# EXACT FINITE-HORIZON MINIMAX PSEUDOREGRET OF THE TWO-ARMED BERNOULLI BANDIT FOR ITS TWO CLASSICAL ADVERSARIES

KOREA SUPERINTELLIGENCE LABS

Abstract. A learner pulls one of two Bernoulli arms for $T$ rounds, sees only the reward of the arm it pulled, may randomise, and pays the pseudoregret (the gap times the expected number of pulls of the worse arm). We determine the minimax pseudoregret over randomised history-dependent learners, exactly, for the two adversary classes of the classical lower-bound arguments: the symmetric class, arms $(\frac{1}{2}+e, \frac{1}{2}-e)$ or their swap, and the one-good-arm class, arms $(\frac{1}{2}+e, \frac{1}{2})$ or their swap; in both the adversary chooses $e \in [0, \frac{1}{2}]$ as well as the labelling, and the learner knows neither. For the symmetric class the minimax at a fixed gap is attained by the follow-the-leader player at every gap (Kobzar–Kohn), so the minimax over the class is the maximum of one polynomial $P_T$; we compute $P_5, \dots, P_8$ exactly and enclose $R^*(5), \dots, R^*(8)$ in rational windows of width $2 \cdot 10^{-15}$, beyond the horizon $T = 4$ at which Fabius and van Zwet stopped in 1970 calling the algebra prohibitive, and we prove that the best and the worst policy sit symmetrically about the fair-coin policy: $P_T(e) + P_T^{\max}(e) = 2Te$ for all $T$ and $e$. For the one-good-arm class no gap-independent optimal player exists, and the value is pinned by a sandwich: the exact value of the game at one rational gap from below, a gap-free policy certified by a polynomial inequality from above. The sandwich closes at $K \le 3$ and at $K = 5$, where $V^*(5) = 27/50$ exactly, at the gap $2/5$; at $K = 4, 6, 7$ it leaves windows of width at most $4 \cdot 10^{-8}$. Every value, polynomial and certificate stated as a theorem is machine-checked in Lean 4 with Mathlib.

## 1. Introduction

1.1. **The objects.** Two Bernoulli arms are pulled for a fixed number of rounds. At each round a learner chooses a distribution over the two arms as a function of the history of its own past arms and observed rewards, an arm is drawn from that distribution, and the reward of that arm alone is revealed. The loss is the *pseudoregret*, the gap between the two means times the expected number of pulls of the worse arm. This is the game of Kobzar and Kohn [11], Section 1, and, with the same loss convention, the game of Fabius and van Zwet [7] and of Feldman [8]. Two adversary classes are considered, and they are the two classes that the standard bandit lower bounds are proved against:

- the *symmetric* class, arms $(\frac{1}{2}+e, \frac{1}{2}-e)$ or $(\frac{1}{2}-e, \frac{1}{2}+e)$, the class of Kobzar and Kohn and, for two arms, the formal class of Bubeck and Cesa-Bianchi's Lemma 3.2 [5];
- the *one-good-arm* class, arms $(\frac{1}{2}+e, \frac{1}{2})$ or $(\frac{1}{2}, \frac{1}{2}+e)$, the class of Appendix A of Auer, Cesa-Bianchi, Freund and Schapire [2], and, for two arms, Feldman's two-point prior with one experiment at $\frac{1}{2}$.

In both classes the adversary chooses the parameter $e \in [0, \frac{1}{2}]$ *and* the labelling, and the learner knows neither. The quantity of interest is the least worst-case bound a single randomised policy can guarantee against the whole class, at a fixed horizon: $R^*(T)$ for the symmetric class and $V^*(K)$ for the one-good-arm class (Definition 2.3).

1.2. **What the printed record contains.** Kobzar and Kohn [11] prove that the follow-the-leader player is minimax-optimal at every fixed gap (their Lemma 3.1) and determine the leading-order asymptotics of the minimax pseudoregret by PDE methods; they write that "the optimal regret has not been determined previously", and every finite-horizon quantity they print carries an error term. Fabius and van Zwet [7] determine the minimax risk exactly for $N \le 4$ over a larger adversary, the whole square of pairs of means; their exact treatment ends on p. 1916 with "the algebra involved

becomes distressingly complicated", and p. 1908 says that it "seems to remain prohibitive already for $N$ as small as 5". For the one-good-arm class the record is thinner still: Feldman [8] proves the finite-horizon Bayes optimality of the myopic rule and prints no number at any finite horizon, and the lower-bound papers [2, 5] print only the constant $\frac{1}{20}$ in front of $\sqrt{KT}$. Section 6 records exactly what each source prints.

1.3. **What is settled here.** For the symmetric class the minimax at a fixed gap is a polynomial $P_T(e)$, the value of a two-line backward induction, attained by the follow-the-leader player at every gap (Theorems 3.1 and 3.2, which are Kobzar–Kohn's Lemma 3.1, re-proved here by backward induction), so the minimax over the class is $\max_{e\in[0,1/2]} P_T(e)$ (Proposition 3.3). The values $R^*(1) = R^*(2) = \frac{1}{2}$ and $R^*(3) = \frac{9}{16}$, the last at the gap $\frac{3}{4}$, are Theorem 3.5; $R^*(3) = \frac{9}{16}$ at $|\alpha - \beta| = \frac{3}{4}$ is printed in Fabius and van Zwet [7], p. 1916, for their larger adversary, and the two minimaxes coincide at $T = 3$, so this theorem is a formal re-derivation and not a new result. At $T = 4$ the polynomial $P_4$ follows from their printed $N = 4$ risk function by a substitution, and $R^*(4)$, a cubic irrational, is enclosed in a window of width $2\cdot10^{-15}$ (Proposition 3.8); this too is attributed, not claimed. The new results for the symmetric class are the exact polynomials $P_5, \dots, P_8$ (Theorem 3.9), the enclosures of $R^*(5), \dots, R^*(8)$ to width $2 \cdot 10^{-15}$ (Theorem 3.10), and the identity $P_T(e) + P_T^{\max}(e) = 2Te$ between the least and the greatest pseudoregret of any policy (Theorem 3.11).

For the one-good-arm class the two actions are different experiments, the backward induction keeps a genuine two-branch minimum, and the optimal player depends on the gap; the minimax over the class is therefore not the maximum of one player's pseudoregret, and it is pinned by a sandwich (Section 4): the exact value $V_K(e_0)$ of the game at one fixed rational gap from below, the exact pseudoregret polynomial $Q_K$ of one explicit gap-free policy from above. The values $V^*(1) = \frac{1}{4}$, $V^*(2) = \frac{3}{8}$, $V^*(3) = \frac{7}{16}$ (Proposition 4.5) are elementary; the new results are $V^*(5) = \frac{27}{50}$ exactly, at the gap $\frac{2}{5}$ (Theorem 4.6), and two-sided windows for $V^*(4), V^*(6), V^*(7)$ (Theorem 4.7). Table 1 and Table 2 collect the two families of values; the controls of Propositions 3.6, 3.7 and 4.8 show that the two classes are different games ($\frac{1}{4}$ against $\frac{1}{2}$ at horizon 1, $\frac{7}{16}$ against $\frac{9}{16}$ at horizon 3) and that randomisation is load-bearing.

Everything stated as a theorem or proposition is verified in Lean 4 with Mathlib (Section 7); the formal statements are reproduced verbatim in Appendix A. Everything that is not part of the formal development — the horizons past $T = 8$ and the asymptotic consistency checks, the uniform-prior Bayes values, and a reduction of a hard-MDP instance to the one-good-arm class — is confined to Remarks 3.13, 3.14 and 4.10, each so marked in its heading, and nothing there is claimed.

## 2. The games

**Definition 2.1** (Histories and policies)**.** A *history* is a finite sequence of pairs $(a, r) \in \{1, 2\}\times\{0, 1\}$, the arm pulled and the reward observed, and $\mathcal{H}$ is the set of histories. A *policy* is a map $q : \mathcal{H} \to [0, 1]$; $q(h)$ is the probability of pulling arm 1 after history $h$. Policies are randomised and history-dependent; nothing else is assumed of them.

**Definition 2.2** (The symmetric game)**.** Fix $e \in [0, \frac{1}{2}]$ and a *world* $w \in \{1, 2\}$, the label of the better arm. Arm $a$ has mean $\mu_w(a) = \frac{1}{2} + e$ if $a = w$ and $\frac{1}{2} - e$ otherwise, and pulling it costs $c_w(a) = 2e$ if $a \neq w$ and 0 otherwise. The expected pseudoregret of $q$ over the next $n$ rounds from history $h$ is defined by $R_0(q; w, e)(h) = 0$ and

$$R_{n+1}(q; w, e)(h) = \sum_{a\in\{1,2\}} \pi_a(h)\Big(c_w(a) + \mu_w(a)\, R_n(q; w, e)(h \cdot (a, 1)) + (1 - \mu_w(a))\, R_n(q; w, e)(h \cdot (a, 0))\Big),$$

where $\pi_1(h) = q(h)$, $\pi_2(h) = 1 - q(h)$ and $h \cdot (a, r)$ appends an entry. We write $R_T(q; w, e)$ for the value at the empty history. This is the sum over the $4^T$ trajectories of the horizon-$T$ game, so no measure theory is involved.

