NeurIPS
2026
POSTER

# Stabilizing RL+Search for Imperfect-Information Games

**Yinghao Li**
China Agricultural University

**Stochastic Horizon Annealing** is one line of code that narrows the across-seed spread of final exploitability by 34% to 61%. It draws the CFR budget of every depth-limited subgame from a truncated Gaussian centred at the nominal budget, and leaves the rest of the ReBeL loop unchanged.

**34-61%**
narrower across-seed exploitability interval

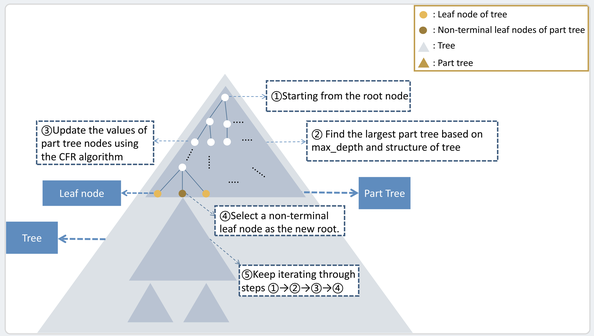
**0.75 → 0.37**
cross-depth error correlation

**≤1.01×**
convergence-rate factor

**1 line**
change to a ReBeL trainer

## 1 Depth-limited solving

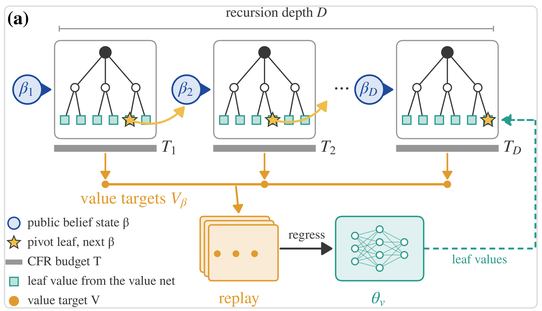
No player sees the whole state, and the game tree grows exponentially with depth. ReBeL searches a few moves ahead, estimates the deeper nodes with a value network, and lifts the game to **public belief states**, the common-knowledge distribution over private information.




## 2 The last fixed constant

Every subgame in ReBeL is solved for the same **N = 1024** linear-CFR iterations.

- One solver maturity feeds every target, so the network cannot separate a nearly converged solve from one that stopped early.
- Each level's error is the input to the level above, so the recursion compounds it.




The solved root value is a regression target for the value network, and the pivot leaf seeds the next public belief state.

ReBeL already randomizes which iteration supplies the next belief state. The budget is the last constant shared by all subgames.

The fixed budget is the **dominant remaining source** of seed-dependent noise.

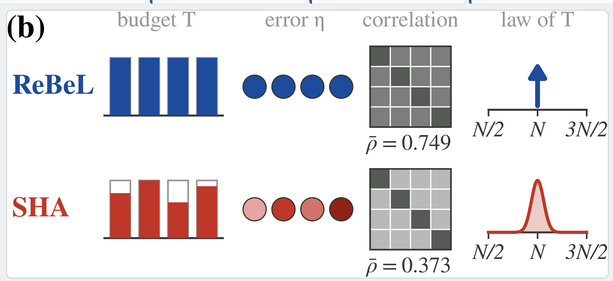
## 3 SHA: randomize one line ★ KEY

Each subgame draws its own budget; the pivot stays uniform, exactly as in ReBeL.

**BUDGET DRAW**

$$T \sim \mathcal{N}(N, \sigma^2) \quad \text{truncated to} \quad [N/2,\, 3N/2], \qquad p \sim \text{Uniform}\{1, \ldots, \lfloor T \rfloor\}$$

**T** drawn budget | **N** nominal budget | **σ** annealing strength | **p** pivot iteration

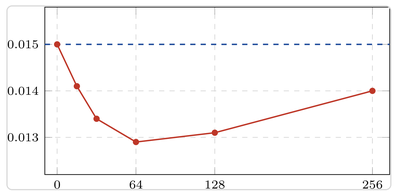



A fixed budget solves every subgame to the same CFR maturity, so the errors share a component; SHA draws each budget and lowers the correlation.

For integer $N$, $\sigma > 0$, and these symmetric bounds, $\mathbb{E}[\lfloor T \rfloor] = N - \frac{1}{2}$, so SHA preserves the expected number of iterations up to half an iteration.

**Implicit ensemble.** Each target averages over a band of solver maturities instead of a single one.

## 4 Annealing strength




Exploitability against $\sigma$ on 1×5f; the dashed line is ReBeL.

| Game | σ=0 | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|---|
| 1×4f | 0.0167 | 0.0165 | 0.0160 | **0.0155** | 0.0158 | 0.0167 |
| 1×5f | 0.0150 | 0.0141 | 0.0134 | **0.0129** | 0.0131 | 0.0140 |
| 1×6f | 0.0238 | 0.0224 | 0.0210 | 0.0197 | **0.0195** | 0.0213 |

The optimum is broad: every value between $N/30$ and $N/8$ is near-optimal, and $\sigma = 64$ stays within 5% of the per-game best.

