Betting Against the Noise

Can reinforcement learning rediscover Kelly bet-sizing
on blackjack's thin, countable edge?

A learned DQN bettor vs. the analytic Kelly ladder, measured over sessions on four axes

Part of AI Journey — Deep Learning & Reinforcement Learning

arda-basarici.github.io/blackjack-betting

Executive summary

Blackjack betting has a rare property for a learning problem: **the optimal rule is derivable.** Once you count cards, the edge at each true count is measurable, and the growth-optimal wager is the Kelly bet. That makes it a clean test of a question that recurs everywhere in applied ML: *can an end-to-end learner rediscover a rule we could also just derive?* We trained a DQN bettor on raw per-hand log-growth reward and measured it against the analytic Kelly ladder and a flat-betting floor, on identical terms, across seeds, in two bankroll regimes.

It does not rediscover Kelly — not from raw hand rewards, not within this experiment's sample budget. The learned bettor converges to approximately flat betting in the growth regime and slightly below flat in the ruin regime; the near-Kelly bet curves it transiently produces measure *worse* than flat. Only the analytic Kelly ladder beats the flat floor — and only barely.

regime	policy	growth /hand (×1e-4)	ruin %	deep-drawdown %
growth	learned BetAgent	-0.19 ± 0.08	0.00	0.18
growth	analytic Kelly	-0.048	0	0.14
growth	Flat (floor)	-0.150	0	0.00
ruin	learned BetAgent	-0.48 ± 0.04	0.03	1.39
ruin	analytic Kelly	-0.348	0	2.34
ruin	Flat (floor)	-0.395	0	0.80

Scoreboard. Agent = mean ± SD over 6 training seeds (2,000 sessions each); Kelly/Flat = the 20,000-session committed baseline. Native bankroll regime for every policy. Ruin is ~zero for every policy shown — it is an over-betting phenomenon (Section 2) — and no policy in this report is judged on growth alone.

The reason is not a bug, a weak architecture, or a bad state encoding — each alternative was isolated and ruled out, twice with controlled experiments the report walks through. The reason is the signal: the entire prize for perfect bet-sizing is ~0.10×10^{-4} log-growth per hand, buried in per-hand noise fifty times larger, carried by counts the shoe rarely deals. Structure — a Kelly rule read off an edge curve measured once, offline — extracts it; end-to-end value learning, re-estimating the same curve through that noise while acting, did not come close, and the sample arithmetic of Section 6 says why. **The practical lesson: when a thin signal has a derivable form, encode it; don't ask a value learner to rediscover it.**

The fourth movement of the project

This report closes a four-part blackjack investigation. The first movement built a Monte Carlo simulator — fixed rules, reproducible seeds, pluggable strategies, millions of hands. The second trained a tabular Monte Carlo agent on it and audited the learned policy cell-by-cell against basic strategy; the finding was that its errors concentrated in rare, coverage-starved cells. The third replaced the table with a DQN and asked whether neural generalization repairs that coverage problem — it does, partially, at the price of distortions a table cannot make.

All three movements share a limitation: they end at the hand. Playing decisions were measured against basic strategy — a reference that is *optimal per hand* but silent about money. Betting is where blackjack actually pays: count the shoe, bet small when the edge is against you, bet big on the rare counts where it turns. And betting brings the one thing the earlier movements never had — a *derivable optimum*. Given a measured edge curve, the growth-optimal bet is the Kelly fraction; no learning required. So the final movement can ask its question cleanly:

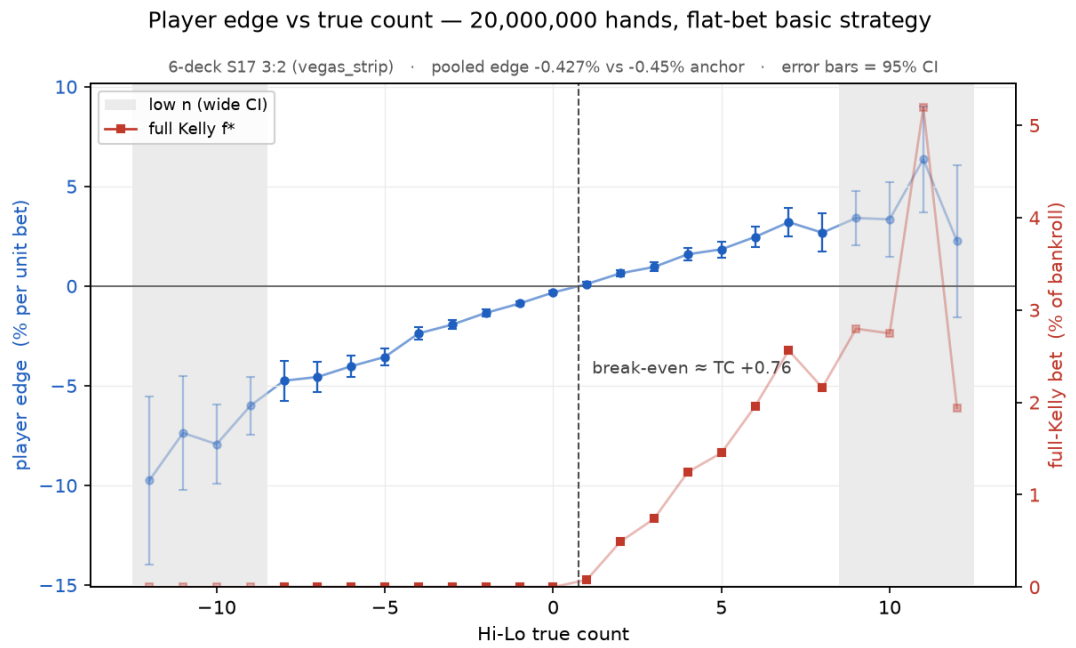
When the optimal rule can be derived, can an end-to-end reinforcement learner rediscover it from raw reward — and if not, what exactly stops it?

The answer occupies the rest of this report. Sections 1–2 build the instrument: the measured edge, the Kelly reference, and what perfect betting is actually worth. Sections 3–5 run the experiment and read the result. Sections 6–7 establish *why* — including a hypothesis of our own that we built two controlled experiments to test, and falsified. Sections 8–9 close the project.

1. The instrument — a countable edge, and a known optimum

The unit of play is a **session**: a bankroll (in table-minimum units) played through 1,000 hands of six-deck blackjack under fixed rules, with basic-strategy play throughout. Playing decisions are frozen — the only free choice is the **wager**, picked before each hand from an arithmetic 1–8 unit spread. Per-hand reward is the log-increment of wealth, so a session's total reward telescopes to log(final/starting bankroll): maximizing it is maximizing compound growth.

The bettor's information is the **Hi-Lo true count**. High counts mean a ten/ace-rich shoe — more player blackjacks at 3:2, better doubles — and the classic counting claim is that the player edge rises roughly half a percent per true count. We did not take that on faith: the edge was **measured**, 20 million basic-strategy hands binned by true count (20 parallel workers, per-worker seeds, merged variance accumulators). Break-even lands at TC +0.76; the neutral count carries -0.32% and the rich counts reach +2.5% at TC +6. This measured curve — not a textbook table — is the reference every result below is audited against.