**Definition 2.3** (Achievable bounds and the minimax)**.** A real $v$ is an *achievable bound at horizon* $T$ if some policy $q$ satisfies $R_T(q; w, e) \le v$ for every $e \in [0, \frac{1}{2}]$ and both $w$; write $\mathcal{A}_T$ for the set of achievable bounds. The *minimax pseudoregret over the symmetric class* is $R^*(T) = \min \mathcal{A}_T$, the least element of $\mathcal{A}_T$, whenever that least element exists. At a fixed gap, $\mathcal{A}_T(e)$ is the set of $v$ such that some $q$ has $R_T(q; w, e) \le v$ for both $w$, and $\min \mathcal{A}_T(e)$ is the minimax at the gap $2e$, the quantity $\bar{R}^*_T$ of Kobzar–Kohn's (1.6).

Every value theorem below asserts that the relevant least element exists and names it; a two-sided enclosure asserts that every achievable bound is at least the lower end and that the upper end is achievable. The adversary in Definition 2.3 picks one instance; it does not mix. Whether the minimax over pure instances equals the maximum over $e$ of the fixed-gap minimax is a separate question, answered affirmatively for the symmetric class (Proposition 3.3) and left open for the one-good-arm class (Remark 4.9).

**Definition 2.4** (The backward inductions)**.** Write $a = \frac{1}{2} + e$ and $b = \frac{1}{2} - e$. For the symmetric game define, for $x, y \in \mathbb{R}$,

$$W_e(0; x, y) = 0, \qquad W_e(n+1; x, y) = 2e \min(x, y) + W_e(n; ax, by) + W_e(n; bx, ay),$$

and $W_e^{\max}$ by the same recursion with max in place of min. Put $P_T(e) = W_e(T; \frac{1}{2}, \frac{1}{2})$ and $P_T^{\max}(e) = W_e^{\max}(T; \frac{1}{2}, \frac{1}{2})$.

The pair $(x, y)$ carries unnormalised posterior weights of the two worlds. A success on arm 1 and a failure on arm 2 are the same evidence, so both actions send $(x, y)$ to the same two children $(ax, by)$ and $(bx, ay)$, and the action decides only the immediate cost; this is why the recursion has one min and no branch on the action. Call an entry $(a, r)$ of a history an *agreement* if $(a, r) = (1, 1)$ or $(2, 0)$ and a *disagreement* otherwise; the likelihood of the history in world 1 is $a^{\#\text{agree}} b^{\#\text{disagree}}$ and in world 2 the same with the exponents exchanged. The *myopic* (follow-the-leader) policy pulls arm 1 if agreements exceed disagreements, arm 2 if disagreements exceed agreements, and each with probability $\frac{1}{2}$ at a tie; the *anti-myopic* policy is 1−myopic.

**Definition 2.5** (The one-good-arm game)**.** Fix $e \in [0, \frac{1}{2}]$ and $w \in \{1, 2\}$. Arm $a$ has mean $\frac{1}{2} + e$ if $a = w$ and $\frac{1}{2}$ otherwise, and pulling $a \ne w$ costs $e$. The pseudoregret $R_K^{\text{og}}(q; w, e)$ is defined by the recursion of Definition 2.2 with these means and costs; $\mathcal{A}_K^{\text{og}}$ and $V^*(K) = \min \mathcal{A}_K^{\text{og}}$ are defined as in Definition 2.3. The backward induction is

$$W_e^{\text{og}}(n+1; x, y) = \min\big( ey + W_e^{\text{og}}(n; ax, \tfrac{y}{2}) + W_e^{\text{og}}(n; bx, \tfrac{y}{2}),\ ex + W_e^{\text{og}}(n; \tfrac{x}{2}, ay) + W_e^{\text{og}}(n; \tfrac{x}{2}, by)\big),$$

with $W_e^{\text{og}}(0; x, y) = 0$, and $V_K(e) = W_e^{\text{og}}(K; \frac{1}{2}, \frac{1}{2})$.

Here pulling arm 1 is informative about $x$ only and pulling arm 2 about $y$ only: the two actions lead to different children, the recursion keeps a two-branch min, and the minimising action depends on $e$.

## 3. The symmetric class

### 3.1. **The minimax at a fixed gap, and its reduction.**

**Theorem 3.1** (Kobzar–Kohn, Lemma 3.1, attainment)**.** *For every $e \in [0, \frac{1}{2}]$, every horizon $T$ and both worlds $w$, the myopic policy has $R_T(\text{myopic}; w, e) = P_T(e)$.*

**Theorem 3.2** (Kobzar–Kohn, Lemma 3.1, minimax at a fixed gap)**.** *For every $e \in [0, \frac{1}{2}]$ and every $T$, $P_T(e)$ is the least element of $\mathcal{A}_T(e)$: every policy has pseudoregret at least $P_T(e)$ in one of the two worlds, and the myopic policy has exactly $P_T(e)$ in both.*

*Proof sketch.* For every policy $q$, all $n$, all $x, y \ge 0$ and every history $h$,

$$W_e(n; x, y) \le x\, R_n(q; 1, e)(h) + y\, R_n(q; 2, e)(h),$$

by induction on $n$: at the head, $q(h)\,y + (1-q(h))\,x \geq \min(x,y)$, and the induction hypothesis is applied at the four children with weights $q(h)$ and $1-q(h)$. The myopic policy attains equality at every history when $(x,y)$ are the two likelihoods, because the smaller likelihood is the one with the larger exponent of $b$, which is exactly the arm the myopic rule avoids; and it is invariant under swapping the arm labels, which swaps the two worlds, so its two pseudoregrets coincide. Kobzar and Kohn prove the same statement by an exchange argument in their Appendix A; the proof here is by backward induction. □

**Proposition 3.3** (Reduction to one polynomial)**.** *If $e_0 \in [0, \frac{1}{2}]$ satisfies $P_T(e) \leq P_T(e_0)$ for all $e \in [0, \frac{1}{2}]$, then $P_T(e_0)$ is the least element of $\mathcal{A}_T$.*

This is one line from Theorem 3.2: the myopic policy is optimal at every gap simultaneously, so it achieves $P_T(e_0)$ against the whole class, and no policy can do better than $P_T(e_0)$ at the gap $2e_0$ alone. The maximiser is a hypothesis and not a conclusion; what is proved is the conditional statement, and each value theorem below supplies its own $e_0$, or its own rational point. A polynomial does attain its maximum on $[0, \frac{1}{2}]$, so the proposition reads in words as $R^*(T) = \max_{[0,1/2]} P_T$; that reading is not itself part of the formal development. Kobzar and Kohn take this maximum only asymptotically; Fabius and van Zwet maximise over the whole square of pairs.

### 3.2. The values at $T \leq 4$.

**Proposition 3.4.** *On $[0, \frac{1}{2}]$,*

$$P_1(e) = e, \quad P_2(e) = 2e - 2e^2, \quad P_3(e) = 3e - 4e^2, \quad P_4(e) = 4e - 7e^2 + 4e^4.$$

In the gap variable $d = 2e$, $P_3$ and $P_4$ read $\frac{3}{2}d - d^2$ and $2d - \frac{7}{4}d^2 + \frac{1}{4}d^4$. These are the $N = 3$ and $N = 4$ risk functions printed by Fabius and van Zwet [7], p. 1916, with the randomisations of the symmetric myopic rule, $\delta(2,1;0,0) = \frac{1}{2}$, $\delta(3,1;0,0) = 0$, $\delta(3,2;0,0) = 1$, substituted and $\alpha + \beta = 1$ imposed; this was verified in exact rational arithmetic. The proposition is therefore routine and attributed.

