WHERE THE WORST CASE LEAVES THE CORNER: THE BAYES-OPTIMAL AND REGULARISED GREEDY POLICIES ON TWO BERNOULLI ARMS AT SMALL HORIZONS

KOREA SUPERINTELLIGENCE LABS

Abstract. For a policy on two Bernoulli arms whose decisions are comparisons of rational numbers, the expected regret at horizon T on the instance with means (p, q) is a piecewise polynomial $R_T(p, q)$ with rational coefficients, so its worst case over the instance square $[0, 1]^2$ is a well-defined number. Three finite invariants of such a policy were introduced for Thompson sampling in a companion paper: the corner-departure horizon T_c, the smallest horizon at which a deterministic instance is no longer a worst case; the edge-departure horizon T_e, the smallest horizon at which the worst case leaves the boundary of the square; and the mechanism of the departure, read off the sign of the regret's derivative at the corner along an edge. We compute them, exactly and with machine-checked proofs, for two further policies: the horizon-T Bayes-optimal policy under independent uniform priors (Bellman's dynamic programme), and the regularised greedy rule $\mathrm{RG}(\alpha, \beta)$ of Zhou, Li and Wang at $(\alpha, \beta) = (0, 0), (1, 2), (2, 4)$. (i) The Bayes-optimal policy has corner regret exactly $1/2$ at every horizon $T \le 12$ and $T_c = 3$, two horizons before the Thompson-sampling value 5; at $T = 4$ its exact worst case is the rational $2/3$, attained at $(2/3, 0)$, with $2/3 - R_4(p, 0) = \frac{3}{2}(p - \frac{2}{3})^2$ along the edge. (ii) Regularised greedy has corner regret exactly 1 at $(0, 1)$ at every horizon $T \le 12$, and $T_c = T_e = 7, 10, 11$ for the three parameter pairs: corner departure and edge departure coincide, whereas for Thompson sampling they are 5 and 18. In particular regularisation strictly delays the departure, $7 < 10 < 11$. (iii) The mechanism separates the two classes. For the Bayes-optimal policy the edge derivative at the corner is 0 at $T = 2$ and $-3/4$ at $T = 3$, so the departure is first-order, as for Thompson sampling; for regularised greedy the same derivative, at every $2 \le T \le 14$, is 2^{2-T} for pure greedy and, for the two regularised rules, 1 at $T = 2$ and $1/2$ thereafter — strictly positive throughout, so the corner remains a strict local maximiser along the edge at and past its own departure horizon and no first-order test at the corner can detect it: the worst case jumps to an interior instance. Every upper bound is a Bernstein certificate of polynomial nonnegativity on the square or on one of its edges, except the $T = 4$ bound, which is a polynomial inequality proved from the explicit degree-four regret; every lower bound is an exact rational evaluation; the derivatives are derivatives of the real polynomial; and every statement is machine-checked in Lean 4. UCB1 and KL-UCB, whose decisions compare transcendental indices, are measured in floating point only and appear as labelled computations outside the formal development.

1. Introduction

The objects. Two Bernoulli arms have means $p = \mu_1$ and $q = \mu_2$; the instance space is the closed square $Q = [0, 1]^2$. A policy plays one arm per round for T rounds, and its expected regret on the instance (p, q) is $|p - q|$ times the expected number of pulls of the worse arm. The companion paper [12] studied this quantity for one algorithm, Thompson sampling with independent $\mathrm{Beta}(1, 1)$ priors, and found that at small horizons the worst case over the instance space is exactly computable and behaves in a structured way: the deterministic instance $(1, 0)$ is a worst case for $T \le 4$ and not at $T = 5$; the worst case stays on the edge $q = 0$ for $T \le 17$ and leaves it at $T = 18$; and both departures are first-order events, visible as a sign change of a derivative of the regret at the departing point. The three numbers — the corner-departure horizon T_c, the edge-departure horizon T_e, and the order of the mechanism — are defined for every policy whose decisions are comparisons of rational numbers, and to our knowledge they have been computed for no policy other than Thompson sampling.

This paper computes them for two further policies. The first is the *Bayes-optimal policy* B_T: for a fixed horizon T and independent $\mathrm{Beta}(1, 1)$ priors, the policy that maximises the expected number of successes over the T rounds, given by Bellman's backward induction on the posterior state (Bellman [3]; Berry and Fristedt [4]; Jacko's survey [7] treats exactly this object). The second is the *regularised greedy* rule $\mathrm{RG}(\alpha, \beta)$ of Zhou, Li and Wang [13]: pull each arm once, then play the arm with the larger score

$(S_i+\alpha)/(N_i+\beta)$, ties broken by a fair coin; $(0,0)$ is the classical greedy rule and $(1,2)$ plays the posterior mean under the uniform prior. Both policies are *counts-based*: the decision in a round is a function of the four counters (S_1, F_1, S_2, F_2) of successes and failures observed so far, and it is a comparison of two rational numbers. That is what makes the regret a piecewise polynomial in (p, q) with rational coefficients and the worst case an exactly certifiable number. Thompson sampling and explore-then-commit are counts-based too, and they appear here only as reproduction controls against the certified values of [12]. UCB1 [2] and KL-UCB [6, 5] compare transcendental indices; they are measured in floating point and reported in Section 6 as computations outside the formal development.

One structural difference from [12] governs the whole paper. Thompson sampling is symmetric under $(p, q) \mapsto (q, p)$, so [12] maximises over the triangle $0 \le q \le p \le 1$. Regularised greedy is not: it pulls arm 1 first, so at $T = 1$ its regret is 0 at $(1, 0)$ and 1 at $(0, 1)$ — the two deterministic instances are not interchangeable, and no symmetry is available, or proved here, to halve the instance space. Every bound in this paper is therefore taken over the whole square, and every certificate comes in two halves, one for each branch of the regret. The reference corner of regularised greedy is accordingly $(0, 1)$, and the interior instances that carry its departures lie in the upper triangle $p < q$: at $T = 7$ the witness is $(7/50, 17/20)$.

What is in print. Zhou, Li and Wang [13] introduce the class $\mathrm{RG}(\alpha, \beta)$ and prove, as their Theorem 2, a two-sided finite-horizon regret *envelope*: in their words, "the first finite-horizon regret envelopes for regularized greedy bandits, showing that finite-horizon regret decomposes into transient exploration costs and a suboptimal convergence term that decays exponentially with the regularization strength", where "the small-o terms are taken with respect to the regularization strength rather than the horizon". No exact value at any horizon is computed there, and the source of the e-print contains no occurrence of "worst case", of "worst-case" or of "corner"; the word "minimax" occurs twice, both times in connection with the algorithm MOSS — the expansion of its name and the title of the paper cited for it. Jacko [7] surveys the finite-horizon two-armed bandit with binary responses, casts it as a Markov decision process with Beta updating and evaluates the Bayes-optimal design among many others, in Bayesian and in frequentist performance measures on named instances; nothing there is a maximum over the instance square, and the source of that e-print contains no occurrence of "worst case", of "worst-case" or of "minimax". Kalvit and Zeevi [8] contrast algorithms by their worst-case behaviour, in diffusion scaling, as $T \to \infty$; nothing there is a finite horizon or a corner. We are not aware of any computation, exact or floating-point, of the worst case of the finite-horizon Bayes-optimal policy or of a greedy rule over the instance square at a fixed horizon.

Results. All statements are for two Bernoulli arms. R^{B}_T is the regret of the horizon-T Bayes-optimal policy under independent $\mathrm{Beta}(1,1)$ priors and $R^{\alpha,\beta}_T$ that of $\mathrm{RG}(\alpha, \beta)$, both defined on the whole square (Section 2).

(i) *The Bayes-optimal policy leaves the corner at* 3. $R^{\mathrm{B}}_T(1,0) = R^{\mathrm{B}}_T(0,1) = 1/2$ for every $T \le 12$ (Proposition 3.1); $(1,0)$ is a worst case over the square for $T \le 2$ and not at $T = 3$, where $R^{\mathrm{B}}_3(99/100, 0) = 2029401/4000000 > 1/2$ (Theorems 3.2 and 3.3); so $T_{\mathrm{c}}(\mathrm{B}) = 3$ (Theorem 3.4), against $T_{\mathrm{c}}(\mathrm{TS}) = 5$ in [12]. At $T = 4$ the exact worst case is $2/3$, attained at $(2/3, 0)$, and along the edge $q = 0$ the deficit is a square, $2/3 - R^{\mathrm{B}}_4(p, 0) = \frac{3}{2}(p - \frac{2}{3})^2$ (Theorem 3.5).

(ii) *Regularised greedy leaves the corner and the boundary together.* $R^{\alpha,\beta}_T(0,1) = 1$ for every $T \le 12$ and each of the three parameter pairs (Proposition 4.1); the corner is a worst case over the square at every horizon below 7, 10, 11 respectively (Theorem 4.2 with Proposition 4.1); at that horizon a single interior instance is strictly worse than every instance of the boundary (Theorem 4.3); so $T_{\mathrm{c}} = T_{\mathrm{e}} = 7$, 10, 11 (Theorem 4.4). For $(0,0)$ and $(1,2)$ the boundary bound at the onset horizon is the corner value itself: the corner is still a global maximiser of the regret restricted to the boundary at the very horizon at which it stops being one of the square.

(iii) *Regularisation strictly delays the departure*: $T_{\mathrm{c}}(0,0) = 7 < T_{\mathrm{c}}(1,2) = 10 < T_{\mathrm{c}}(2,4) = 11$ (Theorem 4.5).