The signature measurement: player edge vs Hi-Lo true count, 95% CIs, with the implied full-Kelly fraction (right axis) and the break-even count. 20,000,000 hands, committed reference (git eab466d). Low-n tails shaded.

The yardstick: Kelly betting (derivation in Appendix A)

Choosing log-growth as the objective is a *risk preference*, not a mathematical necessity — expected-value maximization says bet everything on any positive edge and goes broke on the way. Given that preference, the optimal fraction of bankroll is the **Kelly bet**: to second order, f* = edge / variance of the per-hand return, floored at zero when the edge is negative.

From the measured curve, full Kelly is small: f* ≈ 0.5% of bankroll at TC +2, 1.2% at +4, 2.0% at +6. At a 400-unit bankroll that is bets of roughly 2, 5, and 8 units — which is why the 1–8 arithmetic spread and the ~400u bankroll were *derived together* from the curve, not assumed. Snapped to the spread, the discrete-Kelly ladder over counts (-4…+8) is **1 1 1 2 5 8 8**.

Two refinements matter later: **fractional Kelly** (betting a fraction of f* costs growth only to second order but cuts volatility linearly), and the **ruin-aware bend** (with a hard ruin barrier, the optimum bets below Kelly as wealth approaches it). Appendix A derives both.

Everything runs in two **bankroll regimes**, because bankroll changes what betting is about. The **growth** regime starts rich (400u): ruin is effectively out of reach and the pure question is growth-optimal sizing. The **ruin** regime starts lean (200u) with a hard ruin barrier: over-betting can now end the session, so restraint carries survival value. Each policy is always scored in its native regime.

2. The analytic ladder — what perfect betting is worth

Before any learning, the analytic baselines set the scale. Three reference bettors — **Flat** (always one unit: no betting skill), **discrete Kelly** (the ladder above: perfect skill on the same 1–8 menu), and **flat-8** (always the maximum: reckless aggression) — were each measured over 20,000 sessions per regime. A bettor is never one number here; four axes are held apart throughout: **growth rate** (log-growth per hand, the objective), **ruin %** (the catastrophe), **deep-drawdown %** (losing half the bankroll — pain short of ruin), and the final-bankroll distribution.

regime	bettor	growth /hand (×1e-4)	ruin %	deep-drawdown %
growth (400u)	Flat	-0.150	0	0.00
growth (400u)	Kelly (discrete)	-0.048	0	0.14
growth (400u)	flat-8 (max)	—	20	—
ruin (200u)	Flat	-0.395	0	0.80
ruin (200u)	Kelly (discrete)	-0.348	0	2.34
ruin (200u)	flat-8 (max)	+4.67	53	—

The committed 20k-session bet ladder. flat-8 growth in the growth regime and drawdown columns for flat-8 omitted — ruin dominates every other axis for it.

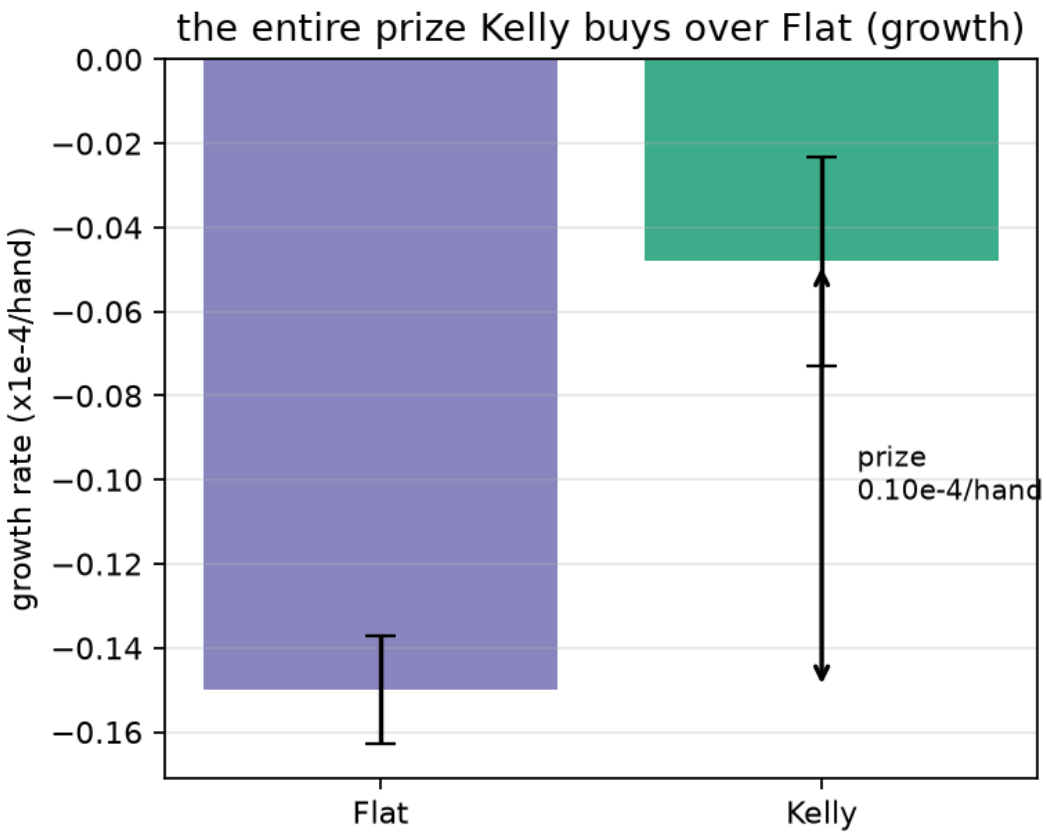
Three findings anchor the whole report. **First, the betting lever works, but the prize is tiny.** Kelly beats Flat on growth in the growth regime by ~0.10×10^{-4} per hand (unpaired $z \approx 7$, decisive); in the ruin regime the edge is real but marginal (~0.05×10^{-4}, $p \approx 0.04$) and bought with more drawdown. **Second, even perfect betting loses money here.** Kelly's growth is negative in both regimes — the table minimum forces a mandatory bet on the frequent negative-edge hands, a tax no sizing rule escapes. 'Kelly beats Flat' means *loses less*; the skill being priced is restraint, not profit. **Third, ruin is purely an over-betting phenomenon.** Kelly never ruins, even on the lean bankroll; flat-8 ruins 20% of sessions rich and 53% lean.

Kelly is optimal under the constraint — and the constraint can still make the game losing.

The flat-8 row is also this project's best argument for the four-axis discipline. In the ruin regime it posts the *best* growth number on the board (+4.67) while ruining 53% of sessions — the growth is pure survivor bias, averaged over the half that lived. Collapse the axes into one score and the worst bettor looks like the best.

A policy can look best on growth and still be unacceptable once ruin is shown — no policy in this report is judged on growth alone.