**Theorem 3.5** ($T \leq 3$; Fabius–van Zwet for $T = 3$)**.** *$R^*(1) = R^*(2) = \frac{1}{2}$, and $R^*(3) = \frac{9}{16}$, attained at $e = \frac{3}{8}$, i.e. at the gap $\frac{3}{4}$ and the arms $(\frac{7}{8}, \frac{1}{8})$; the certificate is $\frac{9}{16} - P_3(e) = 4(e - \frac{3}{8})^2$.*

The value $R^*(3) = \frac{9}{16}$ is printed in Fabius and van Zwet [7], p. 1916, for their unrestricted adversary: "After a little algebra one sees that $\Delta_0$ must have $\delta(2,1;0,0) = 1$ and that $R(\alpha, \beta, \Delta_0)$ attains its maximum $M(\Delta_0) = 9/16$ when $|\alpha - \beta| = 3/4$." Their minimax-risk function at $N = 3$, $\frac{3}{2}|\alpha - \beta| - (\alpha - \beta)^2$, does not depend on $\alpha + \beta$, and on the line $\alpha + \beta = 1$ their general $N = 3$ risk function equals it for every value of the free randomisation, so the unrestricted and the symmetric minimax coincide at $T = 3$. Their risk is the pseudoregret in the present sense (their remark under (1.2): the risk "also equals $|\alpha - \beta|$ times the expected number of trials in which the less favorable experiment is performed"). Theorem 3.5 is thus a formal re-derivation of a 1970 result; the values at $T = 1, 2$, which Fabius and van Zwet do not treat, are routine. The value lives inside a scanned paper with no text layer, and was found by following the citation in Kolnogorov [12] rather than by any text search.

**Proposition 3.6** (Controls at $T = 3$)**.** *$\frac{9}{16} - \frac{1}{1000} \notin \mathcal{A}_3$; $\frac{5}{8} \in \mathcal{A}_3$ but $\frac{5}{8}$ is not the least element of $\mathcal{A}_3$; and $P_3(\frac{1}{4}) = P_3(\frac{1}{2}) = \frac{1}{2} < \frac{9}{16}$, so the worst gap at $T = 3$ is interior and the extreme instance $(1, 0)$ is not the worst one.*

**Proposition 3.7** (Convention controls)**.** *If $q(\emptyset) \in \{0, 1\}$ then every $v$ with $R_1(q; w, e) \leq v$ for all $e \in [0, \frac{1}{2}]$ and both $w$ satisfies $v \geq 1$: over deterministic policies the horizon-1 minimax is 1, twice the randomised value. The policy that always pulls arm 1 is a policy and has $R_3(\,\cdot\,; 2, \frac{1}{2}) = 3$; the constant 2 is not a policy.*

Randomisation is load-bearing in Definition 2.3, and the myopic policy does randomise at the tie, where Kobzar and Kohn (their (3.3)) and Fabius and van Zwet (their (1.5)) both put probability $\frac{1}{2}$.

**Proposition 3.8** ($T = 4$)**.** *Every element of $\mathcal{A}_4$ is at least the lower end, and the upper end belongs to $\mathcal{A}_4$, of*

$$R^*(4) \in [0.605251593991880,\ 0.605251593991882].$$

The value is irrational: $P_4'(e) = 2(8e^3 - 7e + 2)$ and the cubic has no rational root, so the maximiser $e^*(4)$ and $R^*(4) = P_4(e^*(4))$ are cubic irrationals and no exact rational statement exists. The lower end is $P_4(2612/8039)$, the value of the game at a rational gap; the upper end is the myopic player's worst case, certified by the identity

$$U - P_4(e) = c_0 + k\,(e - v)^2 + 4(e - e_1)^2(\tfrac{1}{2} - e)(\tfrac{1}{2} + e + 2e_1),$$

with $e_1 = 2612/8039$, $k = 221891126/64625521$, $v = 289789800605/891891380957$, $c_0 > 0$ and $U = 0.605251593991882$, whose right-hand side is visibly non-negative on $[0, \frac{1}{2}]$. This is *not* the $N = 4$ value .617 of Fabius and van Zwet, which is the minimax over the whole square, attained at $|\alpha - \beta| = .654$; $T = 4$ is the first horizon at which restricting the adversary to the symmetric class lowers the minimax. Since $P_4$ follows from their printed $N = 4$ risk function, Proposition 3.8 is routine and attributed.

### 3.3. **The horizons $5 \le T \le 8$.**

**Theorem 3.9** (The polynomials)**.** *On* $[0, \frac{1}{2}]$,

$$\begin{aligned}
P_5(e) &= 5e - 10e^2 + 8e^4,\\
P_6(e) &= 6e - \tfrac{55}{4}e^2 + 18e^4 - 12e^6,\\
P_7(e) &= 7e - \tfrac{35}{2}e^2 + 28e^4 - 24e^6,\\
P_8(e) &= 8e - \tfrac{175}{8}e^2 + \tfrac{91}{2}e^4 - 66e^6 + 40e^8.
\end{aligned}$$

**Theorem 3.10** (The values)**.** *Each of the following windows contains $R^*(T)$, in the sense that every element of $\mathcal{A}_T$ is at least the lower end and the upper end belongs to $\mathcal{A}_T$:*

$$\begin{aligned}
R^*(5) &\in [0.665598466061581,\ 0.665598466061582],\\
R^*(6) &\in [0.709079607684664,\ 0.709079607684666],\\
R^*(7) &\in [0.760342288803075,\ 0.760342288803077],\\
R^*(8) &\in [0.800891684469692,\ 0.800891684469694].
\end{aligned}$$

*Proof sketch.* The polynomials are the backward induction of Definition 2.4 expanded in the kernel. Every weight pair reached from $(\frac{1}{2}, \frac{1}{2})$ has the form $(a^i b^j, a^j b^i)$, on which $\min(a^i b^j, a^j b^i) = a^{\min(i,j)} b^{\max(i,j)}$ because $0 \le b \le a$; one rewriting rule expands the whole recursion tree, and the result is normalised as a polynomial in $e$. For the values, the lower end of each window is $P_T(e_1)$ at the rational gap $e_1$ of Table 1, a continued-fraction convergent of the maximiser; by Theorem 3.2 no policy beats the fixed-gap value. The upper end is the myopic player's worst case, certified by the Bernstein-type identity of Section 5, with residual degrees $2, 4, 4, 6$ for $T = 5, 6, 7, 8$. Both halves are checked in the kernel. □

### 3.4. **The best and the worst policy.**

**Theorem 3.11.** *For every horizon $T$ and every real $e$, $P_T(e) + P_T^{\max}(e) = 2Te$. Moreover, for $e \in [0, \frac{1}{2}]$, $P_T^{\max}(e)$ is the greatest average pseudoregret of any policy, and it is attained in both worlds by the anti-myopic policy.*

*Proof sketch.* $W_e(n; x, y) + W_e^{\max}(n; x, y) = 2en(x + y)$ by induction on $n$: $\min + \max = x + y$ at the head, and $a + b = 1$ makes the two children's weights sum to $x + y$ again. The attainment statement is the mirror image of Theorem 3.2 with max for min. The fair-coin policy has pseudoregret exactly $Te$, so the least and the greatest pseudoregret sit symmetrically about the coin flip. □

TABLE 1. The symmetric class, $T \le 8$. $R^*(T)$ is exact for $T \le 3$ and a two-sided window of width $\le 2 \cdot 10^{-15}$ for $4 \le T \le 8$, written compactly: $0.60525159399188[0,2]$ is the window $[0.605251593991880, 0.605251593991882]$. $e_1$ is the rational gap parameter at which the lower end $P_T(e_1)$ is evaluated. The column $e^*(T)$ is the numerically computed maximiser of $P_T$ on $[0, \frac{1}{2}]$, given for orientation only; it is not part of the formal development.

| $T$ | $P_T(e)$ | $R^*(T)$ | $e_1$ | $e^*(T)$ |
|---|---|---|---|---|
| 1 | $e$ | $1/2$ | $1/2$ | 0.5 |
| 2 | $2e - 2e^2$ | $1/2$ | $1/2$ | 0.5 |
| 3 | $3e - 4e^2$ | $9/16 = 0.5625$ | $3/8$ | 0.375 |
| 4 | $4e - 7e^2 + 4e^4$ | $0.60525159399188[0,2]$ | 2612/8039 | 0.3249160 |
| 5 | $5e - 10e^2 + 8e^4$ | $0.66559846606158[1,2]$ | 5094/17665 | 0.2883668 |
| 6 | $6e - \frac{55}{4}e^2 + 18e^4 - 12e^6$ | $0.70907960768466[4,6]$ | 2296/8761 | 0.2620705 |
| 7 | $7e - \frac{35}{2}e^2 + 28e^4 - 24e^6$ | $0.76034228880307[5,7]$ | 1913/7909 | 0.2418763 |
| 8 | $8e - \frac{175}{8}e^2 + \frac{91}{2}e^4 - 66e^6 + 40e^8$ | $0.80089168446969[2,4]$ | 2924/12967 | 0.2254955 |

*Remark* 3.12 (Coefficient consequences). On $[0, \frac{1}{2}]$, $P_T(e) = 2e \sum a^{\min(i,j)} b^{\max(i,j)}$ and $P_T^{\max}(e) = 2e \sum a^{\max(i,j)} b^{\min(i,j)}$, the sums running over the nodes of the recursion tree; exchanging $a$ and $b$ is $e \mapsto -e$, so as polynomials $P_T^{\max}(e) = -P_T(-e)$ and Theorem 3.11 says that the odd part of $P_T$ is $Te$. Hence the coefficient of $e$ in $P_T$ is $T$ and every odd coefficient of order $\ge 3$ vanishes, as Table 1 shows. This one-line consequence is stated here as a remark; it is not itself part of the formal development, where only the identity and the attainment are proved.