(iv) *The mechanism is algorithm-dependent.* Along the edge $q = 0$ the derivative of the regret at $p = 1$ is $1/2$, 0, $-3/4$, $-1, \ldots$ for the Bayes-optimal policy: it is 0 at $T = 2$, the last horizon at which the corner is a worst case, and $-3/4 < 0$ at $T = 3$, so the departure is first-order, as for Thompson

sampling. For regularised greedy the same derivative, at every $2 \le T \le 14$, is 2^{2-T} for $(0,0)$ and, for $(1,2)$ and $(2,4)$, 1 at $T = 2$ and $1/2$ thereafter (Proposition 5.2, Theorem 5.3): strictly positive at, before and after the departure, so the corner remains a strict local maximiser along the edge and no first-order test at the corner detects T_c. We call this departure zeroth-order.

(v) *Reproduction.* The same dynamic programme, run with the Thompson-sampling and explore-then-commit decision rules, reproduces the certified corner ladder, the certified corner-derivative ladder and the exact ETC(1) worst-case values of [12] (Proposition 3.7).

What is *not* claimed deserves equal emphasis. Nothing here concerns horizons beyond those stated (12 for the ladders, 14 for the greedy derivatives), other priors, more than two arms, or asymptotics. The closed form 2^{2-T} is verified for $2 \le T \le 14$, and the constant $1/2$ for $3 \le T \le 14$; neither is proved for all T. The edge-departure horizon of the Bayes-optimal policy is not a theorem here: floating-point search puts it at 8 (Section 6), and the certified statements stop at $T = 4$. The regularisation monotonicity is proved at three parameter pairs, not as a statement about (α, β) in general. The exact worst case of $\mathrm{RG}(\alpha, \beta)$ at its own onset horizon is not enclosed: what is proved is that one rational interior instance beats the whole boundary. The UCB1 and KL-UCB rows of Section 6 are floating-point measurements.

Method. Every upper bound on a square or an edge is a *Bernstein certificate* in the sense of [12]: a polynomial with integer coefficients is proved nonnegative on $[0,1]^2$ by a binary subdivision on whose pieces all Bernstein coefficients are nonnegative, and the polynomial is built from the list of posterior states by the checker's own arithmetic. The one exception is the $T = 4$ bound of Theorem 3.5, where the maximiser $(2/3, 0)$ is a tangential zero at an interior point of an edge — exactly what subdivision cannot certify — and the bound is instead a polynomial inequality proved from the explicit degree-four regret. Every lower bound is an exact rational evaluation. The derivatives are derivatives of the real polynomial in the sense of Mathlib's `HasDerivAt`, assembled by the product rule from the same list of states. The Bayes-optimal policy itself is computed inside the formal development by backward induction on the Bellman value table. Everything is checked in Lean 4 with Mathlib (Section 7); the finite searches that found the certificates and the witnesses are described with each proof.

2. Preliminaries

2.1. **Counts-based policies and the state recursion.** A *state* is a quadruple $s = (S_1, F_1, S_2, F_2)$ of nonnegative integers, the successes and failures observed on each arm; its *level* $S_1 + F_1 + S_2 + F_2$ is the number of rounds played. A *counts-based policy* is a map π from states to $[0,1] \cap \mathbb{Q}$, where $\pi(s)$ is the probability of playing arm 1 in state s. For the deterministic policies of this paper $\pi(s) \in \{0, \frac{1}{2}, 1\}$, the value $\frac{1}{2}$ being a fair coin at an exact tie.

Lemma 2.1. *For every counts-based policy π and every state s there is a rational $c_s \ge 0$, independent of (p,q), such that the probability of being in state s after* level(s) *rounds on the instance (p,q) is $c_s\, p^{S_1}(1-p)^{F_1} q^{S_2}(1-q)^{F_2}$; namely $c_{(0,0,0,0)} = 1$ and a state s contributes $c_s\pi(s)$ to each of its two arm-1 successors and $c_s(1-\pi(s))$ to each of its two arm-2 successors. Consequently the expected number of pulls of arm 2 in T rounds is*

$$\mathbb{E}[N_2(T)] = \sum_{\mathrm{level}(s) < T} w_s\, p^{S_1}(1-p)^{F_1} q^{S_2}(1-q)^{F_2}, \qquad w_s = c_s\big(1 - \pi(s)\big), \tag{1}$$

a polynomial in (p,q) with rational coefficients.

Proof. As in [12, Lemma 2.3], with the Thompson comparison probability replaced by $\pi(s)$: each transition multiplies the state probability by $\pi(s)$ or $1-\pi(s)$ and by exactly one of $p, 1-p, q, 1-q$, and the exponents record which. □

The formal development keeps the list of pairs (w_s, s), computed by this forward recursion, and every quantity below is built from that list. In a deterministic policy a branch of weight 0 is dropped, so the list is short: thirteen terms at $T = 4$ for the Bayes-optimal policy.

2.2. **The regret over the square.** Since $N_1(T) + N_2(T) = T$, the expected regret on $(p, q) \in Q$ is

$$R_T(p,q) = \begin{cases} (p-q)\,\mathbb{E}[N_2(T)], & q \le p, \\ (q-p)\,\big(T - \mathbb{E}[N_2(T)]\big), & p < q, \end{cases} \tag{2}$$

with $\mathbb{E}[N_2(T)]$ the polynomial (1). Both branches vanish on the diagonal, so R_T is continuous on Q, and it is a polynomial on each of the two closed triangles. Formula (2) is the definition used throughout, at every real (p, q); on Q it is the expected regret. Since Q is compact the *worst case* $W_T = \max_Q R_T$ is attained, and the points where it is attained are the *worst-case instances*. The *corners* are the two deterministic instances $(1, 0)$ and $(0, 1)$; the *boundary* ∂Q is the union of the four edges $p = 0$, $p = 1$, $q = 0$, $q = 1$; an instance is *interior* if $0 < p < 1$ and $0 < q < 1$.

2.3. **The policies.** *The Bayes-optimal policy* B_T. Under independent Beta$(1, 1)$ priors the posterior mean of arm i in state s is $m_i(s) = (S_i + 1)/(N_i + 2)$ with $N_i = S_i + F_i$. The horizon-T value function is defined by backward induction on the level: $V(s) = 0$ when level$(s) = T$, and otherwise

$$V(s) = \max_{i \in \{1,2\}} v_i(s), \qquad v_i(s) = m_i(s)\big(1 + V(s + \mathrm{succ}_i)\big) + \big(1 - m_i(s)\big)V(s + \mathrm{fail}_i), \tag{3}$$

where $s + \mathrm{succ}_i$ and $s + \mathrm{fail}_i$ add one success, respectively one failure, to arm i. The policy plays arm 1 if $v_1(s) > v_2(s)$, arm 2 if $v_2(s) > v_1(s)$, and a fair coin if $v_1(s) = v_2(s)$. Every quantity in (3) is rational, so B_T is a counts-based policy in the sense above; it is the policy that maximises the expected number of successes in T rounds under the uniform prior, and it depends on T. In the formal development the value table is an array of $(T + 1)^4$ rationals filled level by level from $T - 1$ down to 0, and the policy is read off it. (The empty state is an exact tie, so the first pull is a fair coin.) We write R_T^{B} for its regret (2).

Regularised greedy $\mathrm{RG}(\alpha, \beta)$. We take the rule from Algorithm 1 of [13], whose initialisation reads "Pull each arm once during periods $t = 1, \ldots, K$. Observe rewards $X_1, \ldots, X_K$, set $A_t = t$ for $t = 1, \ldots, K$", and which thereafter forms the scores $\widehat{p}_i(t) = (S_i(t) + \alpha)/(N_i(t) + \beta)$ and "selects an arm with maximal score, breaking ties uniformly at random"; "The classical greedy policy corresponds to $\alpha = \beta = 0$". With $K = 2$: in a state with $N_1 = 0$ play arm 1; in a state with $N_1 > 0$ and $N_2 = 0$ play arm 2; otherwise play arm 1 if $(S_1 + \alpha)/(N_1 + \beta) > (S_2 + \alpha)/(N_2 + \beta)$, arm 2 if the reverse strict inequality holds, and a fair coin at an exact tie. We consider $(\alpha, \beta) \in \{(0, 0), (1, 2), (2, 4)\}$ and write $R_T^{\alpha,\beta}$ for the regret. Because arm 1 is pulled first, $R_1^{\alpha,\beta}(1, 0) = 0$ while $R_1^{\alpha,\beta}(0, 1) = 1$; this asymmetry is why the square, not a triangle, is the instance space of this paper.

The two controls. Thompson sampling TS with Beta$(1, 1)$ priors, in the form of Algorithm 1 of Agrawal and Goyal [1], has $\pi(s) = \mathbb{P}[\theta_1 > \theta_2]$ with $\theta_i \sim \mathrm{Beta}(S_i + 1, F_i + 1)$ independent, a rational number given by the interleaving formula of [12, Lemma 2.1]. Explore-then-commit ETC(m) plays arm 1 for m rounds and arm 2 for m rounds and then commits to the arm with the larger number of successes, by a fair coin at a tie [9, Chapter 6]; it is counts-based because after the exploration phase the committed arm is the one with $N_i > m$, provided that test is made *before* the comparison of successes (Remark 6.3). Both are implemented in the same dynamic programme and serve only to reproduce values certified in [12].

2.4. **The three invariants.** Fix a counts-based policy π and a *reference corner* c_π, which is $(1, 0)$ for the Bayes-optimal policy and $(0, 1)$ for regularised greedy, as in the formal statements.

Definition 2.2. (a) The *corner-departure horizon* $T_{\mathrm{c}}(\pi)$ is the smallest $T \ge 1$ at which c_π is not a worst case: some $(p, q) \in Q$ has $R_T(p, q) > R_T(c_\pi)$.

(b) The *edge-departure horizon* $T_{\mathrm{e}}(\pi)$ is the smallest $T \ge 1$ at which no point of ∂Q is a worst case; since the maximum is attained, this is the smallest T at which some interior instance has strictly larger regret than every boundary instance.

