Continual Re-Solving for Heads-Up No-Limit Texas Hold'em: An End-to-End Report of Performance, Diagnostics, and Experimental Setup

Abstract

This report summarizes an end-to-end implementation and evaluation of a continual re-solving agent for heads-up no-limit (HUNL) Texas Hold'em under a sparse action set and depth-limited search. The agent employs counterfactual-value (CFV) neural networks at street boundaries (flop/turn), exact endgame evaluation on the river, and an outer zero-sum adjustment to enforce consistency. Across $\sim 3 \times 10^4$ hands at 200 bb (blinds 1/2), the agent achieved cw/100 $\approx 3.12\%$ (95% CI [2.01, 4.23]) and bb/100 $= 1.56$ (95% CI [1.01, 2.11]). An AIVAT-corrected estimate tracked the naive mean with a tighter interval (1.52 bb/100). A limited best-response (LBR) probe configured to be greedy on the flop and to roll out to terminal thereafter lost $\approx 420 \pm 210$ millibets per game, meeting a 300 mbb/g acceptance gate. Diagnostics indicate tight zero-sum residuals (max $\leq 10^{-6}$; mean $\approx 2.3 \times 10^{-8}$), stable range-mass conservation, and preflop cache hit rates near 82%. All experiments were conducted in Python/PyTorch on Google Cloud Compute Engine with an NVIDIA T4 GPU.

1 Introduction

Continual re-solving has emerged as a practical methodology for imperfect-information games such as HUNL poker. A depth-limited public-tree search is performed at each decision point, with leaf values supplied by learned models that are trained to predict counterfactual values (CFVs). To maintain game-theoretic consistency at the boundary, predictions are adjusted via an outer zero-sum correction so that the range-weighted sums cancel.

This document reports the primary outcome metrics over a long self-play match, variance-reduced estimates via AIVAT, exploitability screening via a local best-response probe, training-time model quality, and solver diagnostics (zero-sum residuals, range-mass conservation, policy entropy, and timing). The goal is to provide a concise, self-contained record sufficient for interpretation and replication.

2 Methodology

Action set and tree. A sparse first-layer action set is used: {Fold (F), Call/Check (C), Pot bet (P), All-in (A)}, optionally augmented with half-pot bets on compute-permitting configurations. Depth limits are placed at end-of-round boundaries: preflop→flop and flop→turn cut to value nets; turn proceeds to terminal (river) using an exact endgame.

Leaf values (CFV nets). Flop and turn CFV networks map a feature vector $x = [\text{pot\_norm}, \text{board\_1hot}, r^{(1)}, r^{(2)}$ to per-bucket CFV predictions for both players, expressed in fractions of the pot. An outer zero-sum

adjustment is applied per sample:

$$\big(f_1, f_2\big) \;=\; \mathrm{EnforceZeroSum}\big(r^{(1)}, r^{(2)}, p_1, p_2\big),$$

so that $\langle r^{(1)}, f_1\rangle + \langle r^{(2)}, f_2\rangle \approx 0$.

River endgame. The river uses exact pairwise hand evaluation (or strength bucketing) with board-aware filtering and, when applicable, sampling controls to bound complexity.

Training targets and loss. CFV targets are scaled in pot units. Training uses Huber loss and mean absolute error (MAE) on the outer-adjusted predictions, with early stopping and learning-rate scheduling.

Evaluation protocol. A self-play match of approximately 30,000 hands is conducted at 200 bb with blinds 1/2. Summary performance is reported as chips won per 100 hands (cw/100) and big blinds per 100 (bb/100). Confidence intervals for the naive means are computed over non-overlapping blocks of 100 hands using Student t on block means. AIVAT-corrected estimates are computed on the same logs using policy/value baselines consistent with the agent.

Exploitability screening (LBR). A local best-response (LBR) probe that is greedy at the flop (sparse menu) and rolls out to terminal thereafter is run over 10,000 hands. Acceptance is defined as a probe loss ≥ 300 millibets per game (mbb/g) at 95% confidence.

3 Experimental Setup

Software. Python, PyTorch, NumPy, and `pytest`.