*Remark* 3.13 (Beyond $T = 8$, and the asymptotics; computations outside the formal development). The same exact-rational recursion, run outside Lean, gives $P_9, \dots, P_{12}$ and $R^*(9) \approx 0.846000003$, $R^*(10) \approx 0.883626765$, $R^*(11) \approx 0.924321713$, $R^*(12) \approx 0.959430405$; these are not claimed here. Two consistency checks against the sources' own asymptotics were run before any formal work. First, $R^*(T)/\sqrt{T}$ takes the values $0.500, 0.354, 0.325, 0.303, 0.298, 0.289, 0.287, 0.283$ at $T = 1, \dots, 8$ and descends towards the constant 0.265 of the Fabius–van Zwet bracket ([7], p. 1907: the asymptotic minimax risk "must be between $0.265N^{1/2}$ and $0.376N^{1/2}$"). The same numeral 0.265, in the same 0/1 scaling, is what Kobzar and Kohn write as $.530T^{1/2}$ in their centred scaling; they attribute it, not to Fabius and van Zwet, but to Bather [3], and for the Bayesian pseudoregret rather than the minimax. Both attributions are reproduced here, neither is checked against Bather's text, and the two quantities are not identified here. Second, $2e^*(T)\sqrt{T}$ takes the values $1.2997, 1.2896, 1.2839, 1.2799, 1.2756$ at $T = 4, \dots, 8$ and descends towards the maximiser $\gamma \approx 1.274$ that Kobzar and Kohn obtain numerically. Neither limit is claimed.

*Remark* 3.14 (The uniform-prior Bayes values; computation outside the formal development). For orientation, the classical Bayes problem with independent uniform priors on the two means has, at horizon $T$, optimal expected reward $\mathrm{OPT}(T)$ and Bayes regret $\frac{2T}{3} - \mathrm{OPT}(T)$ (since $\mathbb{E}\max(p_1, p_2) = \frac{2}{3}$). In exact rational arithmetic, by two independent programs, $\mathrm{OPT}(T)$ for $T = 1, \dots, 12$ is $\frac{1}{2}$, $\frac{13}{12}$, $\frac{5}{3}$, $\frac{41}{18}$, $\frac{26}{9}$, $\frac{421}{120}$, $\frac{595}{144}$, $\frac{1999}{420}$, $\frac{5431}{1008}$, $\frac{16861}{2800}$, $\frac{11070599}{1663200}$, $\frac{8490113}{1164240}$, and the Bayes regret is $\frac{1}{6}$, $\frac{1}{4}$, $\frac{1}{3}$, $\frac{7}{18}$, $\frac{4}{9}$, $\frac{59}{120}$, $\frac{77}{144}$, $\frac{241}{420}$, $\frac{617}{1008}$, $\frac{5417}{8400}$, $\frac{1126201}{1663200}$, $\frac{823807}{1164240}$. This dynamic programme has been run since Steck (1964) to horizons in the thousands (the census is Table 3 of Jacko [9], which lists publication, horizon and hardware and prints no value), so these rationals are a reference point, not a result; they are not formalised, because a faithful statement needs a prior on $[0,1]^2$.

## 4. The one-good-arm class

### 4.1. **The sandwich.**

**Lemma 4.1** (Minimax at a fixed gap, and the lower half)**.** *For every $e \in [0, \frac{1}{2}]$ and every $K$, $V_K(e)$ is the least element of $\mathcal{A}_K^{\mathrm{og}}(e)$: every policy has pseudoregret at least $V_K(e)$ in one world, and the* Bayes policy *at gap $e$, which at each history plays the branch attaining the* min *of Definition 2.5 and a fair coin when the two branches are equal, has exactly $V_K(e)$ in both worlds. Consequently every $v \in \mathcal{A}_K^{\mathrm{og}}$ satisfies $V_K(e) \le v$ for every $e \in [0, \frac{1}{2}]$.*

*Proof sketch.* The inequality $W_e^{\mathrm{og}}(n; x, y) \le x\, R_n^{\mathrm{og}}(q; 1, e)(h) + y\, R_n^{\mathrm{og}}(q; 2, e)(h)$ is proved by induction as in Theorem 3.2; the new step is that a mixture of the two branches cannot beat the better branch. The Bayes policy is defined by the recursion itself, so it attains equality by construction, and Feldman's theorem is not used. The fair coin at ties makes the policy equivariant under the arm swap, hence indifferent between the two worlds; Feldman's own tie-break (experiment $A$ when $\xi \ge \frac{1}{2}$) is Bayes-optimal but not indifferent, and gives no minimax statement by itself. This lemma is part of the formal development; it is the half of the sandwich that bounds every achievable bound from below. □

**Definition 4.2** (The frozen policy)**.** For $e_0 \in [0, \frac{1}{2}]$ and a history $h$ let $s_a, f_a$ be the numbers of successes and failures recorded on arm $a$, and $\sigma_{e_0}(a) = (1 + 2e_0)^{s_a}(1 - 2e_0)^{f_a}$, the posterior weight (up to a common factor) of the hypothesis that arm $a$ is the good one if the gap were $e_0$. The policy $\mathrm{frozen}_{e_0}$ pulls arm 1 if $\sigma_{e_0}(1) > \sigma_{e_0}(2)$, arm 2 if $\sigma_{e_0}(1) < \sigma_{e_0}(2)$, and each with probability $\frac{1}{2}$ at a tie.

This is Feldman's myopic rule with its posterior comparison frozen at one gap. It never looks at the true gap, so it is a legitimate learner against the whole class; it is equivariant under the arm swap, so its two pseudoregrets agree; and, the policy being independent of $e$, its pseudoregret $Q_K(e) = R_K^{\mathrm{og}}(\mathrm{frozen}_{e_0}; 1, e)$ is a polynomial in $e$ of degree at most $K$. If $Q_K(e) \le v$ on $[0, \frac{1}{2}]$ then $v \in \mathcal{A}_K^{\mathrm{og}}$; this is the upper half of the sandwich, and together with Lemma 4.1,

$$V_K(e_0) \;\le\; V^*(K) \;\le\; \max_{[0,1/2]} Q_K \qquad \text{for every rational } e_0. \tag{1}$$

Where the two ends meet, $V^*(K)$ is pinned exactly. At $e_0 = \frac{1}{2}$ the score is $2^{s_a}$ if $f_a = 0$ and 0 otherwise, so the policy reads: never pull an arm that has failed while the other has not; otherwise pull the arm with more successes; fair coin at a tie.

## 4.2. **The values.**

**Proposition 4.3** (The extreme gap, at every horizon)**.** *For every $n$ and all $x, y \ge 0$, $W_{1/2}^{\mathrm{og}}(n; x, y) = (1 - 2^{-n}) \min(x, y)$; hence, for every horizon $K$, $V_K(\frac{1}{2}) = \frac{1}{2} - 2^{-(K+1)}$.*

At $e = \frac{1}{2}$ the good arm never fails, so a single failure identifies the bad arm and the two-branch recursion collapses; the induction is two lines. The values $\frac{1}{4}, \frac{3}{8}, \frac{7}{16}, \frac{15}{32}, \ldots$ have numerators $2^K - 1$.

**Proposition 4.4** (The frozen policy at $e_0 = \frac{1}{2}$)**.** *For every real $e$,*

$$Q_1(e) = \tfrac{1}{2}e, \qquad Q_2(e) = e - \tfrac{1}{2}e^2, \qquad Q_3(e) = \tfrac{3}{2}e - e^2 - \tfrac{1}{2}e^3,$$

*where $Q_K(e) = R_K^{\mathrm{og}}(\mathrm{frozen}_{1/2}; 1, e)$.*

**Proposition 4.5** ($K \le 3$)**.** *$V^*(1) = \frac{1}{4}$, $V^*(2) = \frac{3}{8}$ and $V^*(3) = \frac{7}{16}$, each as the least element of $\mathcal{A}_K^{\mathrm{og}}$.*

At these horizons the adversary's best gap is the endpoint $e = \frac{1}{2}$: the three polynomials are maximised on $[0, \frac{1}{2}]$ at $\frac{1}{2}$ ($\frac{3}{8} - Q_2(e) = \frac{1}{2}(\frac{1}{2} - e)(\frac{3}{2} - e)$ and $\frac{7}{16} - Q_3(e) = (\frac{1}{2} - e)(\frac{7}{8} - \frac{5}{4}e - \frac{1}{2}e^2)$), where they meet $V_K(\frac{1}{2})$ of Proposition 4.3, and the sandwich (1) closes. The three values are elementary and no searched source prints them; they are recorded as routine.