(c) Along the edge $q = 0$ the regret is the one-variable polynomial $p \mapsto R_T(p, 0) = p\,\mathbb{E}[N_2(T)](p, 0)$; write $D_T = \partial_p R_T(1, 0)$ for its derivative at the corner $(1, 0)$. The departure at T_{c} is *first-order* if $D_{T_{\mathrm{c}}} < 0$: then $R_{T_{\mathrm{c}}}(p, 0) > R_{T_{\mathrm{c}}}(1, 0)$ for all $p < 1$ close enough to 1, so the corner is not even a local maximiser along the edge. It is *zeroth-order* if $D_{T_{\mathrm{c}}} > 0$: then $R_{T_{\mathrm{c}}}(p, 0) < R_{T_{\mathrm{c}}}(1, 0)$ for all $p < 1$ close enough to 1, the corner remains a strict local maximiser along the edge, and the instance that beats it lies at positive distance.

Since a worst case at a corner is in particular on the boundary, $T_{\mathrm{c}} \le T_{\mathrm{e}}$ always. For the Bayes-optimal policy both corners have the same regret at every $T \le 12$ (Proposition 3.1), and for regularised greedy at every $2 \le T \le 12$ (Proposition 4.1); so within those ranges the choice of reference corner does not affect T_{c}, and the derivative at $(1,0)$ in (c) speaks about a corner that is a worst case exactly when c_π is. For Thompson sampling, [12] proves $T_{\mathrm{c}} = 5$ and $T_{\mathrm{e}} = 18$, with $D_4 = 31/120 > 0$ and $D_5 = -847/14400 < 0$.

2.5. **Certificates in square coordinates.** We use the Bernstein certificate checker of [12, Section 2.4] unchanged, now with the certificate box equal to the instance square itself, $x_0 = p$ and $x_1 = q$, and no change of coordinates. In the homogeneous (u, v) form of the checker a term $w_s\, p^{S_1}(1-p)^{F_1} q^{S_2}(1-q)^{F_2}$ of (1) is a single monomial, so the list of states *is* the certificate polynomial. Let D be a common denominator of the w_s at horizon T. For integers M_n and $M_d > 0$ the formal development builds the two polynomials

$$G = M_n D - M_d\, D\,(p-q)\,\mathbb{E}[N_2(T)], \qquad H = M_n D - M_d\, D\,(q-p)\,\big(T - \mathbb{E}[N_2(T)]\big),$$

and the generic theorem is:

(S) if the checker accepts a tree for G and a tree for H on $[0,1]^2$, then $R_T \le M_n/M_d$ on the whole square;

(E) if in the build the pair of factors $(q, 1-q)$ is replaced by the constants $(0,1)$ or $(1,0)$, or the pair $(p, 1-p)$ is, and the checker accepts trees for the resulting G and H, then $R_T \le M_n/M_d$ on the corresponding edge $q = 0$, $q = 1$, $p = 0$ or $p = 1$.

Both are instances of one theorem (`reg_le_of_certs` in Appendix A) whose hypotheses say that the second factor of each pair evaluates to one minus the first; the four edge versions cost no new certificate machinery. As in [12], we count the nodes of a subdivision tree as internal nodes plus leaves, so the trivial certificate has one node. The trees were found by iterative deepening over the subdivision depth in an independent Python model of the checker, and only the tree is handed to the formal checker, which rebuilds the polynomial from the dynamic programme and re-runs the test.

3. The Bayes-optimal policy leaves the corner at horizon three

Proposition 3.1. *For every horizon $1 \le T \le 12$, $R^{\mathrm{B}}_T(1,0) = R^{\mathrm{B}}_T(0,1) = \frac{1}{2}$.*

Proof. Exact rational evaluation of (2) at the two corners for each $T \le 12$, with the policy computed from (3) at each horizon; one Boolean over the twenty-four values. The mechanism is visible in the policy: the first pull is a fair coin, and after one observation on a deterministic instance the Bayes-optimal policy never returns to the dead arm, so exactly one pull of it is wasted with probability $\frac{1}{2}$. Contrast the Thompson ladder $\frac{1}{2}, \frac{5}{6}, \frac{25}{24}, \frac{281}{240}, \ldots$ of [12], which grows. The values at $(1,0)$ for $T \le 4$ are additionally stated over $\mathbb{R}$ (`bayes_corner_half`), which is the form the next theorems use. □

Theorem 3.2. *For $T \in \{1, 2\}$ and every $(p, q) \in Q$, $R^{\mathrm{B}}_T(p,q) \le \frac{1}{2}$.*

Proof. (S) at $T = 1$ and $T = 2$ with $M_n/M_d = 1/2$. All four certificates are trivial: every Bernstein coefficient of G and of H is already nonnegative at the polynomial's own multidegree, with no subdivision. □

Theorem 3.3. *At horizon 3 the corner is not a worst case: some $(p, q) \in Q$ has $R^{\mathrm{B}}_3(1,0) < R^{\mathrm{B}}_3(p,q)$. Explicitly,*

$$R^{\mathrm{B}}_3\big(\tfrac{99}{100}, 0\big) = \frac{2029401}{4000000} = 0.50735025 \;>\; \tfrac{1}{2} = R^{\mathrm{B}}_3(1,0).$$

Proof. Two exact rational evaluations of (2); the value $2029401/4000000$ is the one the formal proof compares. The witness is deliberately taken at distance $1/100$ from the corner rather than at the edge maximiser: it shows that the corner loses to instances next to it, which is the certificate half of the first-order classification of Theorem 5.3. (Along the edge, $R^{\mathrm{B}}_3(p,0) = \frac{3}{2}p - \frac{3}{4}p^2 - \frac{1}{4}p^3$, whose maximum on $[0,1]$ is $\frac{3}{2}\sqrt{3} - 2 = 0.598\ldots$ at $p = \sqrt{3} - 1$; this edge polynomial and its maximum are computed outside the formal development and no statement here depends on them.) □

Theorem 3.4. *Horizon* 3 *is the smallest horizon at which the deterministic instance* $(1,0)$ *is not a worst case of the Bayes-optimal policy: for* $T \in \{1,2\}$, $R_T^{\mathrm{B}} \le R_T^{\mathrm{B}}(1,0)$ *on* Q, *and* $R_3^{\mathrm{B}} \le R_3^{\mathrm{B}}(1,0)$ *fails on* Q. *So* $T_{\mathrm{c}}(\mathrm{B}) = 3$, *whereas* $T_{\mathrm{c}}(\mathrm{TS}) = 5$ [12].

Proof. Proposition 3.1 and Theorems 3.2, 3.3. □

At $T = 4$ the list of states of Lemma 2.1 has thirteen terms and the polynomial (1) collapses to a form that can be handled without certificates.

Theorem 3.5. (a) *For all real* p, q, $\mathbb{E}[N_2(4)] = 2 + \frac{3}{2}q - \frac{3}{2}p + pq^2 - p^2q$ *for the Bayes-optimal policy at horizon* 4.

(b) *For every* $(p,q) \in Q$, $R_4^{\mathrm{B}}(p,q) \le \frac{2}{3}$, *and* $R_4^{\mathrm{B}}(\frac{2}{3},0) = \frac{2}{3}$. *So the exact worst case of the horizon-*4 *Bayes-optimal policy is* $W_4 = 2/3$, *attained at* $(2/3,0)$.

(c) *For every real* $p \ge 0$, $\frac{2}{3} - R_4^{\mathrm{B}}(p,0) = \frac{3}{2}\big(p - \frac{2}{3}\big)^2$.

Proof. (a) The thirteen-term list at $T = 4$ is computed by the dynamic programme and equated, term by term, with an explicit list (one Boolean); expanding it gives the stated polynomial (a ring identity). (b) On the triangle $q \le p$ the claim is $(p-q)(2 + \frac{3}{2}q - \frac{3}{2}p + pq^2 - p^2q) \le \frac{2}{3}$ on $[0,1]^2$, and on the triangle $p \le q$ it is $(q-p)(2 - \frac{3}{2}q + \frac{3}{2}p - pq^2 + p^2q) \le \frac{2}{3}$; each is a polynomial inequality of degree four proved by a Positivstellensatz-style combination of the products of the hypotheses $p, 1-p, q, 1-q \ge 0$ and the squares $(3p-2)^2$, $(p-1)^2$, $(p-q)^2$, q^2 (and their mirror images), found by Mathlib's `nlinarith` from those hints. The lower bound is the exact evaluation $R_4^{\mathrm{B}}(2/3,0) = 2/3$. (c) On the edge, by (a), $R_4^{\mathrm{B}}(p,0) = 2p - \frac{3}{2}p^2$, and $\frac{2}{3} - 2p + \frac{3}{2}p^2 = \frac{3}{2}(p - \frac{2}{3})^2$.

Why not a certificate: the maximiser $(2/3,0)$ is a tangential zero of $\frac{2}{3} - R_4$ at an interior point of an edge, and a Bernstein subdivision cannot certify nonnegativity of a polynomial with an interior zero; the tree search fails at depth 8 even for the loose bound $666667/10^6$. This is the one horizon in the table at which the exact worst case of a policy past its corner departure is a rational number, because the maximiser is rational. The polynomial and the value were computed independently in Python before formalisation. □

Proposition 3.6 (controls)**.** *(a)* $R_3^{\mathrm{B}} \le \frac{1}{2}$ *fails on* Q, *so the bound of Theorem 3.2 does not extend to* $T = 3$. *(b) No* $(p,q) \in Q$ *has* $R_4^{\mathrm{B}}(p,q) > \frac{2}{3}$, *so the value of Theorem 3.5 cannot be pushed up. (c) Some* $(p,q) \in Q$ *has* $R_4^{\mathrm{B}}(p,q) > 0$, *and* $R_4^{\mathrm{B}}(p,p) = 0$ *for every real* p.

