x, x') = \langle \varphi(x), \varphi(x') \rangle$. Moreover, setting

$$\theta_0 = f(x_h), \qquad \theta_j = a^{-1} f(c_j), \quad j \in [N],$$

gives

$$f(x) = \langle \theta, \varphi(x) \rangle, \qquad \|\theta\|_2^2 = f(x_h)^2 + a^{-2} \sum_{j=1}^{N} f(c_j)^2.$$

Thus $\mathcal{F}_T$ is the full Euclidean unit parameter ball for the displayed action set, not merely a subset of a linear model. This is the direct finite-dimensional reduction. The later ℓ_2 construction is needed only to glue these finite blocks into a single horizon-independent kernel.

Let

$$\Gamma_s(T) := \sup_{x_{1:s} \in \mathcal{X}_T^s} \frac{1}{2} \log \det(I + K_{x_{1:s}, x_{1:s}})$$

be the maximal information gain at noise variance one. Consider the following horizon-aware maximal-information calibration

$$\beta_t^{\max} := 1 + \sqrt{2\{\Gamma_{t-1}(T) + 2 \log T\}}, \tag{98}$$

and the zero-mean, unit-ridge GP–UCB rule

$$X_t \in \arg\max_{x \in \mathcal{X}_T} \{ m_t(x) + \beta_t^{\max} s_t(x) \}. \tag{99}$$

The conclusion is insensitive to tie-breaking.

Equation (99) is the finite-horizon prototype of the anytime rule (66): it uses the same posterior-UCB acquisition map and maximal-information mechanism, with the horizon-dependent kernel and logarithmic calibration made explicit. It is not the original schedule (67) of Srinivas et al. (2010). That schedule uses distinct-set information gain and an additional $\log^{3/2}(t/\delta)$ factor. The fixed-kernel proof treats the two calibrations separately.

Theorem E.2 (Small-cloud separation)**.** *Fix $0 < \alpha < 1/4$ and put $a = T^{-\alpha}$. There are constants $0 < c_\alpha < C_\alpha < \infty$ such that, for all sufficiently large T and every $N \ge 4T$,*

$$c_\alpha a T \le \inf_{\mathsf{Alg}} \sup_{f \in \mathcal{F}_T} \mathbb{E}_f \operatorname{Reg}_T(\mathsf{Alg}) \le C_\alpha a T. \tag{100}$$

In contrast, the GP–UCB rule (99) satisfies

$$\sup_{f\in\mathcal{F}_T} \mathbb{E}_f \operatorname{Reg}_T(\text{GP–UCB}) \geq c_\alpha T. \tag{101}$$

Consequently,

$$\frac{\sup_{f\in\mathcal{F}_T} \mathbb{E}_f \operatorname{Reg}_T(\text{GP–UCB})}{\inf_{\mathsf{Alg}} \sup_{f\in\mathcal{F}_T} \mathbb{E}_f \operatorname{Reg}_T(\mathsf{Alg})} \geq c_\alpha T^\alpha. \tag{102}$$

Thus maximal-information-calibrated GP–UCB is polynomially minimax-suboptimal on a family of bounded finite kernels, even though the comparison is made over the entire unit RKHS ball.

Proof. We prove the minimax lower bound, the minimax upper bound, and the GP–UCB lower bound separately.

Minimax lower bound. For $j \in [N]$, let $f^j(c_j) = a$ and let f^j vanish on every other action. By (96), $f^j \in \mathcal{F}_T$. Fix an arbitrary possibly randomized nonanticipating algorithm, let U denote its independent random seed, draw J uniformly from $[N]$, and pre-generate independent Gaussian noise stacks at every action. Couple the experiments under f^J and $f \equiv 0$ using the same U and noise stacks, and let $\tau_J = \inf\{t \geq 1 : X_t = c_J\}$. The two histories agree strictly before τ_J, so τ_J is identical in the coupled experiments. Under the null experiment, if S_T^0 is the set of distinct cloud actions queried by time T, then J is independent of S_T^0 and

$$\mathbb{P}(\tau_J \leq T) = \mathbb{P}(J \in S_T^0) = \frac{\mathbb{E}|S_T^0|}{N} \leq \frac{T}{N} \leq \frac{1}{4}.$$

On $\{\tau_J > T\}$ every selected action has mean zero under f^J, whereas the optimal mean is a. Thus the Bayes regret is at least $(1 - T/N)aT \geq 3aT/4$. This proves the first inequality in (100).

Minimax upper bound. Ignore the cloud and run a horizon-T minimax policy for two unit-sub-Gaussian arms on $\{x_0, x_h\}$; a sub-Gaussian version of MOSS is one example (Audibert and Bubeck, 2009; Lattimore and Szepesvári, 2020). Put

$$g_f = \max\{0, f(x_h)\}, \qquad f^\star_{\max} = \max\{g_f, \max_{j\leq N} f(c_j)\}.$$

Since $|f(x_h)| \leq 1$, $\max_j f(c_j) \leq a$, and $f(x_0) = 0$, we have $0 \leq f^\star_{\max} - g_f \leq a$. The restricted policy has expected regret at most $C\sqrt{T}$ relative to g_f, so its regret relative to the full set is at most

$$C\sqrt{T} + aT \leq C_\alpha aT,$$

where the last inequality uses $\alpha < 1/2$. This proves the second inequality in (100).

GP–UCB lower bound. Take the hub truth

$$f^\dagger(x_h) = \Delta := \frac{1}{4}, \qquad f^\dagger(x_0) = f^\dagger(c_j) = 0, \quad j \leq N. \tag{103}$$

It belongs to $\mathcal{F}_T$. For every $s \leq T$, querying s distinct cloud actions is feasible, while for an arbitrary design the hub contributes at most $\frac{1}{2}\log(1+s)$ and the cloud contributes at most $a^2 s/2$. Since $\log(1+u) \geq u/2$ for $u \in [0,1]$,

$$\frac{sa^2}{4} \leq \Gamma_s(T) \leq \frac{1}{2}\log(1+s) + \frac{sa^2}{2}. \tag{104}$$

Indeed, if $n_h, n_1, \ldots, n_N$ are the design counts, then $n_h + \sum_j n_j \leq s$, pulls of x_0 contribute zero, and diagonalization gives

$$\frac{1}{2}\log(1+n_h) + \frac{1}{2}\sum_{j=1}^N \log(1+a^2 n_j) \leq \frac{1}{2}\log(1+s) + \frac{a^2 s}{2}.$$

An unqueried cloud action has posterior mean zero and posterior standard deviation a, independently of all observations at other actions. By the lower bound in (104), at every $t \geq T/2$ its acquisition score is at least

$$a\beta_t^{\max} \geq a^2\sqrt{\frac{t-1}{2}} \geq cT^{1/2-2\alpha}. \tag{105}$$

This diverges because $\alpha < 1/4$. The upper bound in (104) also gives, uniformly for $t \leq T$,

$$\beta_t^{\max} \leq C\{a\sqrt{T} + \sqrt{\log T}\}. \tag{106}$$

Realize observations through the independent action-specific noise stacks: the rth pull of x reveals $f(x) + \xi_{x,r}$. Let $\xi_{h,r} = \xi_{x_h,r}$ and define

$$\mathcal{E}_T := \left\{ \sum_{r=1}^{n} \xi_{h,r} \leq 2\sqrt{n \log T} \text{ for every } 1 \leq n \leq T \right\}.$$

A Gaussian tail bound and a union bound give

$$\mathbb{P}(\mathcal{E}_T^c) \leq \sum_{n=1}^{T} \exp\left\{ -\frac{(2\sqrt{n \log T})^2}{2n} \right\} \leq T^{-1}.$$

After n hub pulls, diagonal conjugacy gives

$$m_t(x_h) = \frac{n\Delta + \sum_{r=1}^{n} \xi_{h,r}}{1+n}, \qquad s_t(x_h) = \frac{1}{\sqrt{1+n}}.$$

Suppose, toward a contradiction, that fewer than $T/4$ cloud pulls occur by time T. The anchor x_0 is never selected: its acquisition score is zero, while an unqueried cloud has strictly positive score. Hence, for every $t \in [\lceil T/2 \rceil, T]$, the number n of preceding hub pulls is at least $T/4 - 1$. On $\mathcal{E}_T$, (106) yields

$$\begin{aligned} m_t(x_h) + \beta_t^{\max} s_t(x_h) &\leq \Delta + C\sqrt{\frac{\log T}{T}} + C\left(a + \sqrt{\frac{\log T}{T}}\right) \\ &\leq \frac{1}{2} \end{aligned}$$

for all sufficiently large T. On the other hand, (105) is then larger than one. Since $N \geq 4T$, an unqueried cloud is always available. GP–UCB must therefore select a cloud action on every round in the last half of the horizon, contradicting the supposition that it makes fewer than $T/4$ cloud pulls. Thus, on $\mathcal{E}_T$, it makes at least $T/4$ cloud pulls. Each such pull has regret Δ, and consequently

$$\mathbb{E}_{f^\dagger} \operatorname{Reg}_T(\text{GP–UCB}) \geq \frac{\Delta T}{4}(1 - T^{-1}),$$

which proves (101). Combining this lower bound with the minimax upper bound in (100) proves (102). □

Corollary E.3 (Finite-horizon LinUCB separation)**.** *Fix $0 < \alpha < 1/4$. For every sufficiently large integer T, let*

$$N = T - 1, \qquad d = N + 1 = T, \qquad a = T^{-\alpha},$$

and consider the d-dimensional action set

$$\mathcal{A}_T = \{0, e_0, ae_1, \dots, ae_N\} \subset \mathbb{R}^d$$

and parameter class $\Theta_T = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}$. *Observations are*

$$Y_t = \langle \theta, A_t \rangle + \xi_t, \qquad \xi_t \overset{\text{i.i.d.}}{\sim} N(0,1).$$

Thus $|\mathcal{A}_T| = d + 1 = T + 1$, *while* $\|A\|_2 \le 1$ *and* $\|\theta\|_2 \le 1$.

Let

$$V_t = I_d + \sum_{r<t} A_r A_r^\top, \qquad \widehat{\theta}_t = V_t^{-1} \sum_{r<t} A_r Y_r,$$

and define the maximal linear information gain

$$\Gamma_s^{\text{lin}}(\mathcal{A}_T) := \sup_{u_1,\dots,u_s \in \mathcal{A}_T} \frac{1}{2} \log\det\left(I_d + \sum_{r=1}^{s} u_r u_r^\top \right).$$

Consider the maximal-information-calibrated ridge LinUCB rule

$$A_t \in \arg\max_{u \in \mathcal{A}_T} \left\{ \langle \widehat{\theta}_t, u \rangle + \left[1 + \sqrt{2\{\Gamma_{t-1}^{\text{lin}}(\mathcal{A}_T) + 2\log T\}} \right] \|u\|_{V_t^{-1}} \right\}. \tag{107}$$

With

$$\text{Reg}_T(\pi) := \sum_{t=1}^{T} \left\{ \max_{u \in \mathcal{A}_T} \langle \theta, u \rangle - \langle \theta, A_t \rangle \right\},$$

we have

$$\inf_{\pi} \sup_{\theta \in \Theta_T} \mathbb{E}_\theta \,\text{Reg}_T(\pi) = \Theta(T^{1-\alpha}),$$

whereas

$$\sup_{\theta \in \Theta_T} \mathbb{E}_\theta \,\text{Reg}_T(\text{LinUCB}_{\max}) = \Omega(T).$$

Consequently, the worst-case regret of this LinUCB rule divided by the minimax value on the same action set and parameter class is $\Omega(T^\alpha)$. *The same* $\Omega(T)$ *conclusion holds for the finite-action specializations of the anytime multiplier* (66) *and, for every fixed* $\delta \in (0,1)$, *the original-schedule multiplier* (67), *for all sufficiently large* T, *with the threshold allowed to depend on* α *and* δ.

The proof follows after Remark E.4.