**Theorem 4.6** ($K = 5$)**.** *$V^*(5) = \frac{27}{50}$ exactly, as the least element of $\mathcal{A}_5^{\mathrm{og}}$; it is attained at the gap $\frac{2}{5}$, i.e. at the arms $(\frac{9}{10}, \frac{1}{2})$. The frozen policy at $e_0 = \frac{2}{5}$ has pseudoregret $Q_5(e) = \frac{5}{2}e - \frac{19}{8}e^2 - \frac{5}{4}e^3$.*

*Proof sketch.* $Q_5'(e) = \frac{5}{2} - \frac{19}{4}e - \frac{15}{4}e^2$ has the exact root $e = \frac{2}{5}$, and

$$\tfrac{27}{50} - Q_5(e) = \tfrac{5}{4}\,(e - \tfrac{2}{5})^2\,(e + \tfrac{27}{10}) \;\geq 0 \quad \text{on } [0, \tfrac{1}{2}],$$

one polynomial identity, so $\frac{27}{50} \in \mathcal{A}_5^{\mathrm{og}}$. On the other side $V_5(\frac{2}{5}) = \frac{27}{50}$ by an exact-rational evaluation of the two-branch recursion of Definition 2.5, $4^5$ leaves, in the kernel; by Lemma 4.1 every achievable bound is at least $\frac{27}{50}$. The two halves meet from opposite directions. □

This is the one-good-arm analogue of $R^*(3) = \frac{9}{16}$: the largest horizon at which the answer is a small exact rational.

**Theorem 4.7** ($K = 4, 6, 7$)**.** *The frozen policies at* $e_0 = \frac{139}{320}, \frac{373}{1000}, \frac{439}{1250}$ *have pseudoregret*

$$\begin{aligned} Q_4(e) &= 2e - \tfrac{13}{8}e^2 - \tfrac{3}{4}e^3 - \tfrac{1}{2}e^4, \\ Q_6(e) &= 3e - \tfrac{25}{8}e^2 - 2e^3 + \tfrac{3}{2}e^5 + \tfrac{1}{2}e^6, \\ Q_7(e) &= \tfrac{7}{2}e - 4e^2 - \tfrac{81}{32}e^3 + \tfrac{1}{4}e^4 + \tfrac{7}{4}e^5 + \tfrac{7}{4}e^6 + e^7, \end{aligned}$$

*and, with* $V_4(\frac{139}{320}) = \frac{10126581039}{20971520000}$, $V_6(\frac{373}{1000}) = \frac{1185216732367722689}{2000000000000000000}$ *and* $V_7(\frac{439}{1250}) = \frac{1533705581119480917177}{2384185791015625000000}$ *as the exact lower ends, of which the decimals below are the roundings towards zero,*

$$\begin{aligned} V^*(4) &\in [0.482873012495040,\ 0.482873014765140], \\ V^*(6) &\in [0.592608366183861,\ 0.592608403400340], \\ V^*(7) &\in [0.643282745371176,\ 0.643282760257337], \end{aligned}$$

*in the sense that every element of* $\mathcal{A}_K^{\mathrm{og}}$ *is at least the lower end and the upper end belongs to* $\mathcal{A}_K^{\mathrm{og}}$*. The windows have widths* $2.3 \cdot 10^{-9}$*,* $3.7 \cdot 10^{-8}$ *and* $1.5 \cdot 10^{-8}$*.*

*Proof sketch.* The polynomials are the pseudoregret recursion of Definition 2.2 for the frozen policy unfolded over the full $4^K$ trajectory tree and normalised in the kernel. The lower ends are exact-rational evaluations of $W^{\mathrm{og}}$ at the stated gaps; the upper ends are certified by the identity of Section 5 with residual degrees $2, 4, 5$ and expansion points $e_1 = \frac{28346}{65261}, \frac{25390}{68053}, \frac{1761}{5015}$. Unlike the symmetric class the maximiser of $V_K$ is not available from a single gap-independent player, which is why the statement is a sandwich and not an evaluation. □

**Proposition 4.8** (Controls, and the cross-class contrast)**.** $\frac{7}{16} - \frac{1}{1000} \notin \mathcal{A}_3^{\mathrm{og}}$*;* $\frac{1}{2} \in \mathcal{A}_3^{\mathrm{og}}$ *but is not its least element;* $V_3(\frac{1}{4}) = \frac{39}{128} < \frac{7}{16}$*;* $V_4(\frac{1}{2}) = \frac{15}{32}$ *and* $V_4(\frac{1}{2}) < V_4(\frac{139}{320})$*, so the endpoint stops being the worst gap exactly at* $K = 4$*; if* $q(\emptyset) \in \{0, 1\}$ *then every bound valid at horizon* 1 *is at least* $\frac{1}{2}$*, twice the randomised value* $\frac{1}{4}$*; the always-arm-*1 *policy has* $R_3^{\mathrm{og}}(\,\cdot\,; 2, \frac{1}{2}) = \frac{3}{2}$ *and the constant* 2 *is not a policy. Finally, at the same horizons,* $\frac{1}{4}$ *and* $\frac{7}{16}$ *are the least elements of* $\mathcal{A}_1^{\mathrm{og}}$ *and* $\mathcal{A}_3^{\mathrm{og}}$ *while* $\frac{1}{2}$ *and* $\frac{9}{16}$ *are the least elements of* $\mathcal{A}_1$ *and* $\mathcal{A}_3$*, with* $\frac{1}{4} < \frac{1}{2}$ *and* $\frac{7}{16} < \frac{9}{16}$*.*

The last clause is checked in one statement so that a value for one adversary class cannot be read as a value for the other.

*Remark* 4.9 (Mixed adversaries). By Lemma 4.1, $\max_e V_K(e) \leq V^*(K)$ always. Equality would need a single policy optimal at every gap simultaneously, which the symmetric class has and this class does not, or a minimax theorem, which would identify $V^*(K)$ with a supremum over adversary *mixtures* on $[0, \frac{1}{2}]$, possibly larger than $\max_e V_K(e)$. No such equality is claimed or used: what is proved is (1), whatever the value of $\max_e V_K(e)$.

*Remark* 4.10 (A hard-MDP instance; not part of the formal development). Darwiche Domingues, Ménard, Kaufmann and Valko [6] prove their episodic lower bounds on a class of MDPs with parameters $(S, A, H, \bar{H}, d, L)$ (their Section 3.1; all numbering here is that of the arXiv version, the only one) and say in words that these MDPs "behave like multi-armed bandits with $\Theta(HSA)$ arms". The following statement upgrades this to an equality of game values; it is proved on paper and is

TABLE 2. The one-good-arm class, $K \le 7$. $e_0$ is the rational gap at which the lower end $V_K(e_0)$ is evaluated and at which the gap-free policy is frozen; $Q_K$ is that policy's exact pseudoregret. Exact where the sandwich closes, a two-sided window otherwise.

| $K$ | $e_0$ | $Q_K(e)$ | $V^*(K)$ |
|---|---|---|---|
| 1 | 1/2 | $\frac{1}{2}e$ | 1/4 |
| 2 | 1/2 | $e - \frac{1}{2}e^2$ | 3/8 |
| 3 | 1/2 | $\frac{3}{2}e - e^2 - \frac{1}{2}e^3$ | 7/16 |
| 4 | 139/320 | $2e - \frac{13}{8}e^2 - \frac{3}{4}e^3 - \frac{1}{2}e^4$ | [0.482873012495040, 0.482873014765140] |
| 5 | 2/5 | $\frac{5}{2}e - \frac{19}{8}e^2 - \frac{5}{4}e^3$ | $27/50 = 0.54$ |
| 6 | 373/1000 | $3e - \frac{25}{8}e^2 - 2e^3 + \frac{3}{2}e^5 + \frac{1}{2}e^6$ | [0.592608366183861, 0.592608403400340] |
| 7 | 439/1250 | $\frac{7}{2}e - 4e^2 - \frac{81}{32}e^3 + \frac{1}{4}e^4 + \frac{7}{4}e^5 + \frac{7}{4}e^6 + e^7$ | [0.643282745371176, 0.643282760257337] |

deliberately not formalised, since a $K$-episode learning game over trajectories buys no further mathematics. Set $n = \bar{H}LA$ and $c = H - \bar{H} - d \ge 1$. Then for every number of episodes $K$ and every fixed $\varepsilon \in [0, \frac{1}{2}]$, the $K$-episode minimax expected regret over randomised history-dependent policies (their Definition 1) equals $c \cdot V(n, K; \varepsilon)$, the value of the $n$-armed one-good-arm Bernoulli bandit at gap $\varepsilon$, and with $\varepsilon$ free it equals $c$ times the corresponding minimax over the class. At their Figure 1 instance $(S, A, H) = (4, 2, 3)$ with $\bar{H} = d = L = 1$ this is $n = 2$, $c = 1$, so the exact $K$-episode minimax regret of that instance is $V^*(K)$ of Table 2; at $(4, 2, 4)$ with $\bar{H} = 1$ it is $2V^*(K)$. The argument: within an episode every transition before the leaf is deterministic and identical across the class, so an episode is one arm pull; the leaf transition is the Bernoulli sample; the one extra option, staying in the waiting state through stage $\bar{H}$, is a null arm, uninformative in every instance and forfeiting a rewarding stage, hence strictly dominated; and the reference MDP has regret 0 under every policy.