Proof. (a) is the witness of Theorem 3.3; (b) is the upper half of Theorem 3.5(b); (c) from Proposition 3.1 at $(1,0)$ and the factor $p - q$ in (2). □

Proposition 3.7 (reproduction control)**.** *Run with the Thompson-sampling rule, the dynamic programme of Lemma 2.1 gives*

$$R_T^{\mathrm{TS}}(1,0) = \tfrac{1}{2},\ \tfrac{5}{6},\ \tfrac{25}{24},\ \tfrac{281}{240},\ \tfrac{2007}{1600},\ \tfrac{1983559}{1512000} \quad (T = 1,\dots,6),$$

and, run with the explore-then-commit rule ETC(1)*,*

$$R_5^{\mathrm{ETC}(1)}\big(\tfrac{5}{6},0\big) = \tfrac{25}{24}, \qquad R_6^{\mathrm{ETC}(1)}\big(\tfrac{3}{4},0\big) = \tfrac{9}{8}, \qquad R_7^{\mathrm{ETC}(1)}\big(\tfrac{7}{10},0\big) = \tfrac{49}{40}.$$

The first row is the certified corner ladder of [12, Proposition 3.1]*; the second is the exact worst case* $T^2/(8(T-2))$ *of* ETC(1) *at its maximiser* $p - q = T/(2(T-2))$ [12, Proposition 5.1]*, at* $T = 5, 6, 7$. *The corner-derivative ladder of Thompson sampling is reproduced in Proposition 5.2.*

Proof. Exact rational evaluations, in the formal development, of the same polynomial (2) built from a different decision rule. The model of this paper was written independently of [12]: it takes an arbitrary counts-based policy as an argument and works in square rather than triangle coordinates, so the agreement is a check of the model, not a restatement. (Only the values at the maximisers are evaluated here; that they are the worst cases of ETC(1) is the theorem of [12].) □

4. Regularised greedy leaves the corner and the boundary together

Proposition 4.1. *For each* $(\alpha,\beta) \in \{(0,0),(1,2),(2,4)\}$*:* $R_T^{\alpha,\beta}(0,1) = 1$ *for every* $1 \le T \le 12$*, and* $R_T^{\alpha,\beta}(1,0) = 1$ *for every* $2 \le T \le 12$*.*

Proof. Exact rational evaluation, one Boolean per parameter pair. At a deterministic instance the dead arm is pulled exactly once, in its forced initialisation round, and never again, so the regret is 1 from $T = 2$ on at both corners; at $T = 1$ only arm 1 is pulled, so $R_1^{\alpha,\beta}(0,1) = 1$ while $R_1^{\alpha,\beta}(1,0) = 0$ (the latter is immediate from the definition and is not part of the formal statement, which omits $(1,0)$ at $T = 1$). □

Theorem 4.2. *For every* $(p,q) \in Q$*:* $R_T^{0,0}(p,q) \le 1$ *for* $1 \le T \le 6$*;* $R_T^{1,2}(p,q) \le 1$ *for* $1 \le T \le 9$*;* $R_T^{2,4}(p,q) \le 1$ *for* $1 \le T \le 10$*.*

Proof. (S) with $M_n/M_d = 1$ at each of the $6 + 9 + 10 = 25$ horizons, two certificates per horizon, fifty in all. The trees are trivial at every horizon below 6, 7, 7 for the three parameter pairs respectively; the largest trees are 15 nodes for $(0,0)$ at $T = 6$ and 13 nodes for $(1,2)$ at $T = 9$ and for $(2,4)$ at $T = 10$. All fifty Boolean evaluations run in a module that imports nothing from Mathlib (Section 7). □

Theorem 4.3. (a) RG$(0,0)$ *at* $T = 7$*:* $R_7^{0,0}(p,q) \le 1$ *for every* $(p,q) \in \partial Q$*, and some interior instance has* $R_7^{0,0}(p,q) > 1$*; explicitly* $R_7^{0,0}(\frac{7}{50},\frac{17}{20}) > 1$*.*

(b) RG$(1,2)$ *at* $T = 10$*:* $R_{10}^{1,2}(p,q) \le 1$ *for every* $(p,q) \in \partial Q$*, and* $R_{10}^{1,2}(\frac{27}{100},\frac{19}{25}) > 1$*.*

(c) RG$(2,4)$ *at* $T = 11$*:* $R_{11}^{2,4}(p,q) \le \frac{1011}{1000}$ *for every* $(p,q) \in \partial Q$*, and* $R_{11}^{2,4}(\frac{29}{100},\frac{37}{50}) > \frac{1011}{1000}$*.*

In each case the interior instance is strictly worse than every instance of the boundary.

Proof. Upper bounds by (E), four edges, two certificates per edge: eight per policy, twenty-four in all. For $(0,0)$ all eight are trivial; for $(1,2)$ six are trivial and two have 3 nodes; for $(2,4)$ six are trivial and two have 9 nodes. Lower bounds by exact rational evaluation at the stated points, compared with 1, 1 and 1011/1000. The witnesses come from a floating-point grid search over the square with local refinement, which puts the maximisers near $(0.1436, 0.8509)$ for $(0,0)$ at $T = 7$, near $(0.2661, 0.7560)$ for $(1,2)$ at $T = 10$ and near $(0.2895, 0.7393)$ for $(2,4)$ at $T = 11$ — each returned together with its mirror (q,p), at the same value to the precision of the search — with values $1.020772\ldots$, $1.019204\ldots$ and $1.026155\ldots$ at the three rational witnesses; those decimals are evidence about the location, not part of the theorem. For $(2,4)$ the bound 1 is false on the boundary at $T = 11$ (Proposition 4.6(b)), which is why the boundary bound there is 1011/1000. □

Theorem 4.4. *For* $(\alpha,\beta) = (0,0),(1,2),(2,4)$ *let* $T^* = 7, 10, 11$ *respectively. Then:*

(a) *for every* $1 \le T < T^*$ *and every* $(p,q) \in Q$*,* $R_T^{\alpha,\beta}(p,q) \le R_T^{\alpha,\beta}(0,1)$*, and* $R_{T^*}^{\alpha,\beta} \le R_{T^*}^{\alpha,\beta}(0,1)$ *fails on* Q*;*

(b) *at* $T = T^*$ *there is an interior instance* (p,q) *with* $R_{T^*}^{\alpha,\beta}(p',q') < R_{T^*}^{\alpha,\beta}(p,q)$ *for every* $(p',q') \in \partial Q$*.*

Hence $T_{\mathrm{c}} = T_{\mathrm{e}} = 7$, 10, 11 *for the three policies: corner departure and edge departure coincide, whereas for Thompson sampling they are* 5 *and* 18 [12].

Proof. (a) Proposition 4.1 turns the bound 1 of Theorem 4.2 into the corner value at each $T < T^*$; the failure at T^* is the interior witness of Theorem 4.3, whose regret exceeds $1 = R_{T^*}(0,1)$. (b) is Theorem 4.3 read as a strict comparison. For $T < T^*$ the worst case is attained at the boundary point $(0,1)$, and at T^* no boundary point is a worst case, so $T_{\mathrm{e}} = T^*$; and $T_{\mathrm{c}} = T^*$ by Definition 2.2(a). □

Theorem 4.5. *Regularisation strictly delays the departure:* $T_{\mathrm{c}}(0,0) = 7 < T_{\mathrm{c}}(1,2) = 10 < T_{\mathrm{c}}(2,4) = 11$*. As certified facts:* $R_7^{1,2} \le R_7^{1,2}(0,1)$ *on* Q *while* $R_7^{0,0} \le R_7^{0,0}(0,1)$ *fails on* Q*; and* $R_{10}^{2,4} \le R_{10}^{2,4}(0,1)$ *on* Q *while* $R_{10}^{1,2} \le R_{10}^{1,2}(0,1)$ *fails on* Q*.*

Proof. The four facts are the horizons 7 and 10 of Theorem 4.4(a). □

Proposition 4.6 (controls)**.** *(a)* $R_7^{0,0} \le 1$*,* $R_{10}^{1,2} \le 1$ *and* $R_{11}^{2,4} \le 1$ *all fail on* Q*: the corner bound of Theorem 4.2 does not survive the onset horizon for any of the three policies. (b) For* RG$(0,0)$ *at* $T = 7$ *no boundary instance beats the corner:* $R_7^{0,0}(p,q) \le R_7^{0,0}(0,1)$ *for every* $(p,q) \in \partial Q$*. For* RG$(2,4)$ *at* $T = 11$ *the boundary already carries more than the corner value: some* $p \in [0,1]$ *has* $R_{11}^{2,4}(0,1) < R_{11}^{2,4}(p,0)$*,*

namely $p = 417785/10^6$. *So the* $1011/1000$ *of Theorem 4.3(c) is not removable slack. (c)* $R_7^{0,0}(0,1) > 0$, *and* $R_7^{0,0}(p,p) = 0$ *for every real* p.

Proof. (a) the three interior witnesses of Theorem 4.3; (b) the boundary half of Theorem 4.3(a) with Proposition 4.1, and one exact evaluation at $(417785/10^6, 0)$, whose value is $1.010232\ldots$ by the floating-point search; (c) Proposition 4.1 and the factor $p - q$. □

Remark 4.7. Proposition 4.6(b) says that the three departures are not the same event. For $(0,0)$ and $(1,2)$ the corner is still a global maximiser of the regret restricted to the boundary at the onset horizon, and the instance that overtakes it is interior; for $(2,4)$ an edge instance near $(0.418, 0)$ has already passed the corner at $T = 11$, and the interior instance then passes that. In all three cases the corner is not overtaken by anything near it (Theorem 5.3).

5. The mechanism: the corner derivative separates the two classes

Fix the edge $q = 0$. By (1), the regret there is $e_T(p) = p \sum_{S_2=0} w_s\, p^{S_1}(1-p)^{F_1}$, the sum over the states of level $< T$ with $S_2 = 0$ (those with $S_2 > 0$ carry a factor q^{S_2} and vanish); for $p \geq 0$ this is $R_T(p, 0)$ by (2).