Remark E.4 (Scope relative to linear-bandit literature)**.** *Corollary E.3 is nonasymptotic and action-set-specific: for every sufficiently large finite horizon, it supplies one finite action set and compares the specified LinUCB rule with the unrestricted minimax value on that same set and the full unit parameter ball. The action set depends on* T, *its dimension is* $d = T$, *and its geometry is heterogeneous: the hub has norm one, the cloud actions have norm* $T^{-\alpha}$, *and the anchor is zero. The result therefore does not give a polynomial separation in a fixed finite dimension or when* $d = o(T)$. *Proving a comparable separation for a more slowly growing dimension, or for a uniformly normalized and well-conditioned action geometry, remains open.*

Earlier sharp contextual-bandit analyses used the layered SupLinRel and SupLinUCB constructions to recover the independence needed in their proofs (Auer, 2002; Chu et al., 2011). Those works established strong upper bounds but did not prove that the corresponding one-layer rules were

minimax-suboptimal. Lattimore and Szepesvári (2017) proved that a broad class of optimistic policies can be arbitrarily worse in the asymptotic instance-dependent constant on a fixed finite action set. More recently, Song and Zhou (2025) proved a finite-time gap for LinUCB in an i.i.d.-context, arm-specific-parameter model. These results address different criteria or observation structures. To the best of our knowledge, Corollary E.3 is the first polynomial-in-T, finite-horizon separation in a finite-dimensional, finite-action problem between the worst-case cumulative regret of the specified maximal-information-calibrated LinUCB rule and the minimax regret on the same noncontextual, shared-parameter stochastic linear-bandit class. It is not a lower bound for every rule called LinUCB, OFUL, or kernel UCB. By contrast, the standard OFUL confidence radius is based on the realized design determinant (Abbasi-Yadkori et al., 2011); the rule in the corollary uses the global maximal-information envelope, so the corollary does not automatically apply to OFUL. Proposition E.1 also shows that the horizonwise Matérn lower bound of Wang and Zhang (2026) has an infinite-dimensional Hilbert-linear interpretation; the finite-dimensional conclusion here is distinct.

The comparison with Maiti et al. (2026) is complementary rather than nested. They study fixed-confidence ε-best-arm identification and construct a union of k orthogonal r-dimensional balls in $\mathbb{R}^{kr}$. For $r \le k \le r^2$, their construction gives a nonadaptive lower bound

$$\Omega\left(\frac{kr\log(1/\delta) + kr^2}{\varepsilon^2}\right),$$

while an adaptive procedure uses

$$O\left(\frac{kr\log(1/\delta) + kr\log k}{\varepsilon^2}\right)$$

samples. Their lower bound applies to every nonadaptive fixed design and hence exhibits an adaptivity advantage; when $\log(1/\delta)$ does not dominate r, the ratio is of order $r/\log k$. The present result instead concerns cumulative regret and separates one fully adaptive optimistic rule from the unrestricted minimax benchmark. Their statement is broader in algorithm class but concerns terminal identification; ours is algorithm-specific but concerns online regret. Neither theorem implies the other.

Proof of Corollary E.3. Set

$$H := \left\lfloor \frac{T-1}{4} \right\rfloor.$$

Then $N = T - 1 \ge 4H$ and $H \asymp T$. Take the feature map in (97) with this value of N. The coordinate map is an isometry from the RKHS of k_T to $\mathbb{R}^{N+1} = \mathbb{R}^T$, so the coordinate map sends the unit RKHS ball onto the Euclidean unit parameter ball, and rewards, histories, and regret are preserved path by path. For every design sequence, Sylvester's determinant identity gives

$$\det(I_s + K_{x_{1:s},x_{1:s}}) = \det\left(I_d + \sum_{r=1}^{s} \varphi(x_r)\varphi(x_r)^\top\right).$$

Proposition E.1, with $\lambda = 1$, also gives

$$m_t(x) = \langle \widehat{\theta}_t, \varphi(x)\rangle, \qquad s_t^2(x) = \varphi(x)^\top V_t^{-1}\varphi(x).$$

Thus (107) coincides with the GP–UCB rule with the same maximal-information multiplier.

We first prove the minimax rate. For each $j \in [N]$, take the unit parameter $\theta^j = e_j$, for which the unique positive cloud reward is $\langle \theta^j, ae_j\rangle = a$. Draw J uniformly from $[N]$ and couple the

experiments under θ^J and $\theta = 0$ with the same algorithmic seed and action-specific Gaussian noise stacks. If τ_J is the first pull of ae_J, the histories agree before τ_J. Under the null experiment, the first H rounds visit at most H distinct cloud actions, and hence

$$\mathbb{P}(\tau_J \le H) \le \frac{H}{N} \le \frac{1}{4}.$$

On $\{\tau_J > H\}$, every one of the first H pulls has regret a. Therefore every algorithm has Bayes, and hence worst-case, regret at least

$$\left(1 - \frac{H}{N}\right) aH \ge \frac{3}{4} aH = \Omega(T^{1-\alpha}).$$

Conversely, ignoring the cloud and using a two-armed minimax policy on $\{0, e_0\}$ incurs $O(\sqrt{T})$ regret relative to the better retained action and at most aT additional regret relative to the full action set. This is $O(T^{1-\alpha})$.

For the LinUCB lower bound, take the hub truth $\theta^\dagger = \Delta e_0$, $\Delta = 1/4$. For every $s \le H$, the availability of s distinct cloud actions and the same diagonal calculation as in (104) give

$$\frac{sa^2}{4} \le \Gamma_s^{\mathrm{lin}}(\mathcal{A}_T) \le \frac{1}{2}\log(1+s) + \frac{sa^2}{2}.$$

Thus, throughout $H/2 \le t \le H$, an unqueried cloud action has score at least $ca^2\sqrt{H}$, which tends to infinity because $\alpha < 1/4$, while the multiplier in (107) is at most $C(a\sqrt{H} + \sqrt{\log T})$.

Let

$$\mathcal{E}_H := \left\{ \sum_{r=1}^{n} \xi_{0,r} \le 2\sqrt{n \log T} \text{ for every } 1 \le n \le H \right\},$$

where $\xi_{0,r}$ is the rth noise observed at the hub. A union bound gives $\mathbb{P}(\mathcal{E}_H^c) \le H/T^2 \le T^{-1}$. Suppose that fewer than $H/4$ cloud pulls occur by time H. The zero action is never selected because an unqueried cloud action has positive score. Hence, before every round $t \in [\lceil H/2 \rceil, H]$, the hub has been pulled at least $H/4 - 1$ times. On $\mathcal{E}_H$, its score is at most

$$\Delta + C\sqrt{\frac{\log T}{H}} + C\left(a + \sqrt{\frac{\log T}{H}}\right) < \frac{1}{2}$$

for all sufficiently large T, while an unqueried cloud score is larger than one. Since $N \ge 4H$, such a cloud action is always available under the supposition. Every round in the last half of the prefix would therefore be a cloud pull, a contradiction. The rule consequently makes at least $H/4$ cloud pulls on $\mathcal{E}_H$, and

$$\mathbb{E}_{\theta^\dagger} \mathrm{Reg}_T \ge \frac{\Delta H}{4}(1 - T^{-1}) = \Omega(T).$$

The anytime multiplier is handled by the same prefix argument, since $\log(t \vee 2) \asymp \log T$ for $H/2 \le t \le H$. For the original schedule and any fixed $\delta \in (0,1)$, its distinct-set information gain satisfies, for $s \le H$,

$$\frac{sa^2}{4} \le \Gamma_s^{\mathrm{SKKS}}(k_T) \le \frac{1}{2}\log 2 + \frac{sa^2}{2},$$

because $N \ge 4H$. Uniformly on the last half of the prefix, an unqueried cloud score is therefore at least $ca^2\sqrt{H}\log^{3/2} H \to \infty$, whereas after $H/4 - 1$ hub pulls the hub bonus is

$$O\Big(H^{-1/2}\log^{3/2} T + a\log^{3/2} T\Big) = o(1).$$

The same event and contradiction force $\Omega(H) = \Omega(T)$ cloud pulls. □

E.3 Interpretation and scope of the fixed-kernel separation

Remark E.5 (Scope of the theorem)**.** *The kernel is spatially inhomogeneous and its compact domain is countable; the theorem therefore does not settle the corresponding question for translation-invariant kernels on Euclidean domains. Nor does it rule out localized, predictable, or data-dependent exploration multipliers.*

Remark E.6 (AIR Bellman, AMS/EBO, and GP–UCB)**.** *On each predetermined marginal $\mathcal{R}_m$, the robust AIR/AMS saddle supplies the statewise control used by the AIR Bellman supersolution; EBO is related through the stated KL duality, not implemented as a separate functional-estimator algorithm. The comparison with GP–UCB uses the same fixed kernel, RKHS ball, domain, and observation model. The AIR/AMS policy's retained sets and epoch schedule are fixed in advance from public class geometry, whereas GP–UCB acts on the full domain with a global maximal-information multiplier. The theorem does not show that AIR/AMS learns the truncation, that a nonzero future-value term is necessary, or that a GP–UCB rule localized to the same marginal must fail.*

E.4 Horizonwise Matérn lower bounds and effective optimism

The fixed-kernel theorem gives a spatially inhomogeneous counterexample. A complementary result is available for Matérn RKHSs. To align notation, write the GP–UCB acquisition as

$$x_t \in \arg\max_x \{\mu_{t-1}(x;\rho_t) + b_t \sigma_{t-1}(x;\rho_t)\}, \qquad L_t := \rho_t b_t^2.$$

The quantity L_t is the effective optimism of Wang and Zhang (2026); their paper writes $b_t = \sqrt{\beta_t}$.

Corollary E.7 (Polynomial effective optimism is minimax-suboptimal)**.** *Fix d, ν, the Matérn length scale, the RKHS radius $B > 0$, and the Gaussian noise level $\sigma > 0$. Consider the radius-B Matérn-ν RKHS ball on $[0,1]^d$. Assume the conditions of Wang and Zhang (2026, Theorem 1): b_t^2 and ρ_t are positive and nondecreasing, $b_t^2 \geq 1$, and, for every $3 \leq t \leq T$ with constants independent of T,*

$$b_t^2 \asymp t^{\theta_1}(\log t)^{\theta_3}, \qquad \rho_t \asymp t^{\theta_2}(\log t)^{\theta_4}, \qquad \theta_1, \theta_2, \theta_3, \theta_4 \geq 0, \qquad \frac{4\nu + d}{2\nu}\theta_1 + \theta_2 < 1,$$

together with their simultaneous uniform-confidence assumption: for constants $\zeta \in (0,1)$ and $p_0 > 0$ independent of T, the event

$$|\mu_{t-1}(x;\rho_t) - f(x)| \leq \zeta b_t \sigma_{t-1}(x;\rho_t) \quad \text{for every } x \text{ and } 1 \leq t \leq T$$

has probability at least p_0. Then

$$\sup_{\|f\|_{\mathcal{H}_k} \leq B} \mathbb{E}_f \operatorname{Reg}_T(\text{GP–UCB}) = \Omega\left(T^{\frac{\nu+d}{2\nu+d}} L_T^{\frac{\nu}{2\nu+d}}\right). \tag{108}$$

The minimax value on the same standard Matérn class is, up to polylogarithmic factors,

$$\inf_{\mathsf{Alg}} \sup_{\|f\|_{\mathcal{H}_k} \leq B} \mathbb{E}_f \operatorname{Reg}_T(\mathsf{Alg}) = \widetilde{\Theta}\left(T^{\frac{\nu+d}{2\nu+d}}\right). \tag{109}$$

Consequently, any admissible schedule with $L_T = T^{\vartheta}\operatorname{polylog}(T)$, $\vartheta > 0$, is polynomially minimax-suboptimal by the factor

$$\widetilde{\Omega}\left(T^{\vartheta\nu/(2\nu+d)}\right).$$

Proof. See Appendix H.5.2.

The comparison with the Whitehouse–Wu–Ramdas effective-optimism scale is given below; the upper and lower statements concern different schedules and are not asserted to form a single matching theorem.

The Whitehouse effective-optimism scale. For comparison, Whitehouse et al. (2023, Corollary 2) prove that, for $\nu > 1/2$, their horizon-tuned GP–UCB construction with $\rho \asymp T^{d/(2\nu+2d)}$ satisfies

$$\sup_{\|f\|_{\mathcal{H}_k} \le B} \mathbb{E}_f \operatorname{Reg}_T = \widetilde{O}\Big(T^{\frac{\nu+2d}{2\nu+2d}}\Big). \tag{110}$$

Here their high-probability statement implies the expectation bound after taking the failure probability polynomially small and using the uniform RKHS envelope. Example 2 of Wang and Zhang (2026) identifies the corresponding Whitehouse-inspired effective-optimism scale. For every time-indexed schedule satisfying their assumptions and

$$L_T = \widetilde{\Theta}\Big(T^{\frac{d}{2\nu+2d}}\Big),$$

their Theorem 1 gives

$$\sup_{\|f\|_{\mathcal{H}_k} \le B} \mathbb{E}_f \operatorname{Reg}_T = \widetilde{\Omega}\Big(T^{\frac{\nu+2d}{2\nu+2d}}\Big), \tag{111}$$

because

$$\frac{\nu+d}{2\nu+d} + \frac{d}{2\nu+2d}\frac{\nu}{2\nu+d} = \frac{\nu+2d}{2\nu+2d}.$$

The common exponent exceeds the minimax exponent by

$$\frac{\nu d}{2(\nu+d)(2\nu+d)}. \tag{112}$$

Thus the Whitehouse–Wu–Ramdas upper exponent and the Wang–Zhang lower exponent coincide at this effective-optimism scale. The two statements are not a matching $\widetilde{\Theta}$ theorem for one identical implementation: the former holds a horizon-dependent regularizer fixed during the run, whereas the latter assumes time-indexed, two-sided polynomial-times-log schedules.