One sidebar quantifies the tax. Allowing the bettor to *sit out* negative counts (min-wager zero — 'Wonging', after Stanford Wong) flips continuous-Kelly growth from -0.062 to **+0.143**$\times10^{-4}$ at 400u, and from -0.314 to **+0.136** at 200u — counting genuinely beats the house once you are not forced to bet its hands, and with *less* drawdown. This matters in Section 9: it is the analytic answer to a learning experiment we scoped but did not run.



The prize, drawn to scale: Kelly vs Flat growth in the growth regime, 20k-session 95% CIs. The gap between the bars — about 0.10×1e-4 log-growth per hand — is the ENTIRE learnable sizing prize: everything perfect count-based bet-sizing is worth on this table.

3. The question — can a value learner rediscover it?

The learner is deliberately plain: a DQN (two hidden layers of 64) that sees **(true count, decks remaining, bankroll)** and outputs a Q-value per bet level; the greedy argmax is the wager. Reward is the raw per-hand log-growth — exactly the objective Kelly maximizes, so if value learning succeeds, the greedy policy *is* the Kelly ladder. No reward shaping, no curriculum, no oracle hints: the point is to test end-to-end learning on honest terms. Training uses the standard apparatus — replay buffer, target network, Huber loss — hardened over the investigation (reward scaling against the Huber knee, harmonic learning-rate decay, epsilon decay, batch and γ sweeps; the ruin regime gets $\gamma = 0.95$ and the growth regime the myopic $\gamma = 0$, which is optimal there because Kelly is a per-hand optimum).

What is being learned here (vs. the previous report)

Play is frozen to basic strategy — unlike the previous report, no playing decision is learned. **Only the wager is learned.**

State: (Hi-Lo true count, decks remaining, bankroll) · **Actions:** bet sizes 1–8 units · **Reward:** per-hand log-growth of wealth.

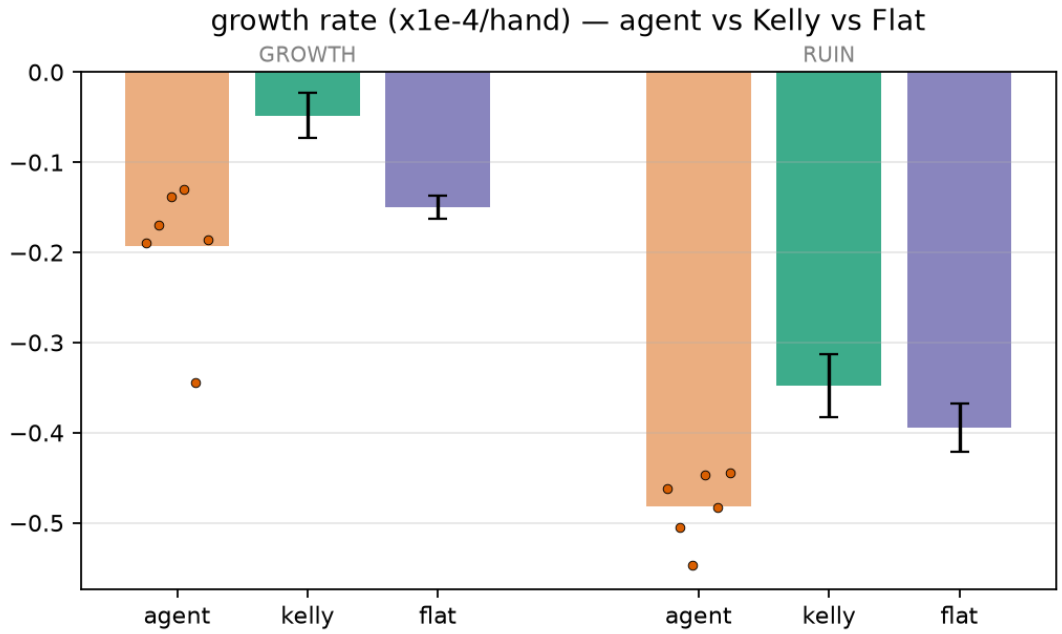
Success criterion: the greedy bet-by-count curve matches the discrete-Kelly ladder (1 1 1 2 5 8 8) *and* the four-axis evaluation matches Kelly's.

Every result reads off the same **three-rung ladder**: Flat is the floor (no skill), analytic Kelly is the ceiling (the derivable optimum), and the BetAgent is the question. '≈ Flat' means no betting skill was learned; '≈ Kelly' would mean the rule was rediscovered. The measurement protocol: each configuration trains under **6 random seeds**; each seed is scored over 2,000 fresh sessions on the four axes; agents carry the SD across seeds, the deterministic baselines carry their 20k-session Monte-Carlo CIs; bettors play independent shoe streams, so comparisons use unpaired z-tests. Figures show seeds as dots rather than whiskers — at n = 6 an SD bar would hide exactly the seed-to-seed story that matters.

One trap shaped the whole evaluation design, and it recurs so often it deserves stating up front: ***a bet curve that looks like Kelly at one probe can be a terrible policy once measured.*** *The report never trusts an eyeballed curve; verdicts come only from the four-axis evaluation at the native bankroll.*

4. The result — never Kelly

The scoreboard in the executive summary is the verdict; here is how to read it. In the **growth** regime the agent lands at -0.19 ± 0.08 against Flat's -0.150 — statistically indistinguishable from the no-skill floor (z ≈ 1.3, n.s.) and nowhere near Kelly's -0.048. In the **ruin** regime it lands *below* the floor: -0.48 vs -0.395 (z ≈ 4 across seeds), with more drawdown — no skill gained, and a small price paid for trying. This holds across every hardening cell (double-DQN on/off, the γ sweep, three state encodings): no configuration approaches the Kelly rung.

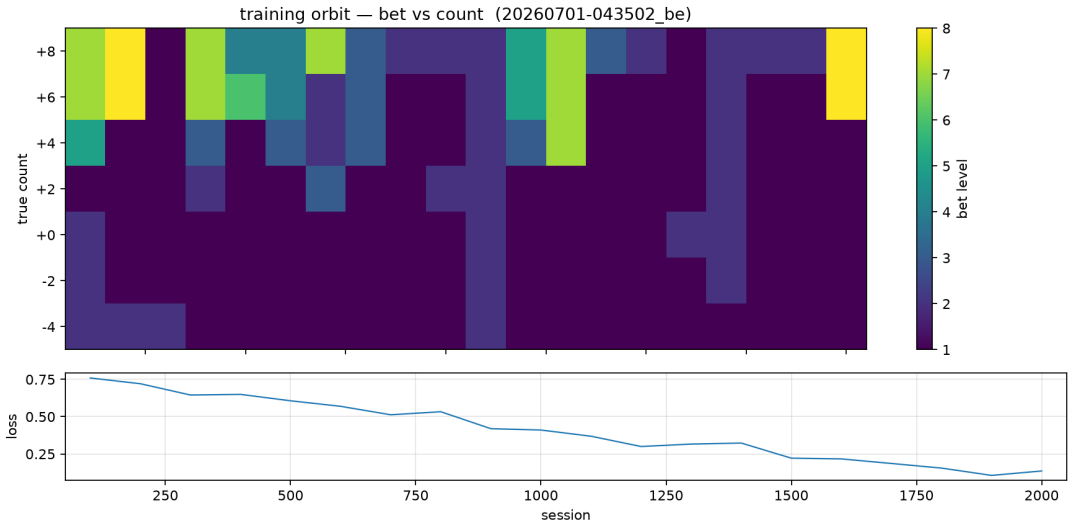


Growth rate by policy and regime — the agent bar (mean over 6 seeds, dots) sits at Flat's height in growth and below it in ruin; the Kelly rung stays out of reach in both. Baselines from the 20k ladder with 95% CIs.