Lemma 5.1. *For every list of states and every real* x, *the function* e_T *is differentiable at* x, *with derivative the product rule applied termwise; at* $x = 1$ *a term* $w_s\, p^{S_1}(1-p)^{F_1}$ *(with* $S_2 = 0$*) contributes* $w_s S_1$ *if* $F_1 = 0$, $-w_s$ *if* $F_1 = 1$ *and* 0 *otherwise, plus the value* $e_T(1)$ *itself from the outer factor* p. *The derivative at* 1 *is therefore an exact rational number* D_T, *computed in rational arithmetic and identified with the derivative of the real function.*

Proof. Each term is a polynomial and Mathlib's product rule applies. The formal statement is a `HasDerivAt` of the real function at every real x, assembled from the list of states (`hasDerivAt_edgeR` in Appendix A); a cast lemma identifies its value at $x = 1$ with the rational termwise derivative. Since e_T agrees with $R_T(\cdot, 0)$ on the half-line $[0, \infty)$, which is a neighbourhood of 1, the number D_T is the derivative of $p \mapsto R_T(p, 0)$ at $p = 1$ in the sense of Definition 2.2(c). □

Proposition 5.2. *The corner derivatives* $D_T = \partial_p R_T(1, 0)$ *are the exact rationals*

Bayes-optimal, $T = 1, \ldots, 12$	$\frac{1}{2},\ 0,\ -\frac{3}{4},\ -1,\ -1,\ -1,\ -1,\ -1,\ -1,\ -1,\ -\frac{3}{2},\ -\frac{3}{2}$
Thompson sampling, $T = 1, \ldots, 6$	$\frac{1}{2},\ \frac{2}{3},\ \frac{13}{24},\ \frac{31}{120},\ -\frac{847}{14400},\ -\frac{131167}{378000}$

The Thompson row is the certified ladder of [12, Proposition 3.5]*, recomputed here by the generic dynamic programme with the Thompson decision rule, sign change between* $T = 4$ *and* $T = 5$ *included; it is the reproduction control for this section. The regularised-greedy ladders are stated in Theorem 5.3(b).*

Proof. Eighteen exact rational evaluations of the termwise derivative of Lemma 5.1. Note that the Bayes-optimal ladder is ≥ 0 exactly while the corner is a worst case ($T \leq 2$) and strictly negative from $T = 3 = T_{\mathrm{c}}$ on. □

Theorem 5.3. (a) First-order for the Bayes-optimal policy. $p \mapsto e_2(p)$ *has derivative exactly* 0 *at* $p = 1$, *and* $p \mapsto e_3(p)$ *has derivative* $-\frac{3}{4}$ *at* $p = 1$, *both as derivatives of the real functions. So at* $T = 2$, *the last horizon at which the corner is a worst case, the edge derivative vanishes, and at* $T = 3 = T_{\mathrm{c}}$ *it is negative: the corner is not a local maximiser along the edge, and instances arbitrarily close to it are strictly worse.*

(b) Zeroth-order for regularised greedy. *For every horizon* $2 \leq T \leq 14$,

$$D_T^{0,0} = 2^{\,2-T}, \qquad D_T^{1,2} = D_T^{2,4} = \begin{cases} 1, & T = 2, \\ \frac{1}{2}, & 3 \leq T \leq 14, \end{cases}$$

all strictly positive; in particular there is no sign change at or near $T_{\mathrm{c}} = 7, 10, 11$.

(c) *The two horizons straddling* $T_{\mathrm{c}} = 7$ *for pure greedy, as real derivatives:* $p \mapsto e_6(p)$ *has derivative* $\frac{1}{16}$ *at* $p = 1$ *and* $p \mapsto e_7(p)$ *has derivative* $\frac{1}{32}$ *at* $p = 1$.

Consequently, for regularised greedy the corner $(1,0)$ *remains a strict local maximiser of the regret along the edge* $q=0$ *at and past its own departure horizon, and no first-order test at the corner can detect the departure: the worst case moves to an interior instance at positive distance (Theorem 4.3).*

Proof. (a) and (c) are Lemma 5.1 at the stated horizons, each value one exact rational comparison. (b) is one Boolean over the 13×3 rational derivatives, comparing each with the stated closed form. The consequence stated after the theorem is the elementary fact that a function with strictly positive derivative at 1 is strictly smaller than its value at 1 on some interval $(1-\varepsilon,1)$, applied to e_T; it is not separately formalised. The closed form 2^{2-T} is verified for $2\le T\le 14$ and the constant $\frac{1}{2}$ for $3\le T\le 14$; neither is proved for all T. □

Remark 5.4. Theorems 4.3 and 4.4 locate the global maximum at the onset horizon off the boundary and say nothing about derivatives; they are compatible with a first-order departure at the corner. Theorem 5.3(b) rules that out: the instrument that *defines* the first-order mechanism in [12] — the sign of the corner derivative, which flips exactly at T_{c} for Thompson sampling and for the Bayes-optimal policy — is exhibited failing to see the departure for a whole class of policies, at the exact horizon at which the departure happens. Conversely (b) does not imply the departure: a positive derivative at one point of the edge says nothing about a distant interior instance. The two descriptions are logically independent, as in [12, Remark 4.9] for the edge departure of Thompson sampling, and together they say why $T_{\mathrm{c}}=T_{\mathrm{e}}$ for regularised greedy.

Proposition 5.5 (controls)**.** *(a)* $D_{11}^{\mathrm{B}}\neq -1$*: the Bayes-optimal ladder does not stay at* -1*. (b)* $D_7^{0,0}\neq 0$*: the pure-greedy derivative at the departure horizon is not zero, so the zeroth-order statement is not the vacuous "the derivative vanishes". (c)* $D_7^{0,0}\neq D_6^{0,0}$*: the derivative is not constant across the departure either.*

Proof. Three exact rational comparisons ($-\frac{3}{2}\neq -1$, $\frac{1}{32}\neq 0$, $\frac{1}{32}\neq\frac{1}{16}$). □

6. The onset table, and computations outside the formal development

Table 1 collects the certified statements of Sections 3–5 with the Thompson-sampling row of [12] and three floating-point rows. Every entry marked $*$ is a theorem of this paper or of [12]; every other entry is a floating-point measurement, computed by a grid search over the square with local refinement in the same Python model that found the witnesses, and no theorem depends on it.

policy	corner value $R_T(1,0)$	T_{c}	T_{e}	mechanism at the corner
TS	$\frac{1}{2},\frac{5}{6},\frac{25}{24},\frac{281}{240},\dots$ [12]	5^*	18^*	first-order* [12]
B_T	$\frac{1}{2}$ at every $T\le 12^*$	3^*	8	first-order*
RG$(0,0)$	1 at every $2\le T\le 12^*$	7^*	7^*	zeroth-order*
RG$(1,2)$	1 at every $2\le T\le 12^*$	10^*	10^*	zeroth-order*
RG$(2,4)$	1 at every $2\le T\le 12^*$	11^*	11^*	zeroth-order*
UCB1	1 on $2\le T\le 6$, 2 on $7\le T\le 15$, …	5	>20	corner returns at $T=7,\dots,10$
KL-UCB	1 at every $T\le 7$	8	10	boundary instance first

Table 1. The onset table. $*$: certified. The three unstarred rows and the unstarred T_{e} of the Bayes-optimal policy are floating-point measurements (Remarks 6.1 and 6.2). For UCB1 and KL-UCB the corner value is at $(1,0)$; the conventions are those of Remark 6.2.

Remark 6.1 (the Bayes-optimal worst case beyond $T=4$; floating point). The floating-point maximum of R_T^{B} over the square is

$$\tfrac{1}{2},\ \tfrac{1}{2},\ 0.598076,\ \tfrac{2}{3},\ 0.716877,\ 0.764695 \qquad (T=1,\dots,6),$$
$$0.816951,\ 0.847054,\ 0.900656,\ 0.937890,\ 0.983479,\ 1.018754 \qquad (T=7,\dots,12),$$

attained on the boundary for $T\le 7$ and at an interior point of the square from $T=8$ on. The entries for $T\le 4$ are the exact values of Proposition 3.1 and Theorem 3.5 (the $T=3$ entry being $\frac{3}{2}\sqrt{3}-2$ at

$(\sqrt{3}-1, 0)$, an edge maximum computed outside the formal development). So the evidence is $T_{\mathrm{e}}(\mathrm{B}) = 8$: five horizons after its own T_{c} and ten before the Thompson value 18. This is a measurement, not a theorem; Section 8 prices the missing certificate.

Remark 6.2 (UCB1 and KL-UCB; floating point). Both indices are transcendental, so nothing in this remark is machine-checked. The conventions, which differ from the benchmark list of [13, Section 3.3] ("$A_t = \arg\max_i S_i/N_i + \sqrt{c \log t / N_i}$ with $c = 2$" and, for KL-UCB, the exploration level $\log t + c \log\log t$ with $c = 3$), are: arm 1 is pulled first, ties are broken by a fair coin, UCB1 uses the index $S_i/N_i + \sqrt{2 \ln n / N_i}$ with n the number of plays already made, and KL-UCB uses the exploration function $\ln t$ with t the current round and no $\log\log$ term.

UCB1. The worst case over the square at $T = 1, \ldots, 12$ is

$$1,\ 1,\ 1,\ 1,\ 1.208876,\ 1.294452,\ 2,\ 2,\ 2,\ 2,\ 2.015427,\ 2.072208.$$

So $T_{\mathrm{c}} = 5$, but the corner is a worst case again at $T = 7, \ldots, 10$: the set of horizons at which a corner is a worst case is not an initial segment. The reason is exact. At $(1, 0)$ the trajectory is deterministic and the regret is the number of re-explorations of the dead arm, which happen at rounds

$$2,\ 7,\ 16,\ 31,\ 54,\ 87,\ 135,\ 205,\ 307,\ 455,\ 670,\ 983,\ 1441,\ 2117$$

(floating point; the sequence was not found in the OEIS), so $R_T(1, 0)$ is 1 on $2 \le T \le 6$, 2 on $7 \le T \le 15$, 3 on $16 \le T \le 30$, and the interior maximum only catches up between the jumps. At every horizon searched, $T \le 20$, the maximum over the square is attained on the boundary — at the corner $(0, 1)$ whenever the corner is a worst case, and otherwise on the edge $p = 0$ or its mirror $q = 0$ — which is the $T_{\mathrm{e}} > 20$ of Table 1.