E.5 AMS/EBO specialization for kernel bandits

Algorithm 3: AMS/EBO on the action index. AMS/EBO approaches the same specialized log-potential bracket through robust posterior optimization. In the finite-action kernel marginal, the natural AIR index is $A^\star$. Let q_t be the current reference distribution on $A^\star$, let $\mathfrak{B}_t(q_t)$ be a calibrated local set of candidate beliefs ν on the active reward vector F, and fix a constant coefficient $\gamma > 0$. For each ν, define $\Delta_\nu^{\mathrm{AIR}}(p)$ and $\mathcal{I}_\nu^{\mathrm{AIR}}(p)$ by the analogues of (59) and (58), using the conditional laws induced by ν. A robust AIR/EBO decision is

$$p_t \in \arg\min_{p \in \Delta(\mathcal{X})} \sup_{\nu \in \mathfrak{B}_t(q_t)} \left\{ \Delta_\nu^{\mathrm{AIR}}(p) - \gamma \mathcal{I}_\nu^{\mathrm{AIR}}(p) - \gamma D_{\mathrm{KL}}(\nu_{A^\star} \| q_t) \right\}. \tag{113}$$

The maximization over ν is essential: without it the display is only the posterior AIR bracket for a chosen belief, not the robust AMS/EBO relaxation. In a finite-index abstraction, the last two terms are the KL-dual form of the log-partition potential (30). If the robust value in (113) is at most ε_t, the required concavity/first-order conditions in Lemma D.2 hold, and the belief set is surely calibrated so that the fixed truth $f^\star$ is an admissible comparison direction at every reached state, then Theorem 4.1 gives

$$\mathbb{E}_{f^\star} \operatorname{Reg}_T \le \gamma \log \frac{1}{q_1(a^\star)} + \mathbb{E}_{f^\star} \sum_{t=1}^{T} \varepsilon_t, \tag{114}$$

where $a^\star = A^\star(f^\star)$. The constant coefficient avoids the false shortcut of replacing a weighted fixed-truth log-gain sum by a single telescope. If a varying coefficient is used, the corresponding weighted telescope must be proved separately. Under a Bayesian reference prior, posterior averaging of the same coordinate bound gives

$$\mathbb{E}\,\mathrm{Reg}_T \le \gamma H(A^\star) + \mathbb{E}\sum_{t=1}^{T} \varepsilon_t.$$

A ratio form gives the familiar Cauchy–Schwarz variant after replacing the fixed coefficient by the appropriate sequential AIR ratio and applying the action-index chain rule. Since $H(A^\star) \le \log n$, this route scales with the finite decision index and the sequential AIR coefficient.

E.6 Numerical protocol for the hub–cloud experiment

The scalar saddle (116) is a two-model Bellman-supersolution, not the full continuous-RKHS finite-horizon program. The AIR/AMS mean under the negative retained hub truth, not shown in Figure 1, is 56.7. The truncation curve is a deterministic counterfactual for one-hot cloud truths outside the binary saddle class; it is not an additional Monte Carlo guarantee. We now give the full protocol and the two fixed-truth bracket checks. For each $T \in \{512, 1024, 2048, 4096, 8192, 16384\}$, the experiment takes $\alpha = 1/5$, $a = T^{-1/5}$, $N = 4T$, unit observation variance, and $\Delta = 1/4$. Because $N \ge T$, the maximal information gain of the diagonal kernel has the closed form

$$\Gamma_s(T) = \frac{1}{2} \max_{h \in \{0,\dots,s\}} \left\{ \log(1+h) + (s-h)\log(1+a^2) \right\}. \tag{115}$$

The integer maximizer lies among the clipped integers adjacent to $1/\log(1+a^2) - 1$. The horizon-aware rule uses (98) without approximation, while the finite-block anytime rule replaces $2\log T$ by $2\log(t \vee 2)$. For the original schedule with $\delta = 0.1$, the exact distinct-set information gain is

$$\Gamma_s^{\mathrm{SKKS}}(T) = \frac{1}{2}\{\log 2 + (s-1)\log(1+a^2)\}, \qquad 1 \le s \le T,$$

because an optimal size-s set contains the hub and $s-1$ distinct cloud points. For a diagonal coordinate of prior variance d, after n pulls with observation sum S, the implementation uses posterior mean $dS/(1+nd)$ and variance $d/(1+nd)$. If a sampled cloud coordinate has posterior mean m and standard deviation $s < a$, its score gap from an unused cloud coordinate is $m - \beta_t(a-s)$, which is nonincreasing because β_t is nondecreasing. Once this gap is nonpositive, the coordinate cannot be selected again. Ties between sampled and unused cloud coordinates are resolved in favor of the unused one; equivalently, under Gaussian noise the pruning gives almost surely the same choices as direct maximization over all $N+2$ actions.

For the localized comparator, let f^+ and f^- have hub means $+\Delta$ and $-\Delta$, respectively, and value zero at the anchor. Their optimal-action indices are the hub and the anchor. Write q for the current reference mass on f^+, r for the candidate pair-belief mass on f^+, and p for the probability of selecting the hub. The following control is a direct two-model Gaussian specialization of the AIR/AMS saddle of Xu and Zeevi (2025). Its relation to Lattimore and György (2021) is through the EBO principle and minimax/KL duality: the numerical code solves the belief-space AIR/AMS saddle, not EBO's functional-estimator mirror-descent implementation. The robust AIR/AMS saddle is

$$\inf_{0\le p\le 1} \sup_{0\le r\le 1} \left[\Delta\{r(1-p) + (1-r)p\} - \gamma p I_r - \gamma\,\mathrm{kl}(r,q)\right], \qquad \gamma = \sqrt{\frac{(1+\Delta^2)T}{\log 2}}, \tag{116}$$

where I_r is the mutual information of the binary Gaussian channel $Z \mid f^{\pm} \sim N(\pm\Delta, 1)$:

$$I_r = r\mathbb{E}_{Z\sim N(\Delta,1)}\left[-\log\{r + (1-r)e^{-2\Delta Z}\}\right] + (1-r)\mathbb{E}_{Z\sim N(-\Delta,1)}\left[-\log\{(1-r) + re^{2\Delta Z}\}\right]. \tag{117}$$

This is a finite instance of (113). The exact saddle is calibrated only to the displayed pair class $\{f^+, f^-\}$, not to the full projected RKHS ball. Sion's identity reduces it to

$$\sup_{0\le r\le 1} \min\left\{\Delta r - \gamma\,\mathrm{kl}(r,q), \Delta(1-r) - \gamma I_r - \gamma\,\mathrm{kl}(r,q)\right\},$$

which is a scalar concave maximization. For completeness, let $v_c = \mathrm{logit}(r_c)$, where r_c solves

$$\Delta(2r_c - 1) + \gamma I_{r_c} = 0,$$

and write $I'(r) = dI_r/dr$. If $u = \mathrm{logit}(q)$ and $v = \mathrm{logit}(r)$, define

$$u_0 = v_c - \Delta/\gamma, \qquad u_1 = v_c + \Delta/\gamma + I'(r_c).$$

The saddle has $p = 0$ and $\mathrm{logit}(r) = u + \Delta/\gamma$ for $u \le u_0$; it has $r = r_c$ and $p = (u - u_0)/(u_1 - u_0)$ for $u_0 < u < u_1$; and it has $p = 1$ for $u \ge u_1$, with $v = \mathrm{logit}(r)$ determined by

$$v - u + \Delta/\gamma + I'\big(\mathrm{logit}^{-1}(v)\big) = 0.$$

After an anchor pull, $u^+ = v$. After a hub observation $Z = z$, Gaussian Bayes updating gives $u^+ = v + 2\Delta z$.

The numerical policy evaluates (117) by 48-node Gauss–Hermite quadrature. Scalar roots are solved to absolute tolerance 2×10^{-13}; the smooth $p = 1$ branch is tabulated on 5001 log-odds points and linearly interpolated, with its limiting shift used beyond the tabulated range. Thus the code approximates the exact scalar saddle. As a diagnostic, for each $\sigma \in \{+, -\}$ it reevaluates the fixed-truth bracket by 96-node quadrature,

$$\Delta_\sigma(p) - \gamma\mathbb{E}_\sigma \log \frac{q^+(Y_\sigma \mid A, Z)}{q(Y_\sigma)} - \frac{1+\Delta^2}{\gamma}$$

Here $\Delta_\sigma(p)$ and Y_σ denote, respectively, the one-step regret and the optimal-action index under f^σ. The expression is evaluated on a grid of more than ten thousand reference log-odds and at every reference state reached in the Monte Carlo runs. The largest static-grid and realized-state values over all displayed horizons are approximately -3.584×10^{-3}, and the largest tabulated first-order residual is 6.81×10^{-8}. Thus no positive fixed-truth Bellman-bracket violation is observed at the reported numerical precision. This numerical check is diagnostic, not an interval-arithmetic proof of a uniform inequality.

Each algorithm–truth pair uses 200 independent NumPy streams spawned from the master seed 20260717; the streams of the original horizon-aware/AIR experiment are retained unchanged, and separate streams are used for the two added GP–UCB schedules. Shaded bands (left) and error bars (right) show two Monte Carlo standard errors. The right-panel truncation curve is the deterministic normalized value $(aT)/T = a$, not a Monte Carlo estimate. The one-hot cloud truths are external to the binary saddle class; their role is only to evaluate the deterministic loss incurred by the predetermined truncation.

F Gaussian MAB: finite-index matching

In this section, we apply our lower-bound approach to multi-armed bandits (MAB). Consider $K \geq 3$ Gaussian arms with unit noise variance. Let the ambient model space be a class $\Omega_{\mathrm{MAB}} \subseteq \mathbb{R}^K$ of mean vectors, and write a model environment as $\omega = (\theta_\omega(1), \dots, \theta_\omega(K))$. Pulling arm a in environment ω produces an observation distributed as $N(\theta_\omega(a), 1)$. Let

$$A^\star(\omega) \in \arg\max_{a \in [K]} \theta_\omega(a)$$

with an arbitrary fixed tie-breaking rule, and define the information index

$$\chi(\omega) = A^\star(\omega) \in [K].$$

Thus the random environment is the full mean vector Ω, while the information target used in the lower bound is only

$$Y = \chi(\Omega) = A^\star(\Omega).$$

For the lower bound, fix $\Delta > 0$ and use the K-point prior μ_Δ supported on the spike environments

$$\omega^j = \Delta e_j, \qquad j \in [K],$$

where e_j is the jth coordinate vector. Equivalently, draw $J \sim \mathrm{Unif}([K])$ and set $\Omega = \omega^J$. On this prior support the optimal arm is unique and

$$Y = \chi(\Omega) = J.$$

The equality $Y = J$ is only a property of this least-favorable prior; it should not be read as identifying the information target with the full model environment. The one-step regret in environment ω^j is

$$\ell_{\omega^j}(a) = \max_b \theta_{\omega^j}(b) - \theta_{\omega^j}(a) = \Delta \mathbf{1}\{a \neq j\}.$$

For brevity write $\ell_j = \ell_{\omega^j}$ on this subfamily.

Lemma F.1 (MAB ghost entropy)**.** *Let $Y = \chi(\Omega)$ under the prior $\Omega \sim \mu_\Delta$. For every algorithm and every reference history H'_T independent of $Y \sim \mathrm{Unif}([K])$,*

$$\mathbb{P}\left(\bar{L}_Y(H'_T) \leq \frac{\Delta}{2}\right) \leq \frac{2}{K},$$

where $\bar{L}_y(H'_T) = T^{-1} \sum_{t=1}^T \ell_y(A'_t)$ and A'_t denotes the arm selected in the reference history at time t.