## 5. Certificates and computations

5.1. **What rests on exhaustive computation.** Three kinds of finite computation are performed inside the kernel, and nothing else is computational. (i) The symmetric polynomials $P_1, \dots, P_8$: the recursion of Definition 2.4 is expanded by the single rewriting rule on pairs $(a^i b^j, a^j b^i)$ and normalised; the tree has $2^T$ leaves. (ii) The frozen-policy polynomials $Q_1, \dots, Q_7$: the pseudoregret recursion is unfolded over the full trajectory tree, $4^K$ leaves, with the policy's case split resolved at every node, and normalised. (iii) The one-good-arm values $V_K(e_0)$ at $e_0 = \frac{1}{2}$ (all $K$, by Proposition 4.3) and at the rational gaps $\frac{139}{320}, \frac{2}{5}, \frac{373}{1000}, \frac{439}{1250}$ for $K = 4, 5, 6, 7$: the two-branch recursion evaluated in exact rational arithmetic, $4^K$ leaves, every min decided by a rational comparison. The $K = 7$ evaluation is the expensive one and lives in its own module.

5.2. **The upper halves.** Let $F$ be $P_T$ or $Q_K$, let $e_1$ be a rational near the maximiser of $F$ on $[0, \frac{1}{2}]$ (a continued-fraction convergent), $D = F'(e_1)$, and $G(e) = (F(e) - F(e_1) - (e - e_1)D)/(e - e_1)^2$, an exact polynomial. Choose a rational $C < 0$ with $G \le C$ on $[0, \frac{1}{2}]$, put $k = -C$, $v = e_1 + D/(2k)$ and $U_0 = F(e_1) + D^2/(4k)$. Then, for any $U \ge U_0$, identically in $e$,

$$U - F(e) = (U - U_0) + k\,(e - v)^2 + (e - e_1)^2\big(C - G(e)\big), \qquad C - G(e) = \sum_{j=0}^{m} \beta_j \binom{m}{j} (2e)^j (1 - 2e)^{m-j},$$

where $C$ is taken to be the largest Bernstein coefficient of $G$ on $[0, \frac{1}{2}]$ at degree $m$, so that every $\beta_j \ge 0$ by construction and the last summand is visibly non-negative on $[0, \frac{1}{2}]$. The identity is verified by polynomial normalisation; the non-negativity of each summand is a product of non-negative factors. The double root at the maximiser is absorbed by the Taylor step, which is why low Bernstein degrees suffice: $m = 2, 4, 4, 6$ for $P_5, \dots, P_8$ and $m = 2, 4, 5$ for $Q_4, Q_6, Q_7$; at $T = 4$

an ad-hoc majorant $e^2 + 2e_1e + 3e_1^2 \le \frac{1}{4} + e_1 + 3e_1^2$ with slack $(\frac{1}{2} - e)(\frac{1}{2} + e + 2e_1)$ was used instead (Proposition 3.8). The window width is governed by $|D|$, hence by the quality of the convergent: denominators $8039, 17665, 8761, 7909, 12967$ give $|D|$ small enough for $2 \cdot 10^{-15}$ in the symmetric class.

5.3. **Independent checks run before the formal work.** For the symmetric class, four exact-rational implementations agree on $P_T$ for $T \le 8$ at seven rational gaps: the weight recursion; the closed form $P_T(e) = e \sum_{t<T} \sum_{u \le t} \binom{t}{u} a^{\min(u,t-u)} b^{\max(u,t-u)}$; enumeration of the $4^T$ trajectory tree under the myopic policy in each world ($T \le 6$); and backward induction over the full history tree with no state collapse ($T \le 6$). The mixed-strategy minimax itself was verified by enumerating all deterministic policies (2 at $T = 1$, 32 at $T = 2$) and minimising the worst case over their convex hull; at $T = 3$ the $2^{21}$ deterministic policies were not enumerated. For the one-good-arm class, three implementations agree on every value of Table 2: two dynamic programmes with different state encodings, and brute force over all deterministic history-dependent policies — exact over the rationals at $K \le 2$, and at $K = 3$ over all $2^{21}$ policies in floating point at $e = \frac{1}{2}$ and $e = 0.4$, agreeing with the recursion to $5.6 \cdot 10^{-17}$. None of these checks is part of the proof.

## 6. The printed record

The numbering and quotations below were read from the arXiv PDFs, or, for the journal papers, from page images of the scans; no number was counted from a source's raw manuscript.

*Kobzar and Kohn* [11], v5. Their Lemma 3.1: "The player $p^m$ given by (3.3) is a minimizer of (1.3) and (1.6). Moreover, this strategy makes the player indifferent about which arm is risky"; their footnote 4: "an optimal player is the same for all feasible values of the gap $\epsilon$". Their Section 1: "we appear to be the first to give a proof of its optimality in the minimax setting … and the optimal regret has not been determined previously"; their Section 4: "We are not aware of the leading order terms of the minimax optimal regret or pseudoregret having being determined previously (as opposed to Bayesian pseudoregret) in the symmetric version of the problem." Their rewards are centred $\pm 1$ from Section 3.1 on, so their constants are twice those of the 0/1 scaling used here, as they say in print. They do not cite Fabius and van Zwet, and no finite-horizon value appears in the paper.

*Fabius and van Zwet* [7]. The $N = 3$ and $N = 4$ quotations are given after Theorem 3.5 and Proposition 3.8; the bracket $0.265N^{1/2}$–$0.376N^{1/2}$ is on p. 1907, the lower constant being obtained by "applying the same method that was used in [5] to the optimal symmetric strategy for $\alpha + \beta = 1$ that was discussed in [4]", their [4] being Vogel [15] and their [5] Vogel [16]. Their Section 2, "Bayes strategies" (pp. 1908–1911), was read for this draft, page by page, from the same scan; it is the part of the paper closest to the one-good-arm results here, because the symmetric two-point prior at $(\alpha, \beta) = (\frac{1}{2} + e, \frac{1}{2})$ is one of the priors it treats and their loss convention is the present one. It contains no number. Its content is structural: their Theorem 1 is a backward recurrence for functions $c_{\alpha,\beta,\Delta}(m, k; n, l)$ through which the coefficient of each randomisation $\delta(m, k; n, l)$ in the expected number of successes is expressed; their Theorem 2 characterises the Bayes strategies against an arbitrary prior $\mu$ on the closed unit square by the sign of an integral $\gamma_\mu(m, k; n, l)$, which p. 1907 describes as "thus generalizing D. Feldman's well-known result in [3]"; and their Theorem 3 prints a closed form for the Bayes risk $\rho(\mu)$ of an arbitrary prior. So the machinery to express $P_T(e)$ and $V_K(e)$ as Bayes risks of two-point priors is in print in 1970 — their Theorem 3 applied to those priors is a formula for both — but no evaluation of it appears at any finite $N$, in Section 2 or anywhere in the paper except the $N = 3$ and $N = 4$ risk functions of p. 1916 already quoted. The residual-novelty flag that stood over Theorems 4.6 and 4.7 while that section was unread is therefore removed.

*Feldman* [8], read in the scan, all ten pages. His Theorem 2.1, p. 851: "For every $N$ the procedure $\pi_N^*$, which chooses $A$ or $B$ at stage $i$ according as $\xi_{i-1} \ge \frac{1}{2}$ or $\xi_{i-1} < \frac{1}{2}$, is optimal." His prior class, the symmetric two-point prior on the ordered pair $(p_1, p_2)/(p_2, p_1)$, covers both classes here; his loss is the expected number of pulls of the inferior arm, the pseudoregret divided by the gap. He prints

no number: no table, no decimal, no risk polynomial at any finite $N$; his Section 4 computes only the $N = \infty$ risk. Keener [10], the literal follow-up, treats the infinite horizon and prints no finite-$N$ value (abstract and zbMATH review read; full text not obtained). Bradt, Johnson and Karlin [4] print Bayes examples on particular priors and no minimax value.