KL-UCB. The worst case is 1 for $T \le 7$ and then 1.042075, 1.108734, 1.140842, 1.178284, 1.230835 for $T = 8, \ldots, 12$, so $T_{\mathrm{c}} = 8$; unlike regularised greedy, the corner is first beaten by a *boundary* instance, near $(0.5523, 0)$ and its mirror $(0, 0.5523)$, and the worst case leaves the boundary only at $T_{\mathrm{e}} = 10$, with maximiser near $(0.0270, 0.5467)$.

Remark 6.3 (explore-then-commit, and a correction). The plan from which this study was carried out printed, as the worst case of the best explore-then-commit rule at $T = 3, \ldots, 12$, the row

$$0.87500,\ 0.97210,\ 1,\ 1,\ 1.03923,\ 1.09924,\ 1.16960,\ 1.24601,\ 1.32622,\ 1.40898.$$

That row is not the worst case of any ETC(m). It is reproduced exactly by a *re-comparing* variant: a counts-based rule that, after the exploration phase, compares S_1 with S_2 afresh in every round, its worst case minimised over m as the printed row is — the first two entries come from $m = 0$, the rest from $m = 1$. ETC(m) commits: once the exploration phase ends the committed arm keeps being pulled, and the counts still identify it ($N_i > m$), provided the commit test is made before the comparison of successes. With the commit test first, the worst case of ETC(1) is $T^2/(8(T-2))$ at every $T = 4, \ldots, 9$ in the Python model, the certified value of [12], and Proposition 3.7 pins it at $T = 5, 6, 7$; with the comparison first one obtains the printed row. The two readings differ already at $T = 3$ (0.875 against 1). For this reason the third invariant of [12] — the smallest horizon at which a policy's worst case falls strictly below that of every explore-then-commit rule — is deliberately not recorded for any policy here: fixing one reading of "explore-then-commit" and certifying it is a separate task.

7. Verification

Every theorem and proposition above is a statement about the piecewise polynomial (2) over the real numbers, or an exact rational evaluation of it, formalised and checked in Lean 4 [10] with Mathlib [11]; Appendix A reproduces the formal statements verbatim. The development consists of five modules: `Model` (277 lines; Mathlib-free: states, the forward dynamic programme for an arbitrary counts-based policy, the Bellman value table and the policy read off it, RG(α, β), the two controls, rational evaluation, and the certificate builds in square coordinates), `Main` (465 lines; the meaning over $\mathbb{R}$, the generic certificate theorem and its five instantiations (S) and (E), Section 3), `Certs` (347 lines; Mathlib-free: the 74 Bernstein certificates of Section 4 and the 56 denominator checks, 130 Boolean evaluations in all), `Greedy` (529 lines; Section 4) and `Deriv` (242 lines; Section 5). The development has no `sorry` and no axiom

beyond Lean's standard three (`propext`, `Classical.choice`, `Quot.sound`) and the axioms introduced by `native_decide`, which is used only to evaluate a Boolean on finite data: the certificate checker on a tree, an exact rational comparison, the equality of two lists of states, or one of the ladders. The statements of Sections 3–5 are carried by 37 declarations — 11 in `Main`, 20 in `Greedy`, 6 in `Deriv` — and Appendix A reproduces those together with nine further ones. Running `#print axioms` on all forty-six gives, in every case, exactly the standard three together with the `native_decide` axioms of the Boolean evaluations that declaration rests on; there is no `sorryAx` and no declared axiom anywhere. Seven of the forty-six — the generic certificate theorem, its five instantiations (S) and (E), and the `HasDerivAt` lemma of Section 5 — depend on the standard three alone. The reasoning layer — the soundness of the checker (itself free of `native_decide`), the identity between the certificate build and $D\,\mathbb{E}[N_2(T)]$, the generic theorem behind (S) and (E), the `HasDerivAt` derivations of Section 5, the polynomial inequalities of Theorem 3.5 and the transfer of rational evaluations to $\mathbb{R}$ — uses only the standard three. A clean check of the whole development in import order takes 49 s of wall time (70 s of CPU) with peak memory 6.86 GB: `Model` 1.6 s at 0.78 GB, `Main` 18.7 s, `Certs` 7.5 s at 0.78 GB, `Greedy` 10.8 s and `Deriv` 10.4 s, the three Mathlib modules at the roughly 6.8 GB baseline of `import Mathlib`. Putting the 130 Boolean evaluations of `Certs` in a module that imports nothing from Mathlib is what keeps them at 0.8 GB. The exploratory computation was about 0.6 CPU-hours of Python, dominated by the floating-point maximisations, and a literature sweep about 0.5 CPU-hours; in all about 1.2 CPU-hours. The exact values, the polynomial of Theorem 3.5, the certificate trees and the witnesses were computed first, independently of the formal development, by three Python programs: the dynamic programme with the four exact and the two floating-point policies, an independent model of the certificate checker in square coordinates (including the multidegree padding without which the Bernstein coefficients of a difference are meaningless), and the iterative-deepening tree search. Nothing from them is trusted: the formal development rebuilds every polynomial from the dynamic programme and re-runs every check. Repository reference to be added at submission.

8. Questions

(1) T_{e} *for the Bayes-optimal policy.* The evidence of Remark 6.1 is $T_{\mathrm{e}}(\mathrm{B}) = 8$. The interior half is cheap — four one-variable boundary certificates and one rational interior instance, the shape of Theorem 4.3. The other half, that the maximum is on the boundary for every $T \le 7$, needs a *dominance* certificate in the triangle coordinates $p = q + (1-q)u$ of [12], $R_T(u,0) - R_T(q + (1-q)u, q) \ge 0$ on the box, for each $T \le 7$, together with its mirror image in the other triangle or a proof that the policy is symmetric, $\mathbb{E}[N_2](p,q) + \mathbb{E}[N_2](q,p) = T$, which the Python model verifies coefficientwise for every $T \le 10$ but the formal development does not. It needs a second coordinate system in the model and was not attempted here.

(2) *UCB1 and KL-UCB as theorems.* Every decision compares $S_i/N_i + \sqrt{2\ln n / N_i}$ across arms, so certifying one horizon needs a rational enclosure of $\ln n$ for $n \le T$ and of two square roots per decision node. The cheapest first target is not the worst case but the corner: at $(1,0)$ the trajectory is deterministic, so one certified ladder of enclosures of $\ln n$ settles $R_T(1,0)$ for every T at once and would make the re-exploration rounds of Remark 6.2 exact.

(3) *Other priors.* The value table (3) is prior-agnostic apart from the posterior means $(S_i+\alpha)/(N_i+\alpha+\beta)$, so T_{c} as a function of a $\mathrm{Beta}(\alpha,\beta)$ prior costs one extra parameter and nothing else. Whether T_{c} moves with the prior is the second axis left open in [12] for Thompson sampling, and it is unanswered for every policy.

(4) *The exact worst case past* T_{c}. $T = 4$ gives the rational $2/3$ because the maximiser is rational. At $T = 3$ the edge maximum is $\frac{3}{2}\sqrt{3} - 2$ at $p = \sqrt{3} - 1$, an algebraic number of degree two, which could be pinned exactly over $\mathbb{Q}(\sqrt{3})$; from $T = 5$ on the worst case needs two-sided enclosures as in [12], and for $T \ge 8$ a genuinely two-variable subdivision. Likewise the exact worst case of $\mathrm{RG}(\alpha,\beta)$ at its own onset horizon: the maximiser is an interior algebraic point of high degree, and the cost of a 10^{-6} enclosure at degree 7 to 11 was not measured.

(5) *Regularisation in general.* Theorem 4.5 is three points. Whether $T_{\mathrm{c}}(\alpha,\beta)$ is monotone along the ray $(\alpha, 2\alpha)$, or in α and β separately, and whether $T_{\mathrm{c}} = T_{\mathrm{e}}$ holds for every (α,β), are well-posed

finite questions per parameter pair; each costs one square certificate per horizon below the onset and eight edge certificates at it.

(6) *Explore-then-commit.* Fix one reading of the rule (Remark 6.3) and certify, for each policy of Table 1, the smallest horizon at which its worst case falls strictly below that of every explore-then-commit rule, as [12] does for Thompson sampling.