Proof. For a fixed reference action sequence $a'_1, \dots, a'_T$, let

$$N'_y = \sum_{t=1}^T \mathbf{1}\{a'_t = y\}.$$

The average regret against the spike environment ω^y is

$$\bar{L}_y(H'_T) = \Delta(1 - N'_y/T).$$

The event $\bar{L}_y(H'_T) \le \Delta/2$ implies $N'_y \ge T/2$. Since Y is uniform and independent of the reference sequence,

$$\mathbb{E}[N'_Y \mid H'_T] = \frac{1}{K}\sum_{y=1}^{K} N'_y = \frac{T}{K}.$$

Markov's inequality gives

$$\mathbb{P}(N'_Y \ge T/2 \mid H'_T) \le \frac{2}{K}.$$

Averaging over H'_T proves the claim. □

Lemma F.2 (MAB information bound)**.** *There is a universal constant $c_0 > 0$ such that if $\Delta^2 T/K \le c_0$, then every adaptive algorithm satisfies*

$$I(Y; H_T) \le \frac{1}{4}\log(K/2),$$

where $Y = \chi(\Omega)$ and $\Omega \sim \mu_\Delta$.

Proof. Let $\mathbb{P}_0$ be the law of the history under the zero-mean reference model, let $\mathbb{P}_j$ be the law under the spike environment ω^j, and let

$$\bar{\mathbb{P}} = \frac{1}{K}\sum_{j=1}^{K} \mathbb{P}_j$$

be the marginal law of H_T under the prior μ_Δ. Under this prior, $Y = j$ is equivalent to $\Omega = \omega^j$, but the mutual information is still the action-index information $I(Y; H_T)$, not information about an ambient full mean vector outside the prior support. Since $\bar{\mathbb{P}} \ge K^{-1}\mathbb{P}_j$, the likelihood ratio $d\mathbb{P}_j/d\bar{\mathbb{P}}$ is bounded by K. The inequality

$$D_{\mathrm{KL}}(P\|Q) \le (2 + \log K) H^2(P, Q), \qquad \text{whenever } Q \ge K^{-1}P,$$

where

$$H^2(P, Q) := \int (\sqrt{dP} - \sqrt{dQ})^2$$

is the standard squared Hellinger distance (twice the h^2 normalization used earlier), together with the triangle inequality and convexity for squared Hellinger distance, gives

$$I(Y; H_T) = \frac{1}{K}\sum_{j=1}^{K} D_{\mathrm{KL}}(\mathbb{P}_j\|\bar{\mathbb{P}}) \le 4(2 + \log K)\frac{1}{K}\sum_{j=1}^{K} H^2(\mathbb{P}_j, \mathbb{P}_0).$$

Using $H^2(P, Q) \le D_{\mathrm{KL}}(Q\|P)$ and the adaptive Gaussian KL chain rule under the zero model,

$$\frac{1}{K}\sum_{j=1}^{K} H^2(\mathbb{P}_j, \mathbb{P}_0) \le \frac{1}{K}\sum_{j=1}^{K} D_{\mathrm{KL}}(\mathbb{P}_0\|\mathbb{P}_j) = \frac{\Delta^2}{2K}\mathbb{E}_0 \sum_{t=1}^{T}\sum_{j=1}^{K} \mathbf{1}\{A_t = j\} = \frac{\Delta^2 T}{2K}.$$

Therefore

$$I(Y; H_T) \le 2(2 + \log K)\frac{\Delta^2 T}{K}.$$

Choosing, for example,

$$c_0 \le \inf_{K \ge 3} \frac{\log(K/2)}{8(2 + \log K)}$$

which is a strictly positive universal constant, proves the displayed bound. □

Theorem F.3 (MAB lower bound)**.** *For Gaussian K-armed bandits with unit noise variance, over any ambient mean-vector class Ω_{MAB} that contains the spike instances $\{\Delta e_j : j \in [K]\}$ used below,*

$$\mathfrak{R}_T^\star \geq c\sqrt{KT}$$

for a universal constant $c > 0$ whenever $T \geq K \geq 3$. In particular, this applies to any bounded class such as $[0,1]^K$ once the universal constant in the choice of Δ is small enough.

Proof. Fix an arbitrary algorithm and abbreviate $p_r = p_r^{\mathsf{Alg}}(\mu_\Delta, \chi)$ and $T\bar{C} = C_\chi^{\mathsf{Alg}}(\mu_\Delta)$. Choose

$$\Delta = c_1\sqrt{K/T}$$

with c_1 a sufficiently small universal constant, and let $r = \Delta/2$. Lemma F.1 gives $p_r \leq 2/K \leq 2/3$. The proof of Lemma F.2 gives the sharper bound

$$T\bar{C} = I(Y; H_T) \leq 2(2 + \log K)c_1^2.$$

Choose c_1 so small that, for every $K \geq 3$,

$$2(2 + \log K)c_1^2 \leq \mathrm{kl}\left(\frac{3}{4}, \frac{2}{K}\right).$$

This is possible because

$$\inf_{K \geq 3} \frac{\mathrm{kl}\left(\frac{3}{4}, \frac{2}{K}\right)}{2(2 + \log K)} > 0.$$

Since $p_r \leq 2/K \leq 2/3 < 3/4$, the exact quantile bound (41) gives

$$\psi(p_r, T\bar{C}) \leq \frac{3}{4}.$$

Therefore

$$\mathfrak{R}_T^\star \geq Tr\left(1 - \frac{3}{4}\right) = \frac{T\Delta}{8} = c\sqrt{KT},$$

where $c > 0$ is a universal constant. □

Proposition F.4 (MAB upper bound and minimax-rate matching)**.** *For Gaussian K-armed bandits with unit noise variance and mean vectors in $[0,1]^K$, there exists a universal constant $C > 0$ and a nonanticipating algorithm, for example a sub-Gaussian version of MOSS, whose regret satisfies*

$$\sup_\omega \mathbb{E}_\omega^{\mathsf{Alg}} L_\omega(H_T) \leq C\sqrt{KT}$$

for all $T \geq K \geq 3$. For another fixed bounded cube, the constant may depend on its mean range. Consequently, together with Theorem F.3,

$$\mathfrak{R}_T^\star(\Omega_{\mathrm{MAB}}) = \Theta(\sqrt{KT})$$

on every ambient class containing the spike hard family and contained in $[0,1]^K$. In the notation of Section 5.4, the hard prior has $M = K$ index values, the upper code length on that same spike subfamily is $L_0 = \log K$, and, because Lemma F.1 holds for every algorithm, the algorithm-uniform ghost entropy satisfies

$$E_{\mu_\Delta, r}^\star = \log \frac{1}{p_{\mu_\Delta, r}^\star} \geq \log(K/2)$$

at radius $r = \Delta/2 \asymp \sqrt{K/T}$. Hence the hard-subfamily entropy ratio $L_0/E_{\mu_\Delta, r}^\star$ is bounded by a universal constant for $K \geq 3$, while the standard minimax upper bound matches the lower rate on the ambient class. This rate comparison does not by itself verify upper-at-lower-entropy calibration of the specific Bellman program for MOSS.

Proof. The lower bound is Theorem F.3. The upper bound is the standard distribution-free minimax upper bound for sub-Gaussian stochastic multi-armed bandits, achieved by a sub-Gaussian MOSS variant and recorded in modern bandit texts (Audibert and Bubeck, 2009; Lattimore and Szepesvári, 2020). The entropy comparison follows directly from the explicit hard prior: under the uniform optimal-arm index prior $q_1(j) = 1/K$, the upper coordinate code length is $\log K$, while Lemma F.1 gives $p^\star_{\mu_\Delta,r} \le 2/K$. Therefore $L_0/E^\star_{\mu_\Delta,r} \le \log K/\log(K/2)$, which is bounded for $K \ge 3$. □

This is the cleanest finite-index example of the Bellman–Fano entropy calculation together with a matching minimax rate. The full environment is the whole mean vector, but the relevant index is the optimal arm. The hard prior, ghost entropy, and information bound operate at the same radius, and a standard minimax algorithm attains the resulting rate; a formal Bellman sandwich for that particular algorithm would additionally identify its upper budget value and verify Definition A.6.

G Hypercube linear bandits: indexed matching

Consider stochastic linear bandits with action set

$$\mathcal{A} = \{a \in \mathbb{R}^d : \|a\|_2 \le 1\}.$$

Let the ambient model class be the Euclidean unit parameter ball

$$\Omega_{\mathrm{lin}} = \{\omega \in \mathbb{R}^d : \|\omega\|_2 \le 1\},$$

and regard an environment $\omega \in \Omega_{\mathrm{lin}}$ as the full mean parameter vector. At time t, after choosing $A_t \in \mathcal{A}$, the learner observes

$$O_t = \langle \omega, A_t \rangle + \xi_t, \qquad \xi_t \sim N(0,1),$$

where the noises are independent over time and independent of the learner's external randomization. For $\omega \neq 0$, the unique optimal action over $\mathcal{A}$ is

$$A^\star(\omega) = \frac{\omega}{\|\omega\|_2}.$$

At $\omega = 0$, fix any deterministic tie-breaking action; this case is not used by the lower-bound prior. The information index is the optimal action

$$\chi(\omega) = A^\star(\omega), \qquad Y = \chi(\Omega),$$

not the full environment Ω. Thus two environments that have the same optimal action are not distinguished by the index unless the prior support itself makes them distinguishable.

We prove the standard $\Omega(d\sqrt{T})$ minimax lower bound through the quantile indexed information theorem. The proof uses a hypercube prior together with an adaptive information bound. The role of the hypercube variable below is only to code the finitely many optimal actions in the prior support. It should not be read as replacing the ambient environment Ω by a model label. Note that for a non-Gaussian prior, one cannot condition on the realized adaptive design matrix and then apply the fixed-design Gaussian-channel capacity formula, because the design itself is a statistic of the past observations and may carry information about the unknown environment and hence about Y. Chen et al. (2024, Section 3.3) prove a closely related lower bound by a different route: they use the interactive Fano method with a Gaussian prior and a unit-sphere

calculation. Both approaches show that action-indexed information arguments can recover the sharp linear-bandit order by charging the optimal direction. Model-indexed DEC analyses charge a different information target and can recover the same order when their localization reflects this decision-relevant geometry.

Let

$$V \sim \mathrm{Unif}(\{\pm 1\}^d), \qquad \Omega = \Delta V.$$

Assume $\Delta\sqrt{d} \le 1$, so that this prior is supported on Ω_{lin}. On this prior support,

$$Y = \chi(\Omega) = A^\star(\Omega) = \frac{V}{\sqrt{d}}. \tag{118}$$

The map $v \mapsto v/\sqrt{d}$ is one-to-one on $\{\pm 1\}^d$. Hence, for this prior only, V may be used as a code for the action index Y, and

$$I(Y; H_T) = I(V; H_T).$$

For an index value $y = v/\sqrt{d}$, write ℓ_v for the regret under the corresponding environment $\omega = \Delta v$. The one-step regret of action $a \in \mathbb{R}^d$, $\|a\|_2 \le 1$, is

$$\ell_v(a) = \Delta\sqrt{d} - \Delta\langle v, a\rangle.$$

Assume $T \ge d^2$ and choose $\Delta \le 1/\sqrt{d}$, so that $\|\Omega\|_2 = \Delta\sqrt{d} \le 1$.

Lemma G.1 (Linear ghost entropy)**.** *There is a universal constant $c_{\mathrm{gh}} > 0$ such that every deterministic reference action sequence $a'_1, \dots, a'_T$, with $\|a'_t\|_2 \le 1$, satisfies*

$$\mathbb{P}_{V \sim \mathrm{Unif}(\{\pm 1\}^d)}\left(\frac{1}{T}\sum_{t=1}^{T}\big(\Delta\sqrt{d} - \Delta\langle V, a'_t\rangle\big) \le \frac{\Delta\sqrt{d}}{4}\right) \le \exp(-c_{\mathrm{gh}} d).$$

Consequently, the same bound holds after averaging over any random reference history H'_T whose induced action sequence is independent of the action index Y, equivalently independent of V under the hypercube prior.