A verdict this negative earns its keep only if the alternatives are ruled out. The next three sections do that in order: the policy's behaviour during training (it is not a near-miss), the signal itself (the wall), and the representation (a hypothesis we falsified ourselves).

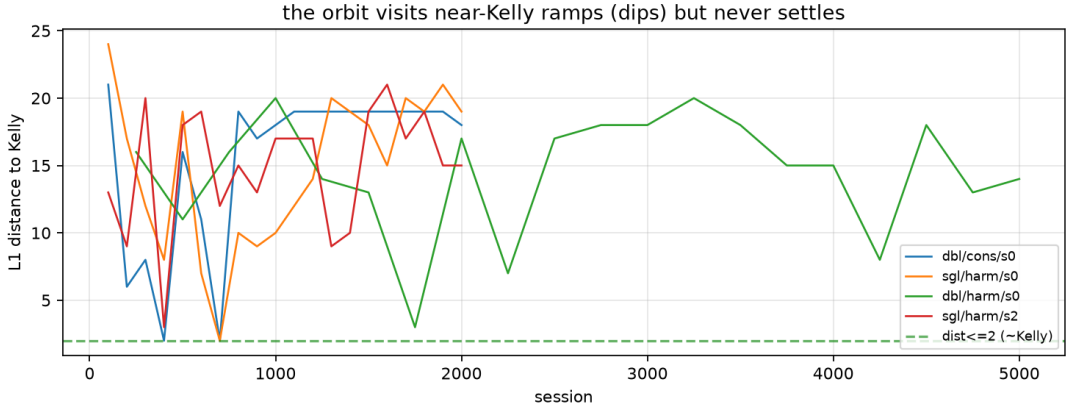
5. The bettor up close — an orbit that visits and leaves

Watching the greedy bet-vs-count curve evolve over training is the most revealing view of what '≈ Flat' actually looks like. The **low counts settle** quickly to the minimum bet and lock. The **high counts never settle** — they flicker between Kelly-like ramps and flat for the entire run. Settling at the low counts is not skill: Kelly also bets the minimum there, so that behaviour is indistinguishable from betting flat everywhere. Only the high counts, where Kelly ramps to 5 and 8 units, can reveal betting skill — and those are exactly the ones that never lock.



The training orbit of a representative ruin-regime run: greedy bet level by true count (heatmap) over training, loss beneath. Low counts lock at the minimum; the high-count rows flicker between ramp and flat for the whole run. Read this as policy INSTABILITY, not near-success: the ramp appears transiently, is not retained, and — Section 6 — does not evaluate well.

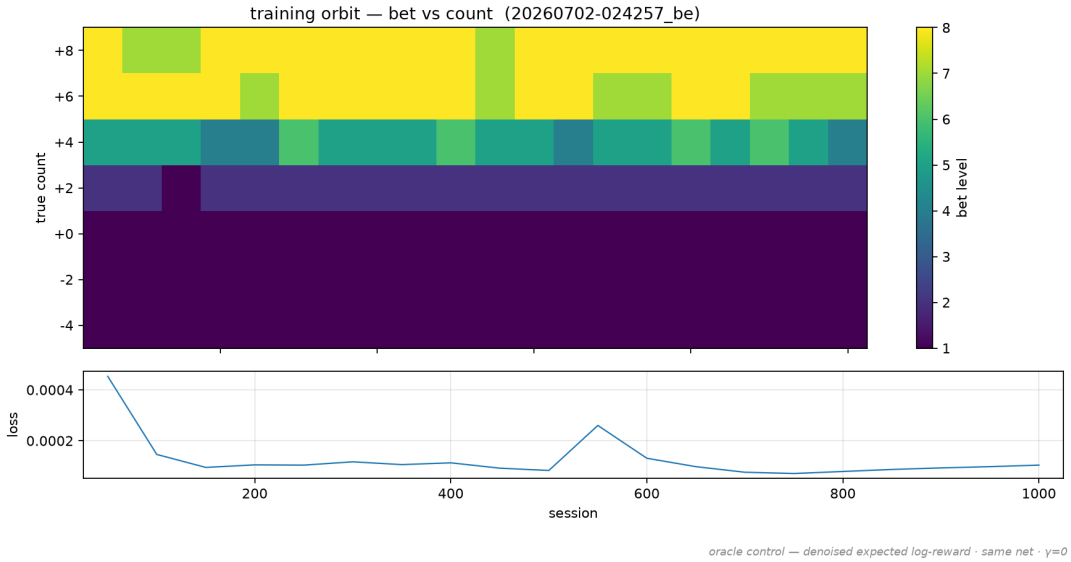
The flicker is not noise around flat — the orbit repeatedly *passes through* genuinely Kelly-shaped curves and leaves them. Tracking the L1 distance between the bet curve and the Kelly ladder over training makes the visits explicit: dips to L1 distance 2–3 (essentially Kelly) that never persist. The network can *express* the target policy; it cannot *hold* it. That observation sharpens the question the next section answers — and it plants a hope the next section kills: perhaps the best mid-training checkpoint, rather than the final policy, is the real prize.



Distance-to-Kelly over training for the four runs that approach it closest — each dip is a checkpoint where the policy was briefly Kelly-shaped. None holds.