Appendix A. The formal statements

The declarations below are reproduced verbatim from the formal development (proofs omitted). Here `bayesRegret T p q` is $R^{\mathrm{B}}_T(p,q)$ of (2) over $\mathbb{R}$ and `rgRegret al be T p q` is $R^{\alpha,\beta}_T(p,q)$, both defined at every real (p,q); `regR T L p q` is (2) over $\mathbb{R}$ for an arbitrary list of states `L`, `regQ` the same over $\mathbb{Q}$, and `n2R` the polynomial (1); `bayesTerms T`, `rgTerms al be T`, `tsTerms T` and `etcTerms m T` are the lists of states of the four policies; `bayesCornerOK 12`, `rgCornerOK al be 12` and `rgDerivOK 13` are the Booleans of Propositions 3.1, 4.1 and Theorem 5.3(b) (the last ranging over $T = 2, \dots, 14$); `edgeR L` is the edge function e_T of Section 5 and `edgeDerivQ L` its rational derivative D_T at $p = 1$; `sumValR` and `dSumR` are the state sum and its termwise derivative over $\mathbb{R}$. In the generic theorem, `checkHP t P` is the certificate checker on the tree `t` and the homogeneous polynomial `P`, `targG` and `targH` the builds G and H of Section 2.5, `hP, hNP, hQ, hNQ, hZero, hOne` the homogeneous factors $p, 1-p, q, 1-q, 0, 1$, `evH x` evaluation at the box point `x`, and `denOKof D L` the denominator side condition. The generic theorem, its five instantiations, the $T = 4$ state list and polynomial, and the real derivative are listed first; then the pinned statements in the order of the text.

```
theorem reg_le_of_certs (T : ℕ) (L : List (Rat × St)) (D : ℕ) (Mn Md : ℤ)
    (a b c d : HP) (t1 t2 : SubTree)
    (hD : 0 < D) (hMd : 0 < Md) (hden : denOKof D L = true)
    (h1 : checkHP t1 (targG a b c d D L Mn Md) = true)
    (h2 : checkHP t2 (targH a b c d D T L Mn Md) = true)
    (x : Nat → ℝ) (hx : InBox x)
    (hb : evH x b = 1 - evH x a) (hd : evH x d = 1 - evH x c) :
    regR T L (evH x a) (evH x c) ≤ (Mn : ℝ) / (Md : ℝ)
theorem reg_le_square (T : ℕ) (L : List (Rat × St)) (D : ℕ) (Mn Md : ℤ) (t1 t2 : SubTree)
    (hD : 0 < D) (hMd : 0 < Md) (hden : denOKof D L = true)
    (h1 : checkHP t1 (targG hP hNP hQ hNQ D L Mn Md) = true)
    (h2 : checkHP t2 (targH hP hNP hQ hNQ D T L Mn Md) = true) :
    ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → regR T L p q ≤ (Mn : ℝ) / (Md : ℝ)
theorem reg_le_edge_q0 (T : ℕ) (L : List (Rat × St)) (D : ℕ) (Mn Md : ℤ) (t1 t2 : SubTree)
    (hD : 0 < D) (hMd : 0 < Md) (hden : denOKof D L = true)
    (h1 : checkHP t1 (targG hP hNP hZero hOne D L Mn Md) = true)
    (h2 : checkHP t2 (targH hP hNP hZero hOne D T L Mn Md) = true) :
    ∀ p : ℝ, 0 ≤ p → p ≤ 1 → regR T L p 0 ≤ (Mn : ℝ) / (Md : ℝ)
theorem reg_le_edge_q1 (T : ℕ) (L : List (Rat × St)) (D : ℕ) (Mn Md : ℤ) (t1 t2 : SubTree)
    (hD : 0 < D) (hMd : 0 < Md) (hden : denOKof D L = true)
    (h1 : checkHP t1 (targG hP hNP hOne hZero D L Mn Md) = true)
    (h2 : checkHP t2 (targH hP hNP hOne hZero D T L Mn Md) = true) :
    ∀ p : ℝ, 0 ≤ p → p ≤ 1 → regR T L p 1 ≤ (Mn : ℝ) / (Md : ℝ)
theorem reg_le_edge_p0 (T : ℕ) (L : List (Rat × St)) (D : ℕ) (Mn Md : ℤ) (t1 t2 : SubTree)
    (hD : 0 < D) (hMd : 0 < Md) (hden : denOKof D L = true)
    (h1 : checkHP t1 (targG hZero hOne hQ hNQ D L Mn Md) = true)
    (h2 : checkHP t2 (targH hZero hOne hQ hNQ D T L Mn Md) = true) :
    ∀ q : ℝ, 0 ≤ q → q ≤ 1 → regR T L 0 q ≤ (Mn : ℝ) / (Md : ℝ)
theorem reg_le_edge_p1 (T : ℕ) (L : List (Rat × St)) (D : ℕ) (Mn Md : ℤ) (t1 t2 : SubTree)
    (hD : 0 < D) (hMd : 0 < Md) (hden : denOKof D L = true)
    (h1 : checkHP t1 (targG hOne hZero hQ hNQ D L Mn Md) = true)
    (h2 : checkHP t2 (targH hOne hZero hQ hNQ D T L Mn Md) = true) :
    ∀ q : ℝ, 0 ≤ q → q ≤ 1 → regR T L 1 q ≤ (Mn : ℝ) / (Md : ℝ)
theorem bayes_terms_four :
    bayesTerms 4 =
      [((1 : Rat)/2, ⟨0,0,0,0⟩), ((1 : Rat)/2, ⟨0,1,0,0⟩), ((1 : Rat)/2, ⟨0,0,1,0⟩),
       ((1 : Rat)/2, ⟨1,1,0,0⟩), ((1 : Rat)/2, ⟨0,1,1,0⟩), ((1 : Rat)/2, ⟨0,1,0,1⟩),
       ((1 : Rat)/2, ⟨0,0,2,0⟩), ((1 : Rat)/2, ⟨1,1,1,0⟩), ((1 : Rat)/2, ⟨0,1,2,0⟩),
       ((3 : Rat)/2, ⟨0,1,1,1⟩), ((1 : Rat)/2, ⟨0,2,0,1⟩), ((1 : Rat)/2, ⟨0,0,3,0⟩),
       ((1 : Rat)/2, ⟨0,0,2,1⟩)]
theorem bayes_n2_four (p q : ℝ) :
    n2R p q (bayesTerms 4) = 2 + (3/2) * q - (3/2) * p + p * q^2 - p^2 * q
theorem hasDerivAt_edgeR (L : List (Rat × St)) (x : ℝ) :
    HasDerivAt (edgeR L) (sumValR x (1 - x) 0 1 L + x * dSumR x L) x
theorem bayes_corner_ladder_flat : bayesCornerOK 12 = true
theorem bayes_corner_half :
    bayesRegret 1 1 0 = 1/2 ∧ bayesRegret 2 1 0 = 1/2 ∧ bayesRegret 3 1 0 = 1/2
      ∧ bayesRegret 4 1 0 = 1/2
theorem bayes_corner_is_max_upto_two (T : ℕ) (hT : 1 ≤ T) (hT2 : T ≤ 2) :
```

```
    ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → bayesRegret T p q ≤ 1/2
theorem bayes_corner_not_max_at_three :
    ∃ p q : ℝ, 0 ≤ p ∧ p ≤ 1 ∧ 0 ≤ q ∧ q ≤ 1 ∧ bayesRegret 3 1 0 < bayesRegret 3 p q
theorem bayes_corner_departure_is_three :
    (∀ T : ℕ, 1 ≤ T → T ≤ 2 →
        ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → bayesRegret T p q ≤ bayesRegret T 1 0)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → bayesRegret 3 p q ≤ bayesRegret 3 1 0)
theorem bayes_worst_case_four :
    (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → bayesRegret 4 p q ≤ 2/3)
      ∧ bayesRegret 4 (2/3) 0 = 2/3
theorem bayes_edge_identity_four (p : ℝ) (hp : 0 ≤ p) :
    2/3 - bayesRegret 4 p 0 = (3/2) * (p - 2/3)^2
theorem control_bayes_corner_bound_fails_at_three :
    ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → bayesRegret 3 p q ≤ 1/2)
theorem control_bayes_no_instance_above_two_thirds :
    ¬ (∃ p q : ℝ, 0 ≤ p ∧ p ≤ 1 ∧ 0 ≤ q ∧ q ≤ 1 ∧ (2/3 : ℝ) < bayesRegret 4 p q)
theorem control_bayes_nonvacuous :
    (∃ p q : ℝ, 0 ≤ p ∧ p ≤ 1 ∧ 0 ≤ q ∧ q ≤ 1 ∧ 0 < bayesRegret 4 p q)
      ∧ (∀ p : ℝ, bayesRegret 4 p p = 0)
theorem control_reproduces_ts_worstcase_corner :
    regQ 1 (tsTerms 1) 1 0 = 1/2 ∧ regQ 2 (tsTerms 2) 1 0 = 5/6
      ∧ regQ 3 (tsTerms 3) 1 0 = 25/24 ∧ regQ 4 (tsTerms 4) 1 0 = 281/240
      ∧ regQ 5 (tsTerms 5) 1 0 = 2007/1600 ∧ regQ 6 (tsTerms 6) 1 0 = 1983559/1512000
      ∧ regQ 5 (etcTerms 1 5) (5/6) 0 = 25/24
      ∧ regQ 6 (etcTerms 1 6) (3/4) 0 = 9/8
      ∧ regQ 7 (etcTerms 1 7) (7/10) 0 = 49/40
theorem rg_corner_ladder_flat :
    rgCornerOK 0 0 12 = true ∧ rgCornerOK 1 2 12 = true ∧ rgCornerOK 2 4 12 = true
theorem rg00_corner_is_max_upto_six (T : ℕ) (hT1 : 1 ≤ T) (hT : T ≤ 6) :
    ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 0 0 T p q ≤ 1
theorem rg00_boundary_at_seven (p q : ℝ) (hp0 : 0 ≤ p) (hp1 : p ≤ 1)
    (hq0 : 0 ≤ q) (hq1 : q ≤ 1) (hb : p = 0 ∨ p = 1 ∨ q = 0 ∨ q = 1) :
    rgRegret 0 0 7 p q ≤ 1
theorem rg00_interior_beats_boundary_at_seven :
    ∃ p q : ℝ, 0 < p ∧ p < 1 ∧ 0 < q ∧ q < 1 ∧ (1 : ℝ) < rgRegret 0 0 7 p q
theorem rg00_onset_is_seven :
    (∀ T : ℕ, 1 ≤ T → T ≤ 6 → ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 0 0 T p q ≤ rgRegret 0 0 T 0 1)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 0 0 7 p q ≤ rgRegret 0 0 7 0 1)
theorem rg12_corner_is_max_upto_nine (T : ℕ) (hT1 : 1 ≤ T) (hT : T ≤ 9) :
    ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 1 2 T p q ≤ 1
theorem rg12_boundary_at_ten (p q : ℝ) (hp0 : 0 ≤ p) (hp1 : p ≤ 1)
    (hq0 : 0 ≤ q) (hq1 : q ≤ 1) (hb : p = 0 ∨ p = 1 ∨ q = 0 ∨ q = 1) :
    rgRegret 1 2 10 p q ≤ 1
theorem rg12_interior_beats_boundary_at_ten :
    ∃ p q : ℝ, 0 < p ∧ p < 1 ∧ 0 < q ∧ q < 1 ∧ (1 : ℝ) < rgRegret 1 2 10 p q
theorem rg12_onset_is_ten :
    (∀ T : ℕ, 1 ≤ T → T ≤ 9 → ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 1 2 T p q ≤ rgRegret 1 2 T 0 1)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 1 2 10 p q ≤ rgRegret 1 2 10 0 1)
theorem rg24_corner_is_max_upto_ten (T : ℕ) (hT1 : 1 ≤ T) (hT : T ≤ 10) :
    ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 2 4 T p q ≤ 1
theorem rg24_boundary_at_eleven (p q : ℝ) (hp0 : 0 ≤ p) (hp1 : p ≤ 1)
    (hq0 : 0 ≤ q) (hq1 : q ≤ 1) (hb : p = 0 ∨ p = 1 ∨ q = 0 ∨ q = 1) :
    rgRegret 2 4 11 p q ≤ 1011/1000
theorem rg24_interior_beats_boundary_at_eleven :
    ∃ p q : ℝ, 0 < p ∧ p < 1 ∧ 0 < q ∧ q < 1 ∧ (1011/1000 : ℝ) < rgRegret 2 4 11 p q
theorem rg24_onset_is_eleven :
    (∀ T : ℕ, 1 ≤ T → T ≤ 10 → ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 2 4 T p q ≤ rgRegret 2 4 T 0 1)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 2 4 11 p q ≤ rgRegret 2 4 11 0 1)
theorem rg00_edge_departure_is_seven :
    (∀ T : ℕ, 1 ≤ T → T ≤ 6 → ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 0 0 T p q ≤ rgRegret 0 0 T 0 1)
      ∧ (∃ p q : ℝ, 0 < p ∧ p < 1 ∧ 0 < q ∧ q < 1 ∧
          ∀ p' q' : ℝ, 0 ≤ p' → p' ≤ 1 → 0 ≤ q' → q' ≤ 1 →
            (p' = 0 ∨ p' = 1 ∨ q' = 0 ∨ q' = 1) →
              rgRegret 0 0 7 p' q' < rgRegret 0 0 7 p q)
theorem rg12_edge_departure_is_ten :
    (∀ T : ℕ, 1 ≤ T → T ≤ 9 → ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
        rgRegret 1 2 T p q ≤ rgRegret 1 2 T 0 1)
      ∧ (∃ p q : ℝ, 0 < p ∧ p < 1 ∧ 0 < q ∧ q < 1 ∧
          ∀ p' q' : ℝ, 0 ≤ p' → p' ≤ 1 → 0 ≤ q' → q' ≤ 1 →
            (p' = 0 ∨ p' = 1 ∨ q' = 0 ∨ q' = 1) →
              rgRegret 1 2 10 p' q' < rgRegret 1 2 10 p q)
theorem rg24_edge_departure_is_eleven :
```

```
    (∀ T : ℕ, 1 ≤ T → T ≤ 10 → ∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 →
       rgRegret 2 4 T p q ≤ rgRegret 2 4 T 0 1)
      ∧ (∃ p q : ℝ, 0 < p ∧ p < 1 ∧ 0 < q ∧ q < 1 ∧
          ∀ p' q' : ℝ, 0 ≤ p' → p' ≤ 1 → 0 ≤ q' → q' ≤ 1 →
            (p' = 0 ∨ p' = 1 ∨ q' = 0 ∨ q' = 1) →
              rgRegret 2 4 11 p' q' < rgRegret 2 4 11 p q)
theorem rg_regularisation_delays_the_departure :
    (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 1 2 7 p q ≤ rgRegret 1 2 7 0 1)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 0 0 7 p q ≤ rgRegret 0 0 7 0 1)
      ∧ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 2 4 10 p q ≤ rgRegret 2 4 10 0 1)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 1 2 10 p q ≤ rgRegret 1 2 10 0 1)
theorem control_rg_corner_bound_fails_at_onset :
    ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 0 0 7 p q ≤ 1)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 1 2 10 p q ≤ 1)
      ∧ ¬ (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → rgRegret 2 4 11 p q ≤ 1)
theorem control_rg_boundary_is_tight :
    (∀ p q : ℝ, 0 ≤ p → p ≤ 1 → 0 ≤ q → q ≤ 1 → (p = 0 ∨ p = 1 ∨ q = 0 ∨ q = 1) →
       rgRegret 0 0 7 p q ≤ rgRegret 0 0 7 0 1)
      ∧ (∃ p : ℝ, 0 ≤ p ∧ p ≤ 1 ∧ rgRegret 2 4 11 0 1 < rgRegret 2 4 11 p 0)
theorem control_rg_nonvacuous :
    (0 : ℝ) < rgRegret 0 0 7 0 1 ∧ (∀ p : ℝ, rgRegret 0 0 7 p p = 0)
theorem control_ts_corner_derivative_ladder :
    edgeDerivQ (tsTerms 1) = 1/2 ∧ edgeDerivQ (tsTerms 2) = 2/3
      ∧ edgeDerivQ (tsTerms 3) = 13/24 ∧ edgeDerivQ (tsTerms 4) = 31/120
      ∧ edgeDerivQ (tsTerms 5) = -(847/14400)
      ∧ edgeDerivQ (tsTerms 6) = -(131167/378000)
theorem bayes_corner_derivative_ladder :
    edgeDerivQ (bayesTerms 1) = 1/2 ∧ edgeDerivQ (bayesTerms 2) = 0
      ∧ edgeDerivQ (bayesTerms 3) = -(3/4) ∧ edgeDerivQ (bayesTerms 4) = -1
      ∧ edgeDerivQ (bayesTerms 5) = -1 ∧ edgeDerivQ (bayesTerms 6) = -1
      ∧ edgeDerivQ (bayesTerms 7) = -1 ∧ edgeDerivQ (bayesTerms 8) = -1
      ∧ edgeDerivQ (bayesTerms 9) = -1 ∧ edgeDerivQ (bayesTerms 10) = -1
      ∧ edgeDerivQ (bayesTerms 11) = -(3/2) ∧ edgeDerivQ (bayesTerms 12) = -(3/2)
theorem bayes_corner_departure_is_first_order :
    HasDerivAt (edgeR (bayesTerms 2)) 0 1 ∧ HasDerivAt (edgeR (bayesTerms 3)) (-(3/4)) 1
theorem rg_corner_departure_is_zeroth_order : rgDerivOK 13 = true
theorem rg00_corner_derivative_at_the_departure :
    HasDerivAt (edgeR (rgTerms 0 0 6)) (1/16) 1
      ∧ HasDerivAt (edgeR (rgTerms 0 0 7)) (1/32) 1
theorem control_deriv_ladders :
    edgeDerivQ (bayesTerms 11) ≠ -1 ∧ edgeDerivQ (rgTerms 0 0 7) ≠ 0
      ∧ edgeDerivQ (rgTerms 0 0 7) ≠ edgeDerivQ (rgTerms 0 0 6)
```