Proof. Let

$$\bar{a} := \frac{1}{T}\sum_{t=1}^{T} a'_t.$$

Since $\|\bar{a}\|_2 \le T^{-1}\sum_t \|a'_t\|_2 \le 1$, the event in the display implies

$$\Delta\sqrt{d} - \Delta\langle V, \bar{a}\rangle \le \frac{\Delta\sqrt{d}}{4},$$

or equivalently,

$$\langle V, \bar{a}\rangle \ge \frac{3\sqrt{d}}{4}.$$

The random variable $\langle V, \bar{a}\rangle = \sum_{i=1}^{d} V_i \bar{a}_i$ is a centered Rademacher sum with variance proxy $\|\bar{a}\|_2^2 \le 1$. Hoeffding's inequality therefore gives

$$\mathbb{P}\left(\langle V, \bar{a}\rangle \ge \frac{3\sqrt{d}}{4}\right) \le \exp\left(-\frac{9d}{32}\right).$$

Thus the first claim holds, for example with $c_{\mathrm{gh}} = 9/32$. If the reference action sequence is random but independent of Y, then because $Y = V/\sqrt{d}$ is a bijective function of V on the prior support, the reference sequence is independent of V. Conditioning on the reference history and applying the deterministic bound gives the second claim. □

Lemma G.2 (Adaptive hypercube information capacity)**.** *For the hypercube prior above, every possibly randomized adaptive algorithm satisfies*

$$I(Y; H_T) = I(V; H_T) \le \Delta^2 T,$$

where $H_T = (A_1, O_1, \dots, A_T, O_T)$.

Proof. Since $Y = V/\sqrt{d}$ is a bijective function of V on the hypercube support, $I(Y; H_T) = I(V; H_T)$. It remains to bound $I(V; H_T)$. Let U denote the learner's internal random seed, independent of V. Then

$$I(V; H_T) \le I(V; H_T, U) = \mathbb{E}_U I(V; H_T \mid U).$$

It is therefore enough to prove the claim after conditioning on U, so that the algorithm is deterministic.

For $v \in \{\pm 1\}^d$, let $\mathbb{P}_v$ denote the trajectory law when $\Omega = \Delta v$. We use the chain rule over coordinates:

$$I(V; H_T) = \sum_{i=1}^{d} I(V_i; H_T \mid V_1, \dots, V_{i-1}).$$

Fix a coordinate i and a prefix $u = (v_1, \dots, v_{i-1})$. Let $\mathbb{P}_{u,+}$ and $\mathbb{P}_{u,-}$ be the trajectory laws after averaging uniformly over the remaining coordinates $V_{i+1}, \dots, V_d$, conditional respectively on $V_i = +1$ and $V_i = -1$. Then

$$I(V_i; H_T \mid V_{<i} = u) = D_{\mathrm{JS}}(\mathbb{P}_{u,+}, \mathbb{P}_{u,-}),$$

where D_{JS} denotes Jensen–Shannon divergence with equal weights. For any two probability measures P, Q,

$$D_{\mathrm{JS}}(P, Q) \le \frac{1}{4}\Big(D_{\mathrm{KL}}(P\|Q) + D_{\mathrm{KL}}(Q\|P)\Big).$$

Next couple the two mixtures by using the same suffix. If $w \in \{\pm 1\}^{d-i}$ and

$$v^+(u, w) = (u, +1, w), \qquad v^-(u, w) = (u, -1, w),$$

then joint convexity of relative entropy gives

$$D_{\mathrm{KL}}(\mathbb{P}_{u,+}\|\mathbb{P}_{u,-}) \le \mathbb{E}_w D_{\mathrm{KL}}\left(\mathbb{P}_{v^+(u,w)} \,\middle\|\, \mathbb{P}_{v^-(u,w)}\right),$$

and the same argument gives the analogous reverse inequality.

For two fixed vertices v^+ and v^- that differ only in coordinate i, the adaptive KL chain rule gives

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_{v^+}\|\mathbb{P}_{v^-}) &= \frac{1}{2}\mathbb{E}_{v^+}\sum_{t=1}^{T}\Big(\Delta\langle v^+ - v^-, A_t\rangle\Big)^2 \\ &= 2\Delta^2\mathbb{E}_{v^+}\sum_{t=1}^{T} A_{t,i}^2. \end{aligned}$$

Here the action kernel contributes no KL because, after conditioning on the learner's internal randomness, the algorithm uses the same decision rule under both environments; only the Gaussian observation kernel changes. Similarly,

$$D_{\mathrm{KL}}(\mathbb{P}_{v^-}\|\mathbb{P}_{v^+}) = 2\Delta^2\mathbb{E}_{v^-}\sum_{t=1}^{T} A_{t,i}^2.$$

Combining the preceding displays yields

$$I(V_i; H_T \mid V_{<i} = u) \le \frac{\Delta^2}{2}\mathbb{E}_w\left[\mathbb{E}_{v^+(u,w)}\sum_{t=1}^{T} A_{t,i}^2 + \mathbb{E}_{v^-(u,w)}\sum_{t=1}^{T} A_{t,i}^2\right]$$
$$= \Delta^2\mathbb{E}\left[\sum_{t=1}^{T} A_{t,i}^2 \,\middle|\, V_{<i} = u\right],$$

where the final expectation is under the original hypercube prior and the algorithm's trajectory law. Averaging over $V_{<i}$ and summing over i gives

$$\begin{aligned} I(V; H_T) &\le \Delta^2\mathbb{E}\sum_{t=1}^{T}\sum_{i=1}^{d} A_{t,i}^2 \\ &= \Delta^2\mathbb{E}\sum_{t=1}^{T}\|A_t\|_2^2 \\ &\le \Delta^2 T. \end{aligned}$$

This proves the claim. □

Remark G.3 (Why the fixed-design log-determinant argument is not used)**.** *For a Gaussian prior on a full parameter vector, posterior conjugacy gives the adaptive information identity*

$$I(\Theta; H_T) = \frac{1}{2}\mathbb{E}\log\det\left(I + \Sigma_0^{1/2}\Big(\sum_{t=1}^{T} A_t A_t^\top\Big)\Sigma_0^{1/2}\right),$$

and this is bounded by a trace-constrained log determinant. For the hypercube prior, however, the posterior is not Gaussian and the realized design matrix is itself a statistic of the observations. Conditioning only on that realized design matrix controls the conditional observation information given the design; it does not by itself control the information carried by the adaptive design. Lemma G.2 therefore uses a coordinatewise adaptive KL argument to control $I(Y; H_T)$ directly.

Theorem G.4 (Linear bandit lower bound)**.** *For stochastic linear bandits with action set $\{a \in \mathbb{R}^d : \|a\|_2 \le 1\}$, unit Gaussian noise, and unit-ball parameter class $\Omega_{\mathrm{lin}} = \{\omega \in \mathbb{R}^d : \|\omega\|_2 \le 1\}$,*

$$\mathfrak{R}_T^\star \ge c\, d\sqrt{T}$$

for a universal constant $c > 0$, whenever $T \ge d^2$.

Proof. Fix an arbitrary algorithm and abbreviate by p_r and $T\bar{C}$ its ghost-good probability and cumulative action-index information under the hypercube prior. Assume first that d is larger than a sufficiently large universal constant d_0 to be fixed below. Choose

$$\Delta = c_1\sqrt{\frac{d}{T}},$$

where $c_1 > 0$ is a sufficiently small universal constant. Since $T \ge d^2$, choosing $c_1 \le 1$ ensures

$$\Delta\sqrt{d} = c_1\frac{d}{\sqrt{T}} \le c_1 \le 1,$$

so the hypercube prior is supported on the unit parameter ball.

Let

$$r = \frac{\Delta\sqrt{d}}{4}.$$

Lemma G.1 gives the reference ghost-good probability bound

$$p_r \le \exp(-c_{\mathrm{gh}} d).$$

Lemma G.2 gives the adaptive action-index information bound

$$T\bar{C} = I(Y; H_T) = I(V; H_T) \le \Delta^2 T = c_1^2 d.$$

Choose c_1 so small that $c_1^2 \le c_{\mathrm{gh}}/4$, and then choose d_0 so large that $\log 2 \le c_{\mathrm{gh}} d/4$ for all $d \ge d_0$. Then, for all $d \ge d_0$,

$$T\bar{C} + \log 2 \le \frac{1}{2} \log \frac{1}{p_r}.$$

Theorem 5.1, applied to the action index $Y = \chi(\Omega) = A^\star(\Omega)$, gives a Bayes regret lower bound under the hypercube prior of at least

$$\frac{Tr}{2} = \frac{T\Delta\sqrt{d}}{8} = \frac{c_1}{8} d\sqrt{T}.$$

Since this prior is supported on Ω_{lin}, the minimax regret over Ω_{lin} is at least this Bayes risk.

It remains to handle the finitely many dimensions $1 \le d < d_0$. For these dimensions, use the two-point prior $\Omega = \sigma\delta e_1$, where $\sigma \sim \mathrm{Unif}(\{\pm 1\})$, e_1 is the first coordinate vector, and $\delta = (2\sqrt{T})^{-1}$. This prior is also supported on the unit parameter ball. Under $\sigma = +1$, the optimal action is e_1; under $\sigma = -1$, the optimal action is $-e_1$. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the laws of the history under these two environments, and let $\mathbb{P}_\pm^{t-1}$ denote the law of the history before action t. For any algorithm,

$$\begin{aligned}\frac{1}{2}\mathbb{E}_+\big[\delta(1 - A_{t,1})\big] + \frac{1}{2}\mathbb{E}_-\big[\delta(1 + A_{t,1})\big] &= \delta\left(1 - \frac{1}{2}(\mathbb{E}_+ A_{t,1} - \mathbb{E}_- A_{t,1})\right) \\ &\ge \delta\left(1 - \mathrm{TV}(\mathbb{P}_+^{t-1}, \mathbb{P}_-^{t-1})\right).\end{aligned}$$

By Pinsker's inequality and the adaptive Gaussian KL chain rule,

$$\mathrm{TV}(\mathbb{P}_+^{t-1}, \mathbb{P}_-^{t-1}) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_+^{t-1} \| \mathbb{P}_-^{t-1})} \le \delta\sqrt{t-1} \le \frac{1}{2}.$$

Therefore each round has Bayes regret at least $\delta/2$, and the total Bayes regret is at least $T\delta/2 = \sqrt{T}/4$. Since $d < d_0$, this is at least $(4d_0)^{-1} d\sqrt{T}$. Reducing the universal constant c, if necessary, completes the proof for all d. □

Proposition G.5 (Linear-bandit upper bound and near matching)**.** *For stochastic linear bandits on the Euclidean unit ball with unit Gaussian noise and unit-ball parameter class, there are algorithms based on confidence ellipsoids, such as OFUL, with regret*

$$\sup_{\omega \in \Omega_{\mathrm{lin}}} \mathbb{E}_\omega^{\mathsf{Alg}} L_\omega(H_T) \le Cd\sqrt{T}\ \mathrm{polylog}(T, d)$$

for a universal constant C and all $T \ge d$. Consequently Theorem G.4 is matched in the polynomial order $d\sqrt{T}$, with only logarithmic factors belonging to the particular tractable upper bound. On the hard hypercube prior, the index support has $M = 2^d$, the initial code length on that same

hypercube subfamily is $L_0 = d \log 2$, and Lemma G.1 gives the algorithm-uniform ghost entropy $E_r^\star := \log(1/p_r^\star) \geq c_{\mathrm{gh}} d$ at radius $r = \Delta\sqrt{d}/4 \asymp d/\sqrt{T}$. Thus the Bellman–Fano hard-subfamily entropy gap is constant, while OFUL supplies minimax-rate matching up to logarithmic factors. This does not by itself verify the upper-at-lower-entropy calibration of the exact log-potential Bellman program for the full unit-ball class.

Proof. The standard self-normalized analysis for stochastic linear bandits gives the upper bound (Abbasi-Yadkori et al., 2011; Lattimore and Szepesvári, 2020). The lower bound and entropy calculation follow directly from the proof of Theorem G.4: the hard index is $Y = V/\sqrt{d}$ with $V \in \{\pm 1\}^d$, so the hypercube-subfamily code length is $L_0 = d \log 2$, while Lemma G.1 yields $p_r^\star \leq e^{-c_{\mathrm{gh}} d}$. □

This example is the high-dimensional analogue of the finite-arm calculation. The hypercube prior is not a model-identification target; it is a finite action-index packing for the optimal direction. The lower proof shows that ghost entropy and information capacity meet at the scale $d\sqrt{T}$. Standard tractable upper algorithms achieve the same scale up to logarithmic terms; proving that the exact log-potential Bellman upper value closes at this entropy scale would give the corresponding formal sandwich theorem. It should be distinguished from the two irregular-action-set results discussed in Section 6.2. Corollary E.3 gives an algorithm-specific cumulative-regret separation for a horizon-dependent small-cloud action set, whereas Maiti et al. (2026) separate adaptive from nonadaptive sampling for fixed-confidence best-arm identification on a union of orthogonal balls. These results complement the Euclidean-ball minimax calculation here; they concern different action geometries, performance criteria, and algorithm classes.