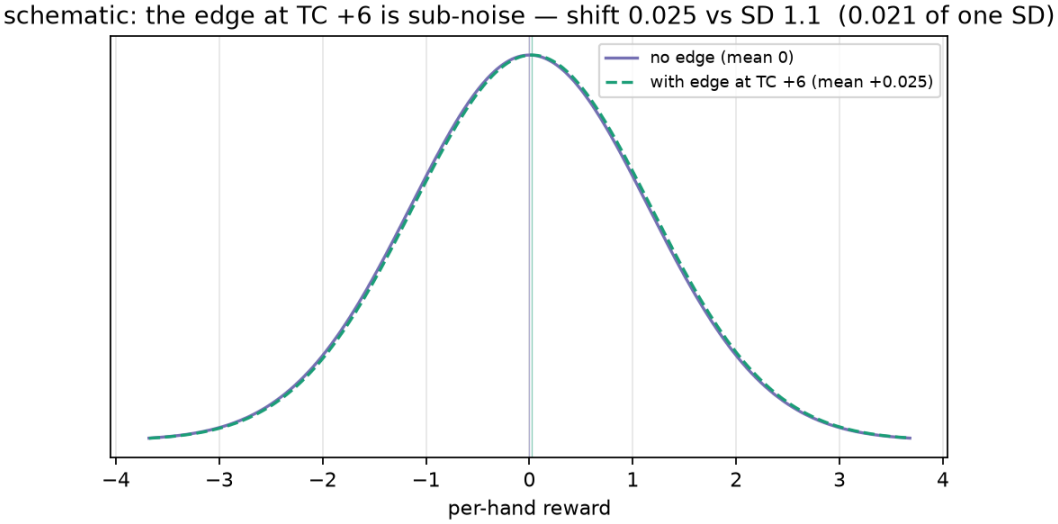
6. Why — the thin edge

Before blaming the signal, rule out the method. The **oracle control** keeps the same network, replay buffer and encode→Q→argmax pipeline at the growth bettor's $\gamma = 0$, and swaps each hand's *realized* log-reward for its *noise-free expectation* from the measured edge curve, rescaled to order one. It runs on plain baseline knobs — batch 128, constant learning rate — none of the noise-averaging machinery the real-reward runs needed. If the flatline were a bug, a capacity limit, or an optimisation failure, the oracle would flatten too. It does not: it locks a clean, stable Kelly ramp and keeps it.



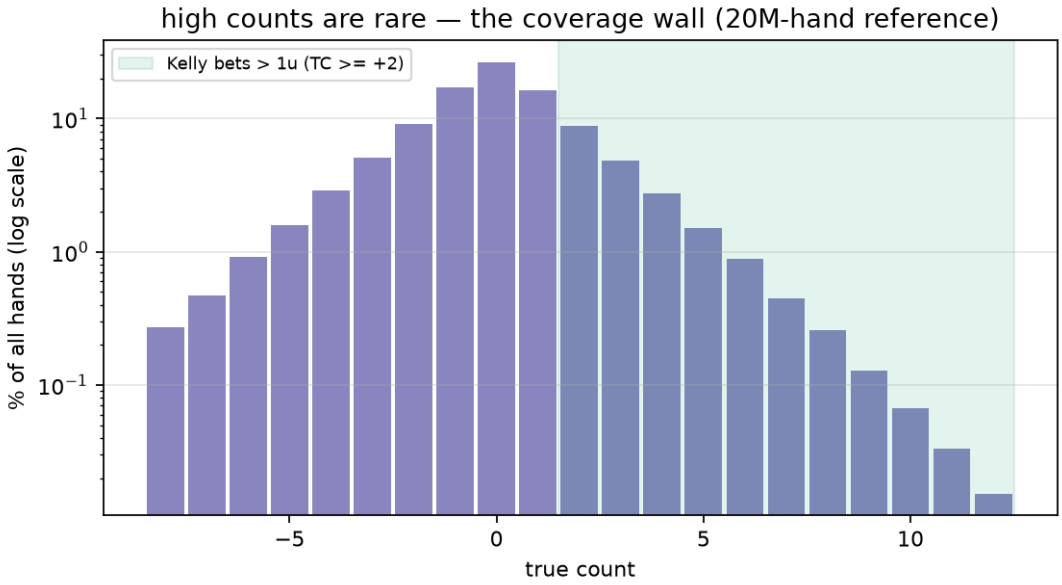
The positive control: same network and pipeline, denoised expected reward. Low counts hold the minimum, high counts ramp to Kelly levels — and stay. The pipeline is sound; what changes on real reward is only the signal.

So the wall is the signal, and it has three compounding parts. **Part one: the prize is tiny.** Everything a learner could win is the Kelly-over-Flat gap of Section 2 — ~0.10×10^{-4} log-growth per hand. **Part two: it is sub-noise.** The bettor must estimate the edge *through* the per-hand payoff, and a single hand's reward has standard deviation ~1.15 — even the strongest max-bet edge (+2.5% at TC +6) is about **2% of one standard deviation**. Resolving a value difference that small takes enormous samples per state.



Schematic, to scale: the per-hand reward distribution at the richest max-bet count against the no-edge baseline. The two curves are visually one — the signal a value learner must resolve is buried in per-hand noise.

Part three: the states that matter are rare. The shoe spends most of its life near neutral; TC +6, where Kelly first bets the maximum, is 0.9% of hands, and TC +8 is 0.26%. The high counts are simultaneously where the edge lives, where the required sample size is largest, and where samples are scarcest. This is the betting-side reprise of the tabular audit's finding two movements ago — errors concentrate where natural play starves the learner of data.



True-count visit frequency (log scale), from the 20M-hand reference. The shaded region — where Kelly bets above the minimum — is a few percent of all hands.

The signal wall, at a glance

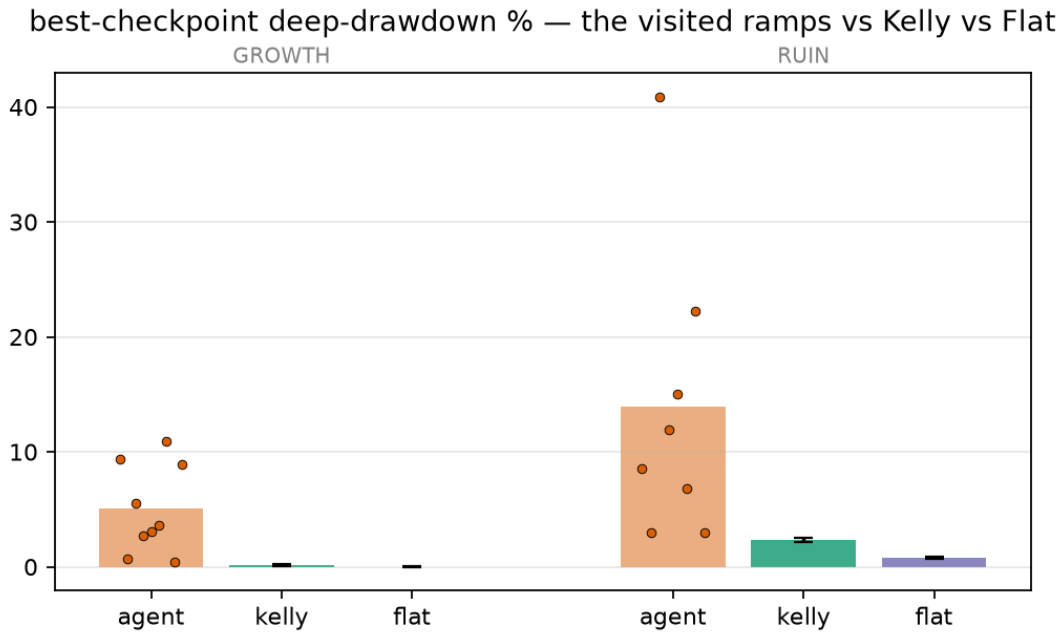
1. The prize is tiny — the entire Kelly-over-Flat gap is ~0.10×10^{-4} log-growth per hand.