References

[1] S. Agrawal and N. Goyal, *Further Optimal Regret Bounds for Thompson Sampling*, arXiv:1209.3353v1 (2012); published in: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013), Proceedings of Machine Learning Research **31** (2013), pp. 99–107.

[2] P. Auer, N. Cesa-Bianchi and P. Fischer, *Finite-time Analysis of the Multiarmed Bandit Problem*, Machine Learning **47** (2002), no. 2–3, 235–256; https://doi.org/10.1023/A:1013689704352.

[3] R. Bellman, *A Problem in the Sequential Design of Experiments*, Sankhyā: The Indian Journal of Statistics **16** (1956), no. 3/4, 221–229; https://www.jstor.org/stable/25048278.

[4] D. A. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*, Monographs on Statistics and Applied Probability, Chapman and Hall, London and New York, 1985.

[5] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos and G. Stoltz, *Kullback–Leibler upper confidence bounds for optimal sequential allocation*, The Annals of Statistics **41** (2013), no. 3, 1516–1541; https://doi.org/10.1214/13-AOS1119.

[6] A. Garivier and O. Cappé, *The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond*, in: Proceedings of the 24th Annual Conference on Learning Theory (COLT 2011), Proceedings of Machine Learning Research **19** (2011), pp. 359–376; arXiv:1102.2490.

[7] P. Jacko, *The Finite-Horizon Two-Armed Bandit Problem with Binary Responses: A Multidisciplinary Survey of the History, State of the Art, and Myths*, arXiv:1906.10173v1 (2019).

[8] A. Kalvit and A. Zeevi, *A Closer Look at the Worst-case Behavior of Multi-armed Bandit Algorithms*, Advances in Neural Information Processing Systems **34** (NeurIPS 2021), pp. 8807–8819; arXiv:2106.02126v3.

[9] T. Lattimore and Cs. Szepesvári, *Bandit Algorithms*, Cambridge University Press, 2020. Chapter 6, "The Explore-Then-Commit Algorithm".

[10] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: Automated Deduction — CADE 28, Lecture Notes in Computer Science **12699**, Springer, 2021, pp. 625–635; https://doi.org/10.1007/978-3-030-79876-5_37.

[11] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, pp. 367–381; https://doi.org/10.1145/3372885.3373824.

[12] Korea Superintelligence Labs, *The exact worst case of Thompson sampling on two Bernoulli arms at horizons up to eighteen*, manuscript, version 2 (2026). Repository reference to be added at submission.
[13] K. Zhou, M. L. Li and K. Wang, *The Greedy Advantage in Finite-Horizon Bandits*, arXiv:2607.29375v1 (2026).