H Proofs

The remaining proofs are collected here.

H.1 State compression and the information-complexity sandwich

H.1.1 Proof of Proposition 2.3

Proof. The Bellman-operator identity follows directly from the state-measurability of the action set, comparison set, one-step cost, transition law, and continuation value. Histories in the same fiber of S_t induce the same local optimization problem. The information inequality is the data-processing inequality applied to the measurable map $S_t = \phi_t(H_{t-1})$. The equality condition is the standard equality case for conditional mutual information, $I(Y; H_{t-1} \mid S_t) = 0$. □

H.1.2 Proof of Lemma A.2

Proof. Since $C_\mu \geq 0$, condition (74) implies $\underline{E}_{\mu,r} \geq 2 \log 2$. Because $\underline{E}_{\mu,r} \leq E_{\mu,r}^\star = -\log p_{\mu,r}^\star$, one has $p_{\mu,r}^\star \leq e^{-\underline{E}_{\mu,r}} \leq 1/4$. The converse statement is only the exact entropy bound $E_{\mu,r}^\star \geq \log(1/\varrho)$; it does not control C_μ. □

H.1.3 Proof of Proposition A.3

Proof. The uniform representative prior gives $q_1(y) = 1/M$, hence the upper code length is $\log M$. If μ_h is Bellman–Fano admissible, Lemma A.2 gives $E_{\mu_h,r}^\star \geq 2 \log 2$, proving (75). Under the

multiplicity condition, every algorithm satisfies

$$p^{\mathsf{Alg}}_{\mu_h,r} = \mathbb{E}_{H'_T \sim \bar{\mathbb{P}}^{\mathsf{Alg}}_{\mu_h}} \frac{|G_r(H'_T)|}{M} \le \frac{m_r}{M}.$$

Taking the supremum over Alg gives $p^\star_{\mu_h,r} \le m_r/M$, and hence $E^\star_{\mu_h,r} \ge \log(M/m_r)$. This proves (76). The final statement follows from $E^\star_{\mu_h,r} \ge \log(1/\varrho)$. □

H.1.4 Proof of Lemma A.5

Proof. Set $\mathsf{U}(x) = 1$ for $x < L$ and $\mathsf{U}(x) = \max\{1, B\}$ for $x \ge L$. Then U is nondecreasing, $\mathsf{U}(E) = 1$, and $\mathsf{U}(L) = \max\{1, B\} \ge B$. Thus a logarithmic entropy gap, even $L/E = O(\log M)$, does not by itself imply any controlled regret gap. Assumption A.4, or an equivalent fixed-point/rate condition, is the additional structure that rules out such jumps. □

H.1.5 Proof of Proposition A.7

Proof. Condition (i) is (78). Condition (ii) implies (i) because $\mathsf{U}_T(E^\star_{\mu,r})$ is the infimum over $\gamma > 0$ of the left side in (77). Condition (iii) is a restatement of (i) with the constant displayed after dividing by T. □

H.1.6 Proof of Theorem A.8

Proof. The upper inequality is Theorem 4.2, optimized over γ, with the uniform coordinate code length bounded by L_0 and the remaining errors included in Δ^γ_T. The lower inequality is Theorem 5.2 applied to the Bellman–Fano admissible prior, followed by Yao's principle because μ is supported on Ω_0. For the final comparison, set

$$a = \max\left\{1, \frac{L_0}{E^\star_{\mu,r}}\right\}.$$

Since U_T is nondecreasing and $aE^\star_{\mu,r} \ge L_0$, Assumption A.4 and (78) give

$$\mathsf{U}_T(L_0) \le \mathsf{U}_T(aE^\star_{\mu,r}) \le C_{\mathrm{gr}} a^\beta \mathsf{U}_T(E^\star_{\mu,r}) \le C_{\mathrm{gr}} C_{\mathrm{base}} a^\beta T r.$$

Combining this with $\mathfrak{R}^\star_T(\Omega_0) \ge Tr/2$ gives the regret-ratio bound. The finite-index statements are Proposition A.3. □

H.2 Indexed AIR/MAIR identities and upper bounds

H.2.1 Proof of Proposition 3.2

Proof. Under the stated assumption on O_0, the mutual-information chain rule gives

$$I_\mu(Y; H_T) = \sum_{t=1}^{T} I_\mu(Y; A_t, O_t \mid H_{t-1}).$$

Given H_{t-1}, the action A_t is sampled by the algorithm using only randomization independent of the environment and is conditionally independent of Y. Therefore

$$I_\mu(Y; A_t, O_t \mid H_{t-1}) = I_\mu(Y; O_t \mid H_{t-1}, A_t).$$

Conditioning further on $A_t = a$, the law of O_t given $Y = y$ is $P^y_{S_t,p_t,a}$, while the posterior predictive law is $P_{S_t,p_t,a}$. Hence

$$I_\mu(Y; O_t \mid H_{t-1}, A_t = a) = \int D_{\mathrm{KL}}(P^y_{S_t,p_t,a} \| P_{S_t,p_t,a}) q_{S_t}(dy).$$

Averaging over $a \sim p_t(\cdot \mid H_{t-1})$ and over the marginal law of the history under the Bayesian mixture law proves the result. Writing this marginal law as the reference history law gives the displayed expression. □

H.2.2 Proof of Lemma 3.3

Proof. For fixed s, p, a, write the joint mixture of (Y, O) under ν as $w_{y,o}$ and the observation mixture as $w_o = \sum_y w_{y,o}$. In the dominated case, the sums below are interpreted as integrals with respect to the fixed dominating measures. The information and reference term can be written as

$$I_\nu(Y; O \mid s, p, a) + D_{\mathrm{KL}}(q_\nu \| q) = \sum_{y,o} w_{y,o} \log \frac{w_{y,o}}{w_o q(y)}.$$

Differentiating with respect to the mass at (ω, y) gives

$$\mathbb{E}_{O \sim P_{\omega,t}(\cdot \mid s,p,a)} \log \frac{q^+_\nu(y \mid s, p, a, O)}{q(y)},$$

because the $+1$ terms from the numerator and denominator cancel. Averaging over $a \sim p$ and adding the derivative of the linear loss term proves (13). Pairing the gradient with $\delta_{(\omega,y)} - \nu$ cancels the posterior-averaged terms and gives (14). □

H.2.3 Proof of Theorem 3.4

Proof. Condition on H_{t-1}. By the update rule (15),

$$\mathbb{E}\left[\log \frac{q_{t+1}(y^\star)}{q_t(y^\star)} \middle| H_{t-1}\right] = \mathbb{E}_{a \sim p_t,\, O \sim P_{\omega^\star,t}(\cdot \mid S_t,p_t,a)} \log \frac{q^+_{\nu_t}(y^\star \mid S_t, p_t, a, O)}{q_t(y^\star)}.$$

Combining this equality with (14) gives the one-step identity between expected loss, bracket, and log-ratio increment. Summing over t telescopes the logarithm. □

H.2.4 Proof of Lemma C.1

Proof. The derivative and the first two specializations are the environment-index specialization of Lemma 3.3. For the square-root claim, fix $\bar{a}$ and write

$$Z_{\bar{a}}(\bar{o}) := \int \sqrt{p_{\omega',\bar{a}}(\bar{o})}\, \rho(d\omega').$$

Then

$$\log \frac{d\mu^{1/2}_{\bar{a},\bar{o}}}{d\rho}(\omega^\star) = \frac{1}{2} \log p_{\omega^\star,\bar{a}}(\bar{o}) - \log Z_{\bar{a}}(\bar{o}).$$

Under $\bar{o} \sim P_{\omega^\star,s,p,\bar{a}}$, Jensen's inequality gives

$$\mathbb{E} \log \frac{d\mu^{1/2}_{\bar{a},\bar{o}}}{d\rho}(\omega^\star) = -\mathbb{E} \log \frac{Z_{\bar{a}}(\bar{o})}{\sqrt{p_{\omega^\star,\bar{a}}(\bar{o})}}$$

$$
\begin{aligned}
&\geq -\log \int \sqrt{p_{\omega^\star,\bar{a}}(o)} Z_{\bar{a}}(o) d\nu_{s,p,\bar{a}}(o) \\
&= -\log\left(1 - \mathbb{E}_{\omega\sim\rho} h^2(P_{\omega^\star,s,p,\bar{a}}, P_{\omega,s,p,\bar{a}})\right) \\
&\geq \mathbb{E}_{\omega\sim\rho} h^2(P_{\omega^\star,s,p,\bar{a}}, P_{\omega,s,p,\bar{a}}).
\end{aligned}
$$

Averaging over $\bar{a} \sim p$, substituting into (82), and dropping the nonnegative KL term proves (86). □

H.2.5 Proof of Theorem 4.1

Proof. Taking the conditional expectation of $\gamma \log(1/q_{t+1}(y^\star)) - \gamma \log(1/q_t(y^\star))$ under $A \sim p$ and $O \sim P_{\omega^\star,t}(\cdot \mid s, p, A)$, with $q_{t+1} = q_\nu^+(\cdot \mid s, p, A, O)$, gives $-\gamma \mathsf{G}_\chi(s, p, \nu; \omega^\star)$. Substituting the bracket inequality into the telescoping identity (26) proves (27).

□

H.2.6 Proof of Theorem 4.2

Proof. By sure calibration, the fixed truth is one of the environments in the robust supremum at every reached state. Therefore the displayed approximate minimization implies almost surely

$$
\ell_{\omega^\star}(S_t, p_{u_t}) + \mathbb{E}_{\omega^\star}[W_{t+1}^\gamma(S^+) \mid S_t, u_t] - W_t^\gamma(S_t) - \gamma \mathsf{G}_\chi(S_t, u_t; \omega^\star) \leq \varepsilon_t.
$$

Apply the log-potential telescope (26) with $V_t = W_t^\gamma$. Since $W_{T+1}^\gamma = 0$ and the terminal index log loss is nonnegative for a finite or countable index, the result follows. □

H.3 Bellman controls and quantile-index lower bounds

H.3.1 Proof of Theorem 4.4

Proof. Rearrange the displayed inequality and telescope U_t^γ. If O_0 is deterministic or independent of Y, Proposition 3.2 identifies $\mathbb{E}\sum_t \mathcal{I}_\chi(S_t, p_t)$ with $I_\mu(Y; H_T)$. If O_0 is informative, the sum is instead $I_\mu(Y; H_T \mid O_0) \leq I_\mu(Y; H_T)$. In either case $I_\mu(Y; H_T) \leq H_\mu(Y)$ gives the displayed bound. The equivalence with (36) follows from the posterior entropy-drop identity. □

H.3.2 Proof of Lemma D.1

Proof. The optimism/calibration hypothesis (87) and Young's inequality give

$$
c_{\mathrm{opt}} \beta_t \sigma_t(a_t) \leq \frac{c_{\mathrm{opt}}^2 \beta_t^2}{4\gamma} + \gamma \sigma_t^2(a_t).
$$

This proves (88). Subtracting $\gamma J_\chi(S_t, \delta_{a_t}; \omega^\star)$ from both sides and using the definition of $B_{\chi,\gamma}$ proves (89). The final sentence is the result of summing the displayed inequality over time. □

H.3.3 Proof of Lemma D.2

Proof. The AIR gradient calculation in Lemma 3.3 identifies the fixed-truth bracket as the first-order value of $\mathfrak{A}_{q_t,\gamma}$ in the direction $\delta_{(\omega^\star,y^\star)} - \bar{\nu}_t$. Concavity and optimality of $\bar{\nu}_t$ make that first-order correction nonpositive for every admissible comparison direction. Hence the bracket is no larger than $\mathfrak{A}_{q_t,\gamma}(s, p_t, \bar{\nu}_t)$, which is at most the robust value and therefore at most ε_t. The regret bound is the fixed-truth log-potential telescope. □

H.3.4 Proof of Theorem 5.1

Proof. Consider the true joint law of (Y, H_T) and the ghost law of (Y, H'_T). Both have the same marginal law of Y, while Y and H'_T are independent under the ghost law. Hence

$$D_{\mathrm{KL}}\big(\mathcal{L}(Y, H_T) \| \mathcal{L}(Y, H'_T)\big) = I_\mu(Y; H_T).$$