*Auer, Cesa-Bianchi, Freund and Schapire* [2], read from the authors' preprint as page images; the numbering in this paragraph is the preprint's. Appendix A (p. 31) defines the one-good-arm class verbatim: "one action $I$ is chosen uniformly at random to be the 'good' action. The $T$ rewards $x_I(t)$ associated with the good action are chosen independently at random to be 1 with probability $1/2 + \epsilon$ and 0 otherwise … The rewards $x_j(t)$ associated with the other actions $j \neq I$ are chosen independently at random to be 0 or 1 with equal odds." Theorem 5.1 (p. 13) prints, for the expected weak regret of any algorithm, the bound "at least" $\frac{1}{20}\min\{\sqrt{KT}, T\}$, and no exact finite-horizon value. The paper of Auer, Cesa-Bianchi and Fischer [1] is a different paper and is not a source of this class: its theorems are upper bounds.

*Bubeck and Cesa-Bianchi* [5], v2. There is no Theorem 3.5 in this version; the minimax lower bound is Theorem 3.4, p. 33, with the constant $\frac{1}{20}$. Its formal class, Lemma 3.2, p. 34, is the centred one: "all arms are i.i.d. Bernoulli of parameter $(1-\epsilon)/2$ but arm $i$, which is i.i.d. Bernoulli of parameter $(1+\epsilon)/2$", which for two arms is the symmetric class of Section 3; only the informal sketch on the same page uses the one-good-arm class. So the two standard lower-bound arguments use different ones of the two classes treated here.

*Lattimore and Szepesvári* [13]. Exercise 15.4 asks to show that $R_n^*(\mathcal{E}_B^k) \geq c\sqrt{nk}$ for a universal $c > 0$ that is never printed; Theorem 15.2, with the constant $\frac{1}{27}$, is the Gaussian statement. $R^*(n)$ bounds $R_n^*(\mathcal{E}_B^2)$ from below and says nothing about an admissible $c$.

*Searches.* No exact finite-horizon minimax value for either class was found in a full-text search of a local corpus of 1,153,045 arXiv papers (phrases such as "symmetric two-armed", "finite-horizon minimax regret", "one good arm", "minimax pseudoregret", and the decimal fingerprints of the values, against positive controls), in the live arXiv API, in the OEIS (no entry mentions bandits; the leading coefficients of $P_T$ appear in A028329 and A240530, the numerators of $V_K(\frac{1}{2})$ in A000225 [14]), in the citing papers of [11] listed by Semantic Scholar, all read in full, or in Kolnogorov's minimax-risk series [12], which treats Gaussian and Poisson arms asymptotically.

*Three items not obtained.* Vogel [15], which Fabius and van Zwet say discussed the optimal symmetric strategy for $\alpha + \beta = 1$, is behind a publisher paywall; its bibliographic data are verified against the publisher's record and against Fabius and van Zwet's own reference list, but its text was not read, and it remains the most plausible earlier home for the polynomials $P_T$. Bather [3], Kobzar and Kohn's [2], was likewise not obtained; its bibliographic data are verified against the publisher's record and zbMATH, and both Kobzar and Kohn (who cite it for the leading term of the Bayesian pseudoregret) and its zbMATH review (which reports an asymptotic minimax lower bound $\liminf_t M(t)/\sqrt{t} \geq .306$ over the whole square) describe it as asymptotic, but that is a secondary description and not a reading of the paper. Finally, a manuscript of Kobzar, *The Symmetric Two-Armed Bandit: the General Case*, dated 2026 and marked "submitted" in the bibliography of Guillen and Kobzar, arXiv:2607.08943, is not posted on arXiv and is not indexed in OpenAlex; nothing about its contents could be checked.

## 7. Verification

Every theorem and proposition above, with the exception of Remark 3.12's one-line consequence and the three remarks whose headings mark them as outside the formal development (Remarks 3.13, 3.14 and 4.10), is a theorem of Lean 4 with Mathlib, in ten modules: five for the symmetric class (the game and the backward induction, the values at $T \leq 3$, the $T = 4$ enclosure, the enclosures at $5 \leq T \leq 8$, and the controls) and five for the one-good-arm class (the game and the sandwich, the extreme gap and $K \leq 3$, the values at $K = 4, 5, 6$, the value at $K = 7$, and the controls). The games are built from scratch over $\mathbb{R}$ as finite sums over trajectory trees; no measure theory, no

probability library and no floating point is used, and no proof appeals to native code evaluation. Every polynomial identity is closed by normalisation in the commutative ring $\mathbb{R}[e]$ and every rational value by exact arithmetic; the modules for $K \geq 5$ raise the elaborator's step budget and nothing else. The axioms reported by `#print axioms` are exactly `propext`, `Classical.choice` and `Quot.sound` for each of the 51 declarations tagged to a result of this paper and for each of the 10 further lemmas through which those results are stated and proved — 61 in all, 28 in the symmetric development and 33 in the one-good-arm development; this was re-checked for this draft. The formal statements are reproduced in Appendix A. Repository reference to be added at submission.

## Appendix A. The formal statements

The Lean statements are reproduced verbatim (line breaks ours), grouped by result. `R q w e T []` is $R_T(q; w, e)$ with `w = true` for world 1; `P T e`, `Pmax T e`, `V K e` are $P_T(e)$, $P_T^{\max}(e)$, $V_K(e)$; `AchievableBound T v` is $v \in \mathcal{A}_T$; `Admissible q` says $q$ takes values in $[0, 1]$; the prefix `OneGood.` marks the one-good-arm development, whose `P_one`, …, `P_seven` are the polynomials written $Q_K$ above.

**Theorems 3.1, 3.2, 3.5, Propositions 3.3, 3.4.**

```
theorem myopic_value (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) (T : ℕ) (w : Bool) :
    R myopic w e T [] = P T e
theorem minimax_at_gap (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) (T : ℕ) :
    IsLeast {v : ℝ | ∃ q, Admissible q ∧ ∀ w : Bool, R q w e T [] ≤ v} (P T e)
theorem minimax_over_gaps (T : ℕ) (e₀ : ℝ) (h0 : 0 ≤ e₀) (h1 : e₀ ≤ 1 / 2)
    (hmax : ∀ e : ℝ, 0 ≤ e → e ≤ 1 / 2 → P T e ≤ P T e₀) :
    IsLeast {v : ℝ | AchievableBound T v} (P T e₀)
theorem P_one (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) : P 1 e = e
theorem P_two (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) : P 2 e = 2 * e - 2 * e ^ 2
theorem P_three (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) : P 3 e = 3 * e - 4 * e ^ 2
theorem P_four (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) :
    P 4 e = 4 * e - 7 * e ^ 2 + 4 * e ^ 4
theorem minimax_one : IsLeast {v : ℝ | AchievableBound 1 v} (1 / 2)
theorem minimax_two : IsLeast {v : ℝ | AchievableBound 2 v} (1 / 2)
theorem minimax_three : IsLeast {v : ℝ | AchievableBound 3 v} (9 / 16)
```

**Propositions 3.6, 3.7, 3.8, Theorems 3.9, 3.10, 3.11.**

```
theorem not_achievable_below_three : ¬ AchievableBound 3 (9 / 16 - 1 / 1000)
theorem not_isLeast_above_three : ¬ IsLeast {v : ℝ | AchievableBound 3 v} (5 / 8)
theorem wrong_gap_quarter : P 3 (1 / 4 : ℝ) = 1 / 2 ∧ (1 / 2 : ℝ) < 9 / 16
theorem wrong_gap_half : P 3 (1 / 2 : ℝ) = 1 / 2 ∧ (1 / 2 : ℝ) < 9 / 16
theorem det_horizon_one (q : Hist → ℝ) (hdet : q [] = 0 ∨ q [] = 1) (v : ℝ)
    (hb : ∀ e : ℝ, 0 ≤ e → e ≤ 1 / 2 → ∀ w : Bool, R q w e 1 [] ≤ v) : 1 ≤ v
theorem always_arm_one : Admissible (fun _ => (1 : ℝ)) ∧
    R (fun _ => (1 : ℝ)) false (1 / 2 : ℝ) 3 [] = 3
theorem not_admissible_two : ¬ Admissible (fun _ => (2 : ℝ))
theorem minimax_four_enclosure :
    (∀ v : ℝ, AchievableBound 4 v → 605251593991880 / 1000000000000000 ≤ v) ∧
      AchievableBound 4 (605251593991882 / 1000000000000000)
theorem P_five (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) :
    P 5 e = (5) * e + (-10) * e ^ 2 + (8) * e ^ 4
theorem P_six (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) :
    P 6 e = (6) * e + (-55/4) * e ^ 2 + (18) * e ^ 4 + (-12) * e ^ 6
theorem P_seven (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) :
    P 7 e = (7) * e + (-35/2) * e ^ 2 + (28) * e ^ 4 + (-24) * e ^ 6
theorem P_eight (e : ℝ) (he : 0 ≤ e) (he2 : e ≤ 1 / 2) :
    P 8 e = (8) * e + (-175/8) * e ^ 2 + (91/2) * e ^ 4 + (-66) * e ^ 6 + (40) * e ^ 8
theorem minimax_five_enclosure :
    (∀ v : ℝ, AchievableBound 5 v → 665598466061581 / 1000000000000000 ≤ v) ∧
      AchievableBound 5 (665598466061582 / 1000000000000000)
theorem minimax_six_enclosure :
    (∀ v : ℝ, AchievableBound 6 v → 709079607684664 / 1000000000000000 ≤ v) ∧
```

```
      AchievableBound 6 (709079607684666 / 1000000000000000)
theorem minimax_seven_enclosure :
    (∀ v : ℝ, AchievableBound 7 v → 760342288803075 / 1000000000000000 ≤ v) ∧
      AchievableBound 7 (760342288803077 / 1000000000000000)
theorem minimax_eight_enclosure :
    (∀ v : ℝ, AchievableBound 8 v → 800891684469692 / 1000000000000000 ≤ v) ∧
      AchievableBound 8 (800891684469694 / 1000000000000000)
theorem P_add_Pmax (T : ℕ) (e : ℝ) : P T e + Pmax T e = 2 * (T : ℝ) * e
```