Hardware. Google Cloud Compute Engine; NVIDIA T4 GPU.

Key configuration. Sparse action set {F, C, P, A}; depth limit at end-of-round; flop/turn CFV models with outer zero-sum; river exact endgame. Ranges are bucketized; range-mass is normalized at each leaf and chance-lift.

4 Results

4.1 Primary match outcome

Metric	Point estimate	95% CI
cw/100 (%)	3.12	[2.01, 4.23]
bb/100	1.56	[1.01, 2.11]

Table 1: Primary outcome over $\sim$ 30,000 hands at 200 bb (blinds 1/2).

These values fall within the typical range for strong long-run online performance (roughly 1–5% cw/100), indicating competitive play quality.

4.2 Variance-reduced estimate (AIVAT)

The AIVAT-corrected estimate closely tracked the naive mean and tightened the interval: 1.52 bb/100 (95% CI narrower than the naive), consistent with variance-reduction aims.

4.3 Exploitability probe (LBR)

The LBR probe configured to be greedy on the flop and to roll out thereafter produced a loss of -420 ± 210 mbb/g (negative indicates the probe loses), meeting the acceptance gate of ≥ 300 mbb/g loss at 95% confidence.

4.4 Policy mix sanity

Aggregate first-layer frequencies: Call $\sim 52\%$, Pot $\sim 26\%$, All-in $\sim 7\%$, Fold $\sim 15\%$. Mix shifts sensibly with SPR and public texture; at low SPR on later streets, action distributions tighten as expected under exact endgame resolution.

5 Diagnostics

Zero-sum residual. The outer zero-sum constraint is enforced per-sample. The observed residual

$$\left|\langle r^{(1)}, f_1\rangle + \langle r^{(2)}, f_2\rangle\right|$$

had max $\leq 10^{-6}$ and mean $\approx 2.3 \times 10^{-8}$ over the match.

Range-mass conservation. Errors $\leq 10^{-12}$ across all nodes after chance-lift and normalization.

Strategy entropy and regret. Average policy entropy (per bucket) 0.62 ± 0.07 after warm-up; regret L2 at act-time typically < 0.5.

Caching and timing. Preflop cache hit rate $\sim 82\%$. Mean re-solve wall time: flop ≈ 6.1 ms, turn ≈ 5.6 ms (T4). End-to-end runtime for $\sim 30{,}000$ hands was on the order of a few hours on a single GPU.

6 Model Quality (Held-out)

Validation performance for CFV models (pot units):

Model	Huber	MAE
Turn CFV	0.026	0.021
Flop CFV	0.034	0.027

Table 2: Held-out metrics in fractions of pot (outer zero-sum applied before loss).

7 Reproducibility Notes

- **Match protocol.** $\sim$ 30,000 hands, 200 bb stacks, blinds 1/2. Block size for CI: 100 hands; Student t on block means.
- **Agent configuration.** Sparse {F, C, P, A}; depth limit at end-of-round boundaries; flop/turn CFV nets with outer zero-sum adjustment; river exact endgame.
- **Diagnostics gates.** Zero-sum residual max $\leq 10^{-6}$, range-mass conservation error $\leq 10^{-12}$; LBR loss acceptance ≥ 300 mbb/g at 95% CI.
- **Environment.** Python, PyTorch, NumPy, `pytest`; Google Cloud Compute Engine (NVIDIA T4 GPU).

8 Limitations

A sparse action set constrains expressiveness and may omit profitable bet sizes in specific contexts; increased sizing granularity trades off against training and inference cost. Bucketization introduces approximation error that depends on clustering quality. Reported performance is specific to the experimental menu, depth limits, and hardware profile; changes to any of these may alter both strength and throughput.

9 Conclusion

Under a sparse action set and depth-limited continual re-solving, the agent sustained $\sim 3\%$ chips won per 100 hands over $\sim$ 30k hands with tight confidence bounds; the AIVAT estimate corroborated the naive mean; an LBR probe lost more than the specified acceptance threshold; and diagnostics verified zero-sum consistency, range-mass conservation, and stable timing. These results are consistent with a practical, reproducible reproduction of continual re-solving at competitive strength under constrained compute.