Data processing applied to the indicator of the event $\{\bar{L}_\chi(Y, \cdot) \le r\}$ gives

$$\mathrm{kl}(q_r^{\mathsf{Alg}}, p_r^{\mathsf{Alg}}) \le I_\mu(Y; H_T).$$

Proposition 3.2 identifies the right-hand side with $C_\chi^{\mathsf{Alg}}(\mu) = T\bar{C}_\chi^{\mathsf{Alg}}(\mu)$, proving (40). The definition of ψ gives $q_r^{\mathsf{Alg}} \le \psi(p_r^{\mathsf{Alg}}, T\bar{C}_\chi^{\mathsf{Alg}})$. By nonnegativity of the cumulative indexed loss and by (4),

$$\begin{aligned}
\mathbb{E}_{\Omega, H_T} L_\Omega(H_T) &= \mathbb{E}_{Y, H_T} L_\chi(Y, H_T) \\
&\ge Tr\, \mathbb{P}\{\bar{L}_\chi(Y, H_T) > r\} \\
&= Tr(1 - q_r^{\mathsf{Alg}}),
\end{aligned}$$

which proves (41). If (42) holds, then $q_r^{\mathsf{Alg}} \le 1/2$, giving (43). Finally,

$$\mathrm{kl}\left(\frac{1}{2}, p\right) = \frac{1}{2}\log\frac{1}{4p(1-p)} \ge \frac{1}{2}\log\frac{1}{p} - \log 2,$$

so (44) implies (42). □

H.3.5 Proof of Theorem 5.2

Proof. For every algorithm, (45) gives $T\bar{C}_\chi^{\mathsf{Alg}}(\mu) = C_\chi^{\mathsf{Alg}}(\mu) \le \overline{C}_1(s_1)$, and (46) gives $p_r^{\mathsf{Alg}}(\mu, \chi) \le e^{-\underline{\mathcal{E}}_1^r(\mu,\chi)}$. Substitute these two inequalities into Theorem 5.1, using the monotonicity of ψ in both arguments. The minimax statement follows from Yao's principle. □

H.4 Finite-marginal kernel-bandit upper bounds

H.4.1 Proof of Theorem 6.2

Proof. Let $x^\star \in \arg\max_{x \in \mathcal{X}} f^\star(x)$. On $\mathcal{E}_{\mathrm{GP}}$, calibration at $x^\star$ and the optimistic choice of X_t give

$$f^\star(x^\star) \le m_t(x^\star) + w_t(x^\star) \le m_t(X_t) + w_t(X_t).$$

Calibration at X_t gives $m_t(X_t) \le f^\star(X_t) + w_t(X_t)$. Hence

$$f^\star(x^\star) - f^\star(X_t) \le 2w_t(X_t) = 2\beta_t s_t(X_t) \le 2\beta_T s_t(X_t).$$

Summing and applying Cauchy–Schwarz yields

$$\mathrm{Reg}_T(f^\star) \le 2\beta_T \sqrt{T \sum_{t=1}^T s_t^2(X_t)}.$$

For $u \in [0, \kappa^2/\lambda]$, monotonicity of $u/\log(1+u)$ on a bounded interval gives

$$\lambda u \le \frac{\kappa^2}{\log(1 + \kappa^2/\lambda)} \log(1+u).$$

Applying this with $u = s_t^2(X_t)/\lambda$ and using (57) proves (64). The expectation bound adds the trivial regret bound T on $\mathcal{E}_{\mathrm{GP}}^c$ and uses Jensen's inequality. □

H.4.2 Proof of Theorem 6.1

Proof. Let $\Theta = \Theta_{\mathcal{R}} \subset [-L, L]^{\mathcal{R}}$ be the compact class of retained reward vectors. To avoid a discontinuous tie-breaking graph in the saddle argument, use the closed optimal relation

$$\widetilde{\Theta} := \{(\theta, y) : \theta \in \Theta,\ y \in \arg\max_{x \in \mathcal{R}} \theta_x\}.$$

This is a compact subset of $\Theta \times \mathcal{R}$. Let $\mathcal{Y}_\Theta$ be the set of actions that occur as the second coordinate of some pair in $\widetilde{\Theta}$, and put $K_0 = |\mathcal{Y}_\Theta| \le K$. The fixed tie-breaking rule used for the true index selects an element of this relation. Candidate pair beliefs range over the weakly compact convex set $\Delta(\widetilde{\Theta})$.

Fix $q \in \operatorname{int} \Delta(\mathcal{Y}_\Theta)$ and $\gamma > 0$. For $\nu \in \Delta(\widetilde{\Theta})$ and $p \in \Delta(\mathcal{R})$, write

$$\mathfrak{A}_{q,\gamma}(p, \nu) = \Delta_\nu(p) - \gamma \mathcal{I}_\nu(Y; Z_A \mid A) - \gamma D_{\mathrm{KL}}(\nu_Y \| q),$$

where $A \sim p$ is independent of $(\theta, Y) \sim \nu$, $Z_A = \theta_A + \xi$ with $\xi \sim N(0, \lambda)$, and

$$\Delta_\nu(p) := \mathbb{E}_{\nu,p}[\theta_Y - \theta_A].$$

The loss is affine in each of p and ν separately. Moreover,

$$\mathcal{I}_\nu(Y; Z_A \mid A) + D_{\mathrm{KL}}(\nu_Y \| q) = \sum_{a \in \mathcal{R}} p(a) D_{\mathrm{KL}}\left(P^\nu_{Y, Z_a} \,\middle\|\, q \otimes P^\nu_{Z_a}\right).$$

The map $\theta \mapsto N(\theta_a, \lambda)$ is weakly continuous, and all its means range over a compact interval. Joint convexity and lower semicontinuity of relative entropy therefore make $\mathfrak{A}_{q,\gamma}$ concave and upper semicontinuous in ν and affine and continuous in p. Both feasible sets are compact, so Sion's theorem gives

$$\inf_p \sup_\nu \mathfrak{A}_{q,\gamma}(p, \nu) = \sup_\nu \inf_p \mathfrak{A}_{q,\gamma}(p, \nu). \tag{119}$$

We bound the right-hand side. Fix ν, let $r = \nu_Y$, take $p = r$ on $\mathcal{Y}_\Theta$ and zero on the remaining actions, and define

$$d_i := \mathbb{E}_\nu[\theta_i \mid Y = i] - \mathbb{E}_\nu \theta_i, \qquad i \in \mathcal{Y}_\Theta,$$

with $d_i = 0$ when $r(i) = 0$. The marginal law of Z_i is $(\lambda + L^2)$-sub-Gaussian about its mean: the Gaussian noise contributes variance proxy λ, while Hoeffding's lemma contributes L^2 for the random mean in $[-L, L]$. The entropy variational formula consequently gives

$$D_{\mathrm{KL}}(P^\nu_{Z_i \mid Y=i} \| P^\nu_{Z_i}) \ge \frac{d_i^2}{2(\lambda + L^2)}.$$

It follows that

$$\mathcal{I}_\nu(Y; Z_A \mid A, p = r) \ge \frac{1}{2(\lambda + L^2)} \sum_{i \in \mathcal{Y}_\Theta} r(i)^2 d_i^2,$$

$$\Delta_\nu(r) = \sum_{i \in \mathcal{Y}_\Theta} r(i) d_i.$$

By Cauchy–Schwarz,

$$\Delta_\nu(r)^2 \le 2(\lambda + L^2) K_0 \, \mathcal{I}_\nu(Y; Z_A \mid A, p = r). \tag{120}$$

Dropping the nonpositive term $-\gamma D_{\mathrm{KL}}(r\|q)$ and maximizing $\sqrt{2(\lambda+L^2)K_0 z}-\gamma z$ over $z\ge 0$ now yields

$$\inf_p \sup_\nu \mathfrak{A}_{q,\gamma}(p,\nu) \le \frac{(\lambda+L^2)K_0}{2\gamma} \le \frac{(\lambda+L^2)K}{2\gamma} =: c_\gamma. \tag{121}$$

Take a saddle pair $(p_q, \bar\nu_q)$ in (119) and update q by the posterior optimal-action marginal generated by $\bar\nu_q$. The identity

$$-\mathcal{I}_\nu(Y; Z_A \mid A) - D_{\mathrm{KL}}(\nu_Y\|q) = H_\nu(Y \mid A, Z_A) + \mathbb{E}_\nu \log q(Y)$$

shows that a maximizing belief assigns positive mass to every attainable index. Indeed, suppose $\bar\nu_{q,Y}(y)=0$, choose $(\theta^y, y)\in\widetilde\Theta$, and put

$$\nu_\epsilon = (1-\epsilon)\bar\nu_q + \epsilon\delta_{(\theta^y,y)}.$$

Let B indicate the new mixture component and let $h(\epsilon)$ be binary entropy. Since the new label did not occur under $\bar\nu_q$,

$$\begin{aligned} H_{\nu_\epsilon}(Y \mid A, Z_A) &= (1-\epsilon)H_{\bar\nu_q}(Y \mid A, Z_A) + H(B \mid A, Z_A), \\ H(B \mid A, Z_A) &\ge h(\epsilon) - \epsilon \sup_{a\in\mathcal{R}} D_{\mathrm{KL}}\Big(N(\theta_a^y,\lambda)\,\Big\|\, P_{Z_a}^{\bar\nu_q}\Big) \\ &\ge h(\epsilon) - \frac{2L^2}{\lambda}\epsilon. \end{aligned}$$

The last step uses convexity in the second argument of KL and the uniform pairwise Gaussian bound. Thus the conditional-entropy contribution increases by $\epsilon\log(1/\epsilon) - O(\epsilon)$, while the loss and $\mathbb{E}\log q(Y)$ change by only $O(\epsilon)$. This contradicts maximality. Hence the generated posterior coordinate remains positive for every feasible fixed-truth index after every observation.

Because $\lambda>0$, the predictive mixture densities are strictly positive. Bounded Gaussian means and the preceding KL bound also make the relevant log likelihood ratios integrable. The one-sided Gateaux derivative toward every $\delta_{(\theta,y)}-\bar\nu_q$ therefore exists, so Lemma 3.3 applies to this dominated continuous-belief problem.

At the maximizing belief, the directional derivative of the concave AIR functional toward any $(\theta,y)\in\widetilde\Theta$ is nonpositive. Lemma 3.3 and (121) therefore give

$$\Delta_\theta(p_q) - \gamma\mathbb{E}_\theta \log \frac{q^+(y \mid A, Z_A)}{q(y)} \le c_\gamma \qquad \text{for every } (\theta,y)\in\widetilde\Theta. \tag{122}$$

Thus $\overline{W}_t = (T-t+1)c_\gamma$ is a supersolution of the finite-marginal action-index Bellman recursion: at every state, insert the robust AIR saddle control in the Bellman bracket. Take the comparison set to be the full class Θ at every state and restrict the admissible AIR updates to those that remain positive on every attainable index. Backward induction gives $W_t^\gamma \le \overline{W}_t$. A measurable approximate selector of the exact W recursion and Theorem 4.2 therefore give the bound below, up to arbitrarily small optimization error. The robust saddle policy itself satisfies the same estimate: with continuation $V_t = \overline{W}_t$, inequality (122) makes its Bellman bracket nonpositive, so Theorem 4.1 applies directly. In either case, the uniform initial law on $\mathcal{Y}_\Theta$ gives

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \operatorname{Reg}_T^{\mathcal{R}} \le Tc_\gamma + \gamma\log K_0 \le \frac{(\lambda+L^2)KT}{2\gamma} + \gamma\log K.$$

Standard measurable-selection results apply to the compact saddle problem; equivalently, measurable approximate saddles may be used and their arbitrarily small optimization errors added to the display.