2. The noise is huge — per-hand reward SD ~1.15; even the strongest max-bet edge is ~2% of one standard deviation.

3. The carriers are rare — the counts that hold the edge are ~0.9% of hands (TC +6) and 0.26% (TC +8).

Each factor alone is surmountable; their **product** is the wall. The oracle isolates it: remove factor 2 and the same pipeline learns Kelly immediately.

What about the ramps the orbit visits — is the best checkpoint secretly a good policy? **No, and measurably so.** Taking each run's closest-to-Kelly checkpoint (selected by curve shape, never by performance — an honest 'is the ramp real' probe, not cherry-picking) and scoring it on the four axes: in the ruin regime the visited ramps breach the half-bankroll mark 14–18% of the time versus Flat's 0.8%, no better on growth; in the growth regime they run ~5% drawdown versus zero and are decisively worse on growth (-0.52 ± 0.23 vs -0.150, $z \approx 5$). They look like Kelly at a single-count probe and over-bet in the full policy — the Section-3 trap, caught red-handed. The wandering itself was also chased to mechanism: an early 'limit cycle' hypothesis (the agent forgets punishments once it stops over-betting) was ruled out by construction at $\gamma = 0$, leaving static argmax instability on a near-flat Q-surface — harmonic learning-rate decay collapses the orbit, and it collapses to *flat*, not to the ramp. Flat is the loss-attractor; the ramps are the part that averages away.

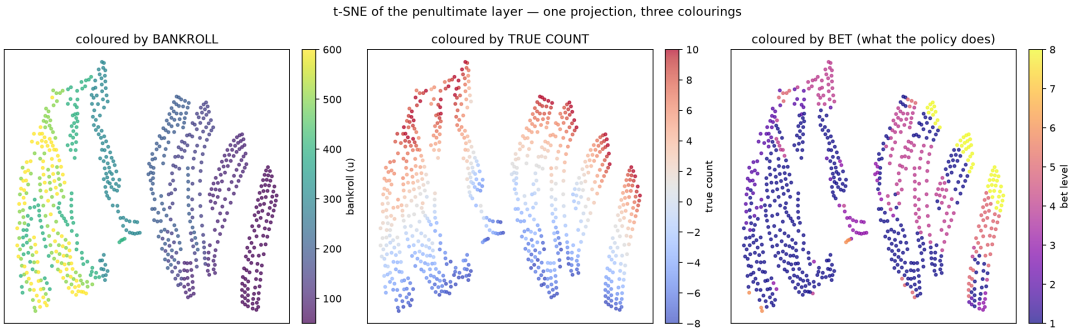


The visited ramps, measured: best-checkpoint deep-drawdown vs the baselines. The Kelly-shaped excursions are over-betting policies the training objective correctly declines to keep.

Put together, the evidence points to the signal as the wall — and the correct reading of '≈ Flat' stops being 'the agent failed to find the answer' and becomes **flat is the loss-optimal answer for the signal it could actually resolve**. One escape hatch remains: maybe the signal is fine and the *representation* is the problem. That hypothesis got the full treatment.

7. A hypothesis of our own, tested and falsified

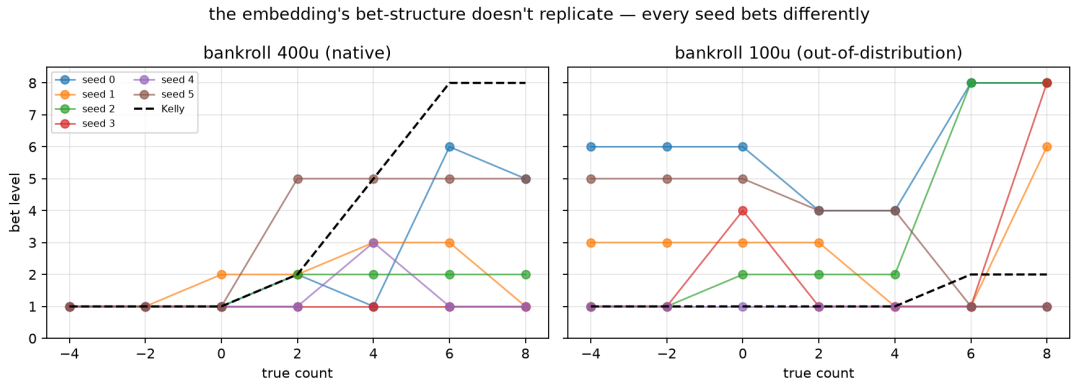
It started with a picture. Projecting the trained network's penultimate-layer activations over a grid of (count, depth, bankroll) states, three colourings tell three tempting stories: **bankroll separates the clusters** ('it keyed on wealth'), **count is a clean gradient within them** ('no, it encoded the count'), and the **greedy bet tracks the wealth clusters** — which seems to settle it for wealth. The growth and ruin networks even mirror: one bets big on its low-bankroll side, the other on its high-bankroll side. The hypothesis practically writes itself: *the agent learned the ×bankroll half of Kelly but not the count gate — the wall is representational, not fundamental.*



The picture that launched the hypothesis: one growth network's embedding, coloured by bankroll, count, and greedy bet. Clusters split by wealth; the bet appears to track them.

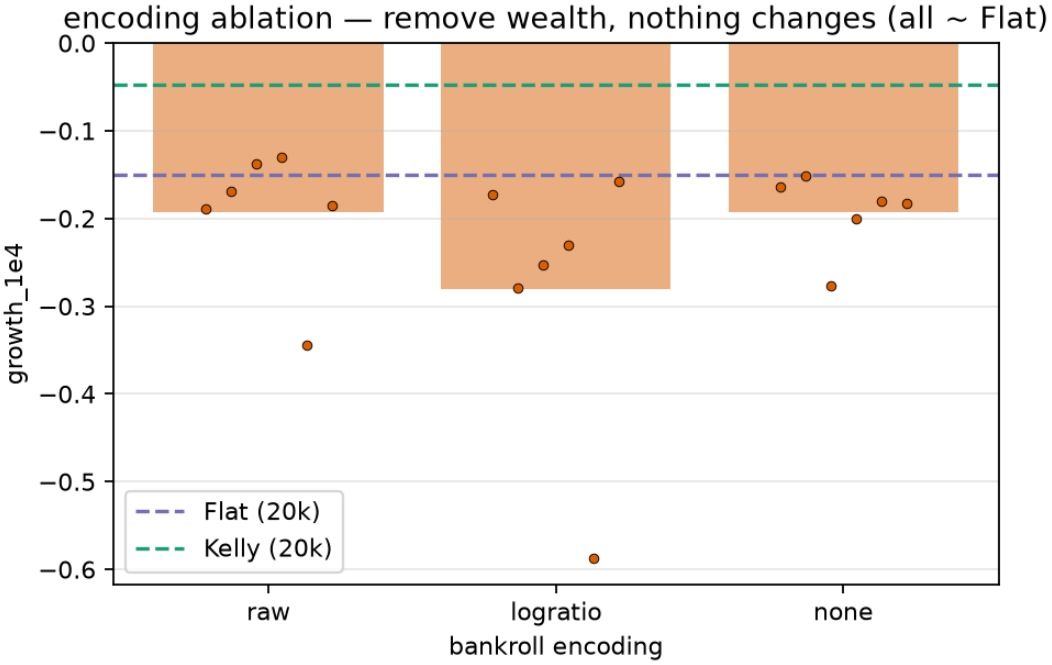
Two disciplines dissolved the picture before any conclusion was drawn from it. **Out-of-distribution:** the probe grid sweeps bankrolls 50–600u, but each agent lives in a narrow band around its start — most of

that wealth axis is states the policy never occupies, so the clean separation is the network extrapolating where no gradient ever corrected it. **Replication:** probing every seed's bet-vs-count curve at the native bankroll, the seeds disagree completely — dead-flat, coarse gating, erratic — and none is the Kelly ramp. The elegant structure was one seed's idiosyncrasy. An embedding is suggestive, not decisive.