### Propositions 4.3, 4.4, 4.5, Theorems 4.6, 4.7.

```
theorem OneGood.V_half (K : ℕ) : V K (1 / 2 : ℝ) = 1 / 2 - (1 / 2) ^ (K + 1)
theorem OneGood.P_one (e : ℝ) : R (frozen (1 / 2)) true e 1 [] = (1 / 2) * e
theorem OneGood.P_two (e : ℝ) : R (frozen (1 / 2)) true e 2 [] = e - (1 / 2) * e ^ 2
theorem OneGood.P_three (e : ℝ) :
    R (frozen (1 / 2)) true e 3 [] = (3 / 2) * e - e ^ 2 - (1 / 2) * e ^ 3
theorem OneGood.minimax_one : IsLeast {v : ℝ | AchievableBound 1 v} (1 / 4)
theorem OneGood.minimax_two : IsLeast {v : ℝ | AchievableBound 2 v} (3 / 8)
theorem OneGood.minimax_three : IsLeast {v : ℝ | AchievableBound 3 v} (7 / 16)
theorem OneGood.P_five (e : ℝ) :
    R (frozen (2/5)) true e 5 [] = (5/2) * e + (-19/8) * e ^ 2 + (-5/4) * e ^ 3
theorem OneGood.minimax_five : IsLeast {v : ℝ | AchievableBound 5 v} (27/50)
theorem OneGood.P_four (e : ℝ) :
    R (frozen (139/320)) true e 4 []
      = (2) * e + (-13/8) * e ^ 2 + (-3/4) * e ^ 3 + (-1/2) * e ^ 4
theorem OneGood.minimax_four_enclosure :
    (∀ v : ℝ, AchievableBound 4 v → 10126581039/20971520000 ≤ v) ∧
      AchievableBound 4 (24143650738257/50000000000000)
theorem OneGood.P_six (e : ℝ) :
    R (frozen (373/1000)) true e 6 []
      = (3) * e + (-25/8) * e ^ 2 + (-2) * e ^ 3 + (3/2) * e ^ 5 + (1/2) * e ^ 6
theorem OneGood.minimax_six_enclosure :
    (∀ v : ℝ, AchievableBound 6 v → 1185216732367722689/2000000000000000000 ≤ v) ∧
      AchievableBound 6 (29630420170017/50000000000000)
theorem OneGood.P_seven (e : ℝ) :
    R (frozen (439/1250)) true e 7 []
      = (7/2) * e + (-4) * e ^ 2 + (-81/32) * e ^ 3 + (1/4) * e ^ 4 + (7/4) * e ^ 5
        + (7/4) * e ^ 6 + (1) * e ^ 7
theorem OneGood.minimax_seven_enclosure :
    (∀ v : ℝ, AchievableBound 7 v
        → 153370558111948091717 7/238418579101562500000 0 ≤ v) ∧
      AchievableBound 7 (643282760257337/1000000000000000)
```

### Proposition 4.8.

```
theorem OneGood.og_not_achievable_below_three : ¬ AchievableBound 3 (7 / 16 - 1 / 1000)
theorem OneGood.og_not_isLeast_above_three :
    ¬ IsLeast {v : ℝ | AchievableBound 3 v} (1 / 2)
theorem OneGood.og_wrong_gap_quarter :
    V 3 (1 / 4 : ℝ) = 39 / 128 ∧ (39 / 128 : ℝ) < 7 / 16
theorem OneGood.og_boundary_not_worst :
    V 4 (1 / 2 : ℝ) = 15 / 32 ∧ V 4 (1 / 2 : ℝ) < V 4 (139 / 320 : ℝ)
theorem OneGood.og_det_horizon_one (q : Hist → ℝ) (hdet : q [] = 0 ∨ q [] = 1) (v : ℝ)
    (hb : ∀ e : ℝ, 0 ≤ e → e ≤ 1 / 2 → ∀ w : Bool, R q w e 1 [] ≤ v) : 1 / 2 ≤ v
theorem OneGood.og_always_arm_one : Admissible (fun _ => (1 : ℝ)) ∧
    R (fun _ => (1 : ℝ)) false (1 / 2 : ℝ) 3 [] = 3 / 2
theorem OneGood.og_not_admissible_two : ¬ Admissible (fun _ => (2 : ℝ))
theorem OneGood.og_vs_symmetric_one :
    IsLeast {v : ℝ | AchievableBound 1 v} (1 / 4) ∧
      IsLeast {v : ℝ | _root_.LeanProblemSpec.Sym2abMinimax.AchievableBound 1 v} (1 / 2) ∧
      (1 / 4 : ℝ) < 1 / 2
theorem OneGood.og_vs_symmetric_three :
    IsLeast {v : ℝ | AchievableBound 3 v} (7 / 16) ∧
```

```
IsLeast {v : ℝ | _root_.LeanProblemSpec.Sym2abMinimax.AchievableBound 3 v} (9 / 16) ∧
(7 / 16 : ℝ) < 9 / 16
```

## References


[1] P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem, *Machine Learning* 47 (2002), 235–256.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem, *SIAM J. Comput.* 32 (2002), no. 1, 48–77.

[3] J. A. Bather, The minimax risk for the two-armed bandit problem, in: *Mathematical Learning Models — Theory and Algorithms* (Proc. Conf., Bad Honnef, 1982), Lecture Notes in Statistics 20, Springer, New York, 1983, 1–11.

[4] R. N. Bradt, S. M. Johnson, S. Karlin, On sequential designs for maximizing the sum of $n$ observations, *Ann. Math. Statist.* 27 (1956), 1060–1074.

[5] S. Bubeck, N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems, arXiv:1204.5721v2 (2012).

[6] O. Darwiche Domingues, P. Ménard, E. Kaufmann, M. Valko, Episodic reinforcement learning in finite MDPs: minimax lower bounds revisited, *Proc. Algorithmic Learning Theory (ALT 2021)*, 578–598; arXiv:2010.03531 (2020).

[7] J. Fabius, W. R. van Zwet, Some remarks on the two-armed bandit, *Ann. Math. Statist.* 41 (1970), 1906–1916.

[8] D. Feldman, Contributions to the "two-armed bandit" problem, *Ann. Math. Statist.* 33 (1962), 847–856.

[9] 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:1906.10173 (2019).

[10] R. Keener, Further contributions to the "two-armed bandit" problem, *Ann. Statist.* 13 (1985), 418–422.

[11] V. A. Kobzar, R. V. Kohn, A PDE-based analysis of the symmetric two-armed Bernoulli bandit, arXiv:2202.05767, version 5 (2023).

[12] A. V. Kolnogorov, Batch data processing and Gaussian two-armed bandit, arXiv:1704.03631 (2017).

[13] T. Lattimore, C. Szepesvári, *Bandit Algorithms*, Cambridge University Press, 2020.

[14] The OEIS Foundation, *The On-Line Encyclopedia of Integer Sequences*, entries A000225, A028329, A240530, `https://oeis.org`.

[15] W. Vogel, Ein Irrfahrten-Problem und seine Anwendung auf die Theorie der sequentiellen Versuchs-Pläne, *Arch. Math.* 11 (1960), 310–320.

[16] W. Vogel, An asymptotic minimax theorem for the two armed bandit problem, *Ann. Math. Statist.* 31 (1960), no. 2, 444–451.