Finally, for a policy confined to $\mathcal{R}$,

$$\sup_{x\in\mathcal{X}_{\text{all}}} f(x) - f(X_t) \le \varepsilon_{\mathcal{R}} + \max_{x\in\mathcal{R}} f(x) - f(X_t).$$

Summing proves (60). Optimizing in γ proves (61). □

H.5 Fixed-kernel separations

H.5.1 Proof of Theorem 6.3

Proof. Choose increasing integers T_m recursively, starting with T_1 sufficiently large. Put

$$a_m = T_m^{-\alpha}, \qquad N_m = \lceil 4T_m \rceil, \qquad M_{m-1} = \sum_{\ell<m} N_\ell,$$

and choose T_m sufficiently large that

$$(M_{m-1}+2)\log(1+T_m) \le a_m^2 T_m = T_m^{1-2\alpha}. \tag{123}$$

This is possible because $1-2\alpha > 0$. Let

$$\mathcal{X} = \{x_0, x_h\} \cup \bigcup_{m\ge1} B_m, \qquad B_m = \{c_{m,1}, \dots, c_{m,N_m}\},$$

and define a diagonal kernel by

$$k(x_h, x_h) = 1, \qquad k(c_{m,j}, c_{m,j}) = a_m^2,$$

with every other entry zero. Let $D = \bigcup_{m\ge1} B_m$ have the discrete topology and endow $D \cup \{x_0\}$ with its one-point compactification: a neighborhood base at x_0 is

$$U_F = \{x_0\} \cup (D \setminus F), \qquad F \subset D \text{ finite.}$$

Adjoin x_h as an isolated point. The resulting $\mathcal{X}$ is compact, metrizable, and countable, with x_0 as its only accumulation point. Since T_m is strictly increasing, $a_m \downarrow 0$. Every point other than x_0 is isolated. If $z_r \to x_0$, then the block index of z_r tends to infinity and hence $k(z_r, z_r) \to 0$. All off-diagonal values are zero, so the same observation verifies continuity at (x_0, x_0) and at every (x_0, x) and (x, x_0). Thus k is jointly continuous on $\mathcal{X}\times\mathcal{X}$ and bounded by one. Its unit RKHS ball is

$$\mathcal{F} = \left\{ f : f(x_0) = 0, \quad f(x_h)^2 + \sum_{m\ge1} a_m^{-2} \sum_{j=1}^{N_m} f(c_{m,j})^2 \le 1 \right\}. \tag{124}$$

For the linear-bandit interpretation, take

$$\mathcal{H} = \ell_2\big(\{h\} \cup \{(m,j) : m \ge 1,\ j \in [N_m]\}\big)$$

with its canonical orthonormal basis and define

$$\varphi(x_0) = 0, \qquad \varphi(x_h) = e_h, \qquad \varphi(c_{m,j}) = a_m e_{m,j}.$$

Then $k(x,x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$, and (124) is the unit ball of parameters $\theta \in \mathcal{H}$ under the identification $f(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$. This makes the glued theorem one fixed infinite-dimensional Hilbert-space linear-bandit instance.

Fix the horizon T_m and put

$$\mathcal{R}_m := \{x_0, x_h\} \cup \bigcup_{\ell<m} B_\ell, \qquad K_m := |\mathcal{R}_m| = M_{m-1} + 2.$$

The one-hot alternatives supported on B_m and the coupling argument from Theorem E.2 give the lower bound $3a_mT_m/4$. For the upper bound, run a minimax finite-armed sub-Gaussian policy on $\mathcal{R}_m$. Its regret relative to the best retained action is at most $C\sqrt{K_mT_m}$. Every point in the ignored blocks B_ℓ, $\ell \geq m$, has $|f(x)| \leq a_\ell \leq a_m$, while the retained anchor has value zero. By (123), the full regret is therefore at most

$$C\sqrt{K_mT_m} + a_mT_m \leq C'a_mT_m.$$

This proves (68).

We next define one AIR Bellman policy for the entire time axis. Set $T_0 = 0$ and $L_m = T_m - T_{m-1}$. During the predetermined epoch $(T_{m-1}, T_m]$, the policy discards data from earlier epochs, restricts its actions to $\mathcal{R}_m$, initializes the optimal-action reference uniformly, and runs the L_m-round finite-marginal action-index AIR Bellman selector from Theorem 6.1 with

$$\gamma_m = \sqrt{\frac{K_mL_m}{\log K_m}}.$$

The control invoked here is the finite-marginal AIR/AMS saddle of Xu and Zeevi (2025). Its connection to Lattimore and György (2021) is the EBO optimization principle described after Theorem 6.1; the proof does not use EBO's functional-estimator implementation. The epoch schedule and all its controls are fixed in advance, so this is one nonanticipating policy and does not take the evaluation horizon as input. The projection of the unit RKHS ball onto $\mathcal{R}_m$ is a compact ellipsoid contained in $[-1, 1]^{K_m}$, and the approximation error of $\mathcal{R}_m$ is at most a_m. Therefore the regret accumulated during epoch m is at most

$$a_mL_m + 2\sqrt{K_mL_m \log K_m} \leq 3a_mT_m,$$

where the last inequality uses

$$K_m \log K_m \leq K_m \log(1 + T_m) \leq a_m^2T_m.$$

Here the first inequality follows because the scale condition itself implies $K_m \leq T_m$ for all sufficiently large m. The RKHS envelope bounds per-round regret by two. For $m \geq 2$, moreover, $K_m \geq N_{m-1} \geq 4T_{m-1}$, so the regret before the current epoch is at most

$$2T_{m-1} \leq \frac{K_m}{2} \leq \frac{a_m^2T_m}{2\log(1 + T_m)} \leq a_mT_m.$$

Consequently, the single epochwise AIR Bellman policy has worst-case regret at most $4a_mT_m = O(T_m^{1-\alpha})$ at time T_m. This proves (69).

Let $\overline{\Gamma}_s$ denote the maximal information gain of this fixed kernel. For $s \leq T_m$, selecting distinct points of B_m and using $\log(1 + u) \geq u/2$ gives

$$\frac{sa_m^2}{4} \leq \overline{\Gamma}_s \leq \frac{M_{m-1} + 1}{2} \log(1 + s) + \frac{a_m^2 s}{2} \leq a_m^2T_m. \tag{125}$$

Indeed, the hub and each coordinate in the earlier blocks contribute at most $\frac{1}{2}\log(1+s)$, while all coordinates in B_ℓ, $\ell \geq m$, contribute in total at most $a_m^2 s/2$. Use the anytime multiplier

$$\overline{\beta}_t = 1 + \sqrt{2\{\overline{\Gamma}_{t-1} + 2\log(t \vee 2)\}}.$$

At round t, the fixed-kernel GP–UCB rule selects a maximizer of $m_t(x) + \overline{\beta}_t s_t(x)$. This is one policy, independent of the evaluation horizon. After every finite history, $m_t(x)$ and $s_t^2(x)$ are finite algebraic combinations of the continuous kernel sections $k(x, X_r)$ and $k(x, x)$. Hence $m_t + \overline{\beta}_t s_t$ is continuous on compact $\mathcal{X}$ and attains its maximum. The lower bound in (125) makes the score of an unqueried point of B_m at least $a_m^2\sqrt{(t-1)/2}$, which diverges uniformly for $t \geq T_m/2$. For $t \leq T_m$, the upper bound and (123) give $\overline{\beta}_t \leq C a_m \sqrt{T_m}$.

Finally take the fixed hub truth $f^\dagger(x_h) = 1/4$ and zero elsewhere. Repeat the Gaussian-noise-stack event and last-half argument in the proof of Theorem E.2. For every $t \leq T_m$, an unqueried point of B_m remains available because $N_m \geq 4T_m$. Its score is $a_m\overline{\beta}_t > 0$, whereas the anchor x_0 has score zero; consequently, x_0 is never selected. Hence, if fewer than $T_m/4$ nonanchor, nonhub pulls occur, the hub has been sampled at least $T_m/4 - 1$ times before every round in the last half. On an event of probability at least $1 - T_m^{-1}$ its acquisition score is at most $1/4 + O(a_m) + O(\sqrt{\log T_m / T_m}) < 1/2$. An unqueried point of B_m has score larger than one for all sufficiently large m. Thus every round in the last half must select a cloud action, a contradiction. At least $T_m/4$ pulls therefore incur regret $1/4$, proving (70) and the ratio claim for (66).

For the original schedule (67), the same argument applies with only two scale estimates changed. The distinct-set information gain obeys the same bounds needed here:

$$\frac{s a_m^2}{4} \leq \Gamma_s^{\mathrm{SKKS}}(k) \leq \frac{M_{m-1}+1}{2}\log 2 + \frac{a_m^2 s}{2} \leq a_m^2 T_m, \qquad s \leq T_m.$$

The lower bound selects s distinct points of B_m; the upper bound follows because a set uses every earlier coordinate at most once and all later diagonal values are at most a_m^2. Hence, uniformly for $T_m/2 \leq t \leq T_m$,

$$a_m b_t^{\mathrm{SKKS}} \geq c\, a_m^2 \sqrt{T_m}\, \log^{3/2} T_m \longrightarrow \infty,$$

because $\alpha < 1/4$. Uniformly for $t \leq T_m$, it also gives

$$b_t^{\mathrm{SKKS}} \leq C\{1 + a_m \sqrt{T_m}\log^{3/2} T_m\}.$$

After at least $T_m/4 - 1$ hub pulls, the corresponding hub bonus is

$$O\Big(T_m^{-1/2} + a_m \log^{3/2} T_m\Big) = o(1),$$

while an unqueried point of B_m has score larger than one. The same noise-stack event and contradiction force at least $T_m/4$ cloud pulls. This proves (70) and the ratio claim for (67) as well. □

H.5.2 Proof of Corollary E.7

Proof. Equation (108) is Wang and Zhang (2026, Theorem 1) after translating their squared exploration coefficient into b_t. The minimax lower bound is due to Scarlett et al. (2017); matching upper bounds under the standard compact-domain regularity assumptions are achieved by localized kernel-bandit algorithms such as GP–ThreDS (Salgia et al., 2021). For fixed kernel parameters, the auxiliary Hölder condition in the latter result follows uniformly from the reproducing property and the Matérn kernel increment bound, while its range condition follows from the uniform RKHS envelope. Its high-probability guarantee gives the displayed expectation rate by taking the failure probability polynomially small and bounding regret on failure by that envelope. Dividing (108) by (109) proves the polynomial ratio. □

Acknowledgements

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

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado P. van Hasselt, and Satinder P. Singh. A definition of continual reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Peter Auer. Using confidence bounds for exploitation–exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Andrew R. Barron and Thomas M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

Andrew G. Barto, Richard S. Sutton, and Christopher J. C. H. Watkins. Learning and sequential decision making. Technical report, University of Massachusetts, Amherst, MA, 1989.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Fan Chen, Dylan J. Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. Assouad, Fano, and Le Cam with interaction: A unifying framework for lower bounds, 2024. arXiv:2410.05117.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making, 2021. arXiv: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*, 2022.

Dylan J. Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient, 2023. arXiv:2301.08215.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Tor Lattimore and András György. Mirror descent and the information ratio. In *Proceedings of the 34th Annual Conference on Learning Theory*, 2021.

Tor Lattimore and Csaba Szepesvári. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Shaojie Li and Yunbei Xu. Pointwise generalization in deep neural networks, 2026. arXiv:2605.18598.

Michael L. Littman. *Algorithms for Sequential Decision-Making*. PhD thesis, Brown University, 1996.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Decision making in hybrid environments: A model aggregation approach. In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, 2025.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. An improved model-free decision-estimation coefficient with applications in adversarial MDPs. In *Proceedings of the International Conference on Learning Representations*, 2026.

Arnab Maiti, Yunbei Xu, and Kevin Jamieson. On the power of adaptivity for ε-best arm identification in linear bandits. In *Proceedings of the 39th Annual Conference on Learning Theory*, 2026.

Alexander Rakhlin and Karthik Sridharan. BISTRO: An efficient relaxation-based method for contextual bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. In *Advances in Neural Information Processing Systems*, 2012.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, 2014.

Sudeep Salgia, Sattar Vakili, and Qing Zhao. A domain-shrinking based Bayesian optimization algorithm with order-optimal regret performance. In *Advances in Neural Information Processing Systems*, 2021.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Proceedings of the 30th Annual Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1723–1742, 2017.

David Silver and Richard S. Sutton. Welcome to the era of experience, 2025. Preprint of a chapter to appear in *Designing an Intelligence*, MIT Press.

Yanglei Song and Meng Zhou. Truncated LinUCB for stochastic linear bandits, 2025. arXiv:2202.11735v4.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Sattar Vakili, Jonathan Scarlett, and Tara Javidi. Open problem: Tight online confidence intervals for RKHS elements. In *Proceedings of the 34th Annual Conference on Learning Theory*, 2021.

Wenjia Wang and Xiaowei Zhang. On the suboptimality of GP–UCB under polynomial effective optimism, 2026. arXiv:2312.01386v2.

Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. On the sublinear regret of GP–UCB. In *Advances in Neural Information Processing Systems*, 2023.

Yunbei Xu. Pointwise complexity for Gaussian fields: Upper envelopes, algorithmic lower bounds, and separation, 2026. arXiv:2606.07931.

Yunbei Xu and Assaf Zeevi. Bayesian design principles for frequentist sequential learning. *Journal of the ACM*, 72(5):Article 37, 2025.

Yuhong Yang and Andrew R. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.