The replication check that killed the first read: every growth seed's greedy bet curve at native (left) and out-of-distribution (right) bankroll, discrete Kelly dashed. No common structure survives across seeds.

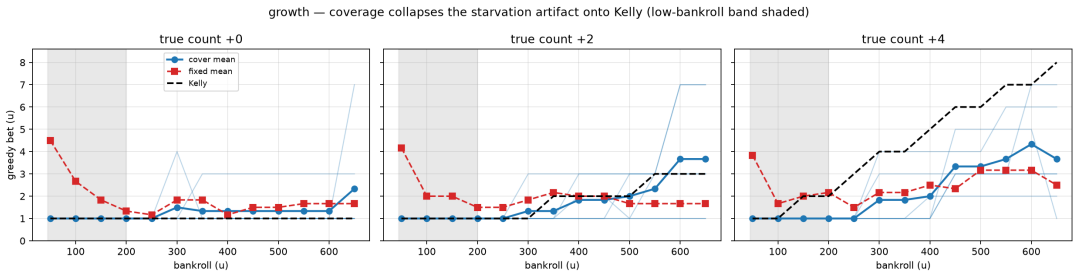
So the hypothesis was promoted from picture-reading to **prediction**, and tested twice, from opposite sides. **Test one — remove wealth.** The bet encoder was made configurable: bankroll fed raw, as a scale-free log-ratio, or **removed from the input entirely**. If the agent keyed on wealth, blinding it to wealth must change its behaviour. It changes nothing: **none** (-0.19) is statistically identical to **raw** (-0.19), log-ratio no better (-0.28); all three sit at the Flat floor.



The encoding ablation: growth across seeds for raw / log-ratio / no-bankroll encoders, Kelly and Flat dashed. Removing wealth from the input entirely moves nothing.

Test two — fill wealth in. If the embedding's mirror is starvation — each regime over-betting exactly the wealth band it never trains in — then training *across* the wealth axis (each session's starting bankroll cycling 100–600u) should collapse the artifact; and if that out-of-band over-betting was costing growth, erasing it should pay. The first prediction **confirmed**: the coverage-trained agents play 26% of hands in the formerly-starved band (vs 0.05%) and the over-betting humps collapse onto Kelly's floor — the mirror was untrained-wealth extrapolation, not strategy. The second prediction **falsified**: growth does not move

(p ≈ 0.89 growth regime, 0.90 ruin), nor does any other axis. The artifact lived where native play never goes; erasing it is cosmetic.



Filling in the wealth axis: bet at a fixed count across bankroll, fixed-start vs coverage-trained agents vs Kelly. The starvation hump collapses; the four-axis outcome does not move.

A footnote the method earned: the first coverage wave (3 seeds) showed a significant ruin-regime gain (p ≈ 0.005). It did not survive three more seeds — a small-sample fluke of exactly the kind this project's protocol exists to catch, recorded rather than erased.

Two independent experiments, opposite directions, one verdict: **nothing wealth-shaped is load-bearing.** Remove wealth — nothing changes; fill it in — the artifact dissolves and still nothing changes. With that, both representational doors are closed: the oracle of Section 6 closed the *capacity* door (the network can represent and hold the Kelly ladder when the signal is clean), and the two ablations closed the *input-encoding* door. The wall is not the representation; it is the thin, sub-noise, rare-state edge of Section 6. And the arc matters as much as the verdict: a hypothesis formed from an internal representation, an over-read caught by out-of-distribution and replication checks, and a falsification delivered by our own controlled experiments.

8. Synthesis — structure beats end-to-end on a sub-noise signal

The edge exists; Kelly captures it; the end-to-end learner, in these runs and under this reward, did not. The asymmetry is sample budgets. The analytic bettor reads the edge from a curve measured *once, offline*, over 20 million pooled hands — then applies a two-line formula. The DQN must re-estimate that same curve *online*, through per-hand noise fifty times larger than the signal, from the few percent of hands that carry it, while simultaneously using the estimate to act. Same edge, opposite sample economics. Where a signal is thin, noisy and rare, structure is not a convenience — it is the difference between extractable and not.

This closes a through-line the whole project kept meeting. The tabular audit found errors pooling in coverage-starved cells; the DQN study found generalization repairing coverage while distorting boundaries; the betting study finds the extreme case — a signal so thin that no amount of end-to-end estimation within the experiment's budget resolves it, while a derived rule captures it outright. The skill the environment actually prices is *restraint* — not over-betting a mostly-negative game — and the learned agent achieves it only in the trivial sense of converging to the minimum bet everywhere: a mistake avoided, not a skill gained.

A result like this is only as trustworthy as the checks it survived, and the method is the other half of what this project set out to practice: build the analytic baseline first, test the learner against it on identical terms, chase the tempting hypothesis, falsify your own interpretation, and record the false positive instead of hiding it. Four times the first read of the evidence was wrong. Recording these is the point — the conclusion stands *because* they were caught:

first read (tempting)	the check that caught it	corrected finding
double-DQN looked 'safe' (low drawdown)	multi-seed CIs, not one run	seed luck — dd 2.6 ± 3.05%, unstable
the visited near-Kelly ramps are the real policy	best-checkpoint four-axis eval	far riskier than Flat, no better on growth
it keyed on wealth, not count (embedding)	OOD + replication + encoding ablation	seed-specific artifact; falsified twice
wealth-coverage recovers growth (n=3, p=0.005)	three more seeds, four-axis re-test	gain vanished (p≈0.9); small-sample fluke

The honesty trail — four tempting over-reads, and the discipline that caught each.

Four movements end here. A simulator became a measurement instrument; a table exposed where learning starves; a network showed what generalization buys and breaks; and a bettor drew the clean line: **when the optimum is derivable, derive it — and know how to prove that the learner's failure is the signal's fault, not your pipeline's.** The second half of that sentence is the transferable skill.

9. What we scoped and set aside

Each entry below was designed far enough to know what it would cost and predict what it would show — the value here is demonstrating the judgment, not promising the work.

Count-coverage (the missing sibling). Section 7 filled in the *wealth* axis; the exact analog for the *count* axis — oversampling high-count states via engineered shoes or stratified replay — would directly separate rarity from noise. Prediction: the ramp stabilizes but real-play growth does not move, because oversampling reweights gradients without growing the ~0.10×10^{-4} prize. Deferred because engineered shoes must first be validated to preserve the measured conditional edge (count is a linear summary of composition; forcing it can silently change the game) — a study of its own.

Learnable abstention (drop the forced minimum bet). The table-minimum tax dominates the growth numbers, and Section 2's Wonging sidebar already gives the analytic answer: sitting out negative counts flips growth positive. Adding a bet-0 action would test whether a value learner can discover abstention — a larger, cleaner signal than sizing (the tax is ~$0.2–0.45\times10^{-4}$ per hand, several times the sizing prize), which makes it the most promising learning experiment on this list.

Prioritized / stratified replay. The standard remedy for rare-state starvation (Schaul et al.); the cheap version of count-coverage. Same prediction, same caveat — it reweights, it does not add information.

Paired evaluation with common random numbers. Our bettors play independent shoes, so tiny gaps need 20k sessions to resolve. Sharing shoe streams across policies would collapse the variance of the *difference* and resolve the ruin-regime margins cleanly at a fraction of the cost; it requires a small evaluator change and per-session artifacts, and would be the first infrastructure investment of any follow-up.

Multi-step credit (TD(λ)), horizon-relative ruin, factored-vs-monolithic, learning to count. Considered and set aside deliberately: the bet's reward lands the same hand it is placed, so multi-step returns buy little; a 'ruined-for-the-remaining-horizon' metric is more realistic but needs an arbitrary recovery model; joint play-and-bet learning and learning the count statistic itself from raw composition are natural extensions of the question, at respectively higher compute and lower expected yield.

Close

What was tested: whether an end-to-end value learner, given the same information a card counter uses and the exact objective Kelly optimizes, rediscovers the derivable betting rule. What happened: it did not — it converged to the no-skill floor, and its Kelly-shaped excursions measured worse than the floor. Why: the prize is tiny, the noise is fifty times larger, and the states that carry the signal are rare — while the analytic rule reads the same edge from a curve measured once, offline. **The transferable problem is not blackjack betting; it is deciding when to learn a policy end-to-end and when to encode known structure — and knowing how to prove which regime you are in. RL was the experiment; structure was the lesson.**

Appendix A — the Kelly criterion, derived

A.1 Why log-growth is the objective — a chosen risk preference. Let f be the fraction of bankroll wagered and R the per-unit return of a hand (win +1, lose –1, blackjack +1.5, doubles ±2, …). Wealth multiplies: after the hand it is W(1 + fR). Maximizing *expected wealth* E[W(1+fR)] is linear in f — on any positive edge it says bet everything, every hand, and a single loss ends the game; almost-sure ruin with maximal expectation is the classic pathology of the arithmetic mean in a multiplicative process. Over T hands wealth is a *product* of factors, and what controls its long-run behaviour is the mean of the *logs*: log W(T) = log W(0) + sum of log(1 + f·R(t)). Maximizing E[log W] maximizes the compound growth *rate* and, by the law of large numbers, almost surely dominates any other strategy eventually. Choosing it is still a preference — it accepts large swings that a mean-variance bettor would not — and the honest framing is: *given* log-growth as the objective, Kelly is a theorem; the objective itself is a decision. This project declares it as the objective and prices the swings separately on the risk axes.

A.2 The Kelly fraction. The exact object is the growth-maximizing fraction at each count — concave in f, so the optimum is unique:

$$f^*(TC) \;=\; \arg\max_f \;\mathrm{E}[\,\log(1 + f R_{TC})\,]$$

Writing the expectation outcome by outcome (the only place the count enters is the outcome probabilities) and Taylor-expanding the logarithm — below, the mean of R is the *edge* and its variance is ≈ 1.3 for blackjack's payoff mix:

$$g(f) \;=\; \mathrm{E}[\log(1 + fR)] \;\approx\; f\mu \;-\; \tfrac{1}{2} f^2 \sigma^2 \qquad \Longrightarrow \qquad f^* \;=\; \frac{\mu}{\sigma^2}$$

(floored at zero when the edge is negative — don't bet a losing count). A term of order edge-squared beside the variance is dropped along the way (the edge is ~0.01, the variance ~1.3) — the small-bet approximation, and the only approximation involved; it is why the Kelly fraction has the memorable *edge over variance* form. The moments have no closed form (the game tree is too rich), so they are Monte-Carlo estimated — the 20M-hand curve of Section 1. From it: f* ≈ 0.5% of bankroll at TC +2, 1.2% at +4, 2.0% at +6 — full Kelly on this game is *small*, under 2% of bankroll across the realistic range. At a 400-unit bankroll that is 2, 5, and 8 units; snapped to the 1–8 spread this is the discrete ladder **1 1 1 2 5 8 8** used throughout, and the measured cost of the snapping is nil (discrete Kelly ≈ continuous Kelly on every axis). The spread top itself is derived the same way — top ≈ f*(+6) × bankroll ≈ 8 units at 400u — and an arithmetic ladder matches the near-linear ramp of f* in the productive band, where a geometric (1,2,4,8) menu would over-resolve the bottom and starve the 5–7u bets Kelly actually wants.

A.3 Fractional Kelly — why backing off is cheap. Bet a fraction c of full Kelly, for c between 0 and 1. Substituting c·f* into g(f), with g_{max} the full-Kelly growth rate:

$$g(c) \;=\; g_{\max}\,(2c - c^2) \;=\; g_{\max}\,\big(1 - (1-c)^2\big)$$

Half Kelly keeps 75% of the growth at half the volatility: the growth penalty near the peak is second-order (the top of a concave curve is flat — g'(f*) = 0) while the volatility cut is first-order (volatility scales linearly with f). This asymmetry is why practical bettors play fractional Kelly, and why over-betting is so much worse than under-betting: past f* the same second-order flatness turns against you, and by 2f* growth is back to zero with volatility doubled. Section 2's ladder shows the live version — flat-8 is far beyond f* at most counts, and it ruins.

A.4 The ruin-aware bend. Without a ruin barrier, maximizing E[log W(T)] is *myopic*: full Kelly every hand, horizon irrelevant — finite time alone does not change the bet. A hard barrier (bankroll below the table minimum ends the session permanently) breaks the myopia: hitting it forfeits every future positive-edge hand, so capital near the barrier carries option value beyond the current hand's log term. The optimum becomes a function b*(W, TC, hands left) that equals Kelly when W is far from the barrier and bends *below* Kelly as W approaches it — under-bet to survive. This is why the project runs two regimes: at 400u the barrier is out of reach and pure Kelly is the target; at 200u restraint acquires survival value, which is exactly the regime where the learned agent's over-betting excursions get priced (Section 6).

A.5 What is exact and what is estimated. The definition of f* is exact; the edge-over-variance form is the small-bet approximation (fine here — the dropped term is four orders below the kept one); the moments are Monte-Carlo estimates whose standard errors shrank as $1/\sqrt{n}$ to two significant figures over the 20M-hand run. The reference the agent is graded against is therefore itself a measured object with quantified uncertainty — matching the report's standing rule that a number without provenance and error bars is not yet a result.