Lookahead depth 2, $N = 1024$ linear-CFR iterations, $\sigma = 64$ with truncation $[512, 1536]$ unless noted.

## 5 Three results

**Rate preserved.** For a finite two-player zero-sum subgame with perfect recall and leaf values held fixed during the solve,

$$\mathbb{E}_T\big[\text{expl}(\bar{\pi}^H)\big] \le \frac{K_\beta}{\sqrt{N}}\Big(1 + \tfrac{1}{2}\big(\tfrac{\sigma}{N}\big)^2\Big)\Big(1 + \tfrac{2}{N}\Big),$$

at most **1.01×** the fixed-budget bound when $\sigma \le N/8$ and $N \ge 1024$ (Thm. 5.3).

**Small per-target cost.** Under a regular budget response,

$$\text{Var}_T[V_\beta(H)] \le \frac{L_\beta^2(\sigma+1)^2}{N^3},$$

about $4 \times 10^{-3}\,(L_\beta/K_\beta)^2$ of the squared error envelope $K_\beta^2/N$ of a single solve at $N = 1024$, $\sigma = 64$ (Thm. 5.4).

## 6 Gain grows with depth

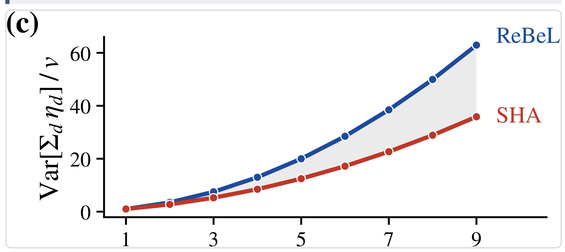
Under the cross-depth correlation model, the local errors $\eta_d$ along the chain have common variance $v$ and pairwise correlation $\rho$.

**RECURSIVE VARIANCE**

$$\text{Var}\Big[\sum_{d=1}^{D} \eta_d\Big] = D\,v\,\big(1 + (D-1)\rho\big)$$

For fixed model parameters with $\rho_s < \rho_b$ (s: SHA, b: ReBeL), the variance ratio falls monotonically with the recursion depth $D$ (Cor. 5.6):

$$r(D) = \frac{v_s}{v_b} \cdot \frac{1 + (D-1)\rho_s}{1 + (D-1)\rho_b}$$




Variance of the summed error, $Dv(1 + (D-1)\rho)$, at the two measured correlations.

## 7 Liar's Dice ★ Headline

Final exploitability at epoch 100, mean ± $CI_{95}$; 30 seeds (Deep CFR 3). Lower is better.

| Game | ReBeL | SHA | Deep CFR |
|---|---|---|---|
| 1×4f | 0.0167 ± 0.0014 | **0.0155 ± 0.0009** | 0.101 ± 0.037 |
| 1×5f | 0.0150 ± 0.0013 | **0.0129 ± 0.0006** | 0.148 ± 0.006 |
| 1×6f | 0.0238 ± 0.0020 | **0.0197 ± 0.0008** | 0.253 ± 0.034 |

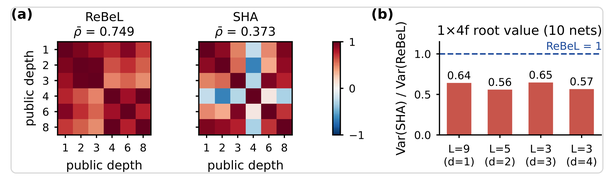
**34%** interval 1×4f | **52%** interval 1×5f | **61%** interval 1×6f

The mean also ends lower, by 7%, 14% and 17%. Both effects grow with the size of the variant. Larger variants have longer recursion chains, where Cor. 5.6 predicts the variance ratio to fall. At $N = 1024$ the variance SHA injects at a single subgame is tiny, and the gain comes from the recursion chain. Deep CFR, a search-free reference, trails the recursive methods by roughly an order of magnitude.

## 8 Deeper Leduc domains

Leduc-30 and Leduc-40 add a different information structure and recursion depths of 3.2 to 3.4, testing Cor. 5.6 beyond the Liar's Dice range. Exact exploitability at the matched budget, mean ± $CI_{95}$, $n = 20$ runs per arm.

| Game | ReBeL | SHA |
|---|---|---|
| Leduc-30 | **0.228 ± 0.036** | 0.242 ± 0.021 |
| Leduc-40 | 0.563 ± 0.100 | **0.432 ± 0.047** |




**(a)** Across-run correlation of value errors by public depth, on Leduc-40, whose tree supports chains of up to nine subgames: mean off-diagonal 0.749 (ReBeL) vs 0.373 (SHA). **(b)** Across-run variance of the root value in nested solves on 1×4f, SHA relative to ReBeL.

**40%** across-run std Leduc-30 | **53%** across-run std Leduc-40 | **23%** mean Leduc-40 | **0.56-0.65×** root-value variance, 1×4f

The variance-coupling ratio falls from 3.17 under ReBeL to 2.50 under SHA. The residual correlation under SHA is the shared-weight component. Variance reduction is the consistent effect; the mean gain appears only in the larger game. Both variance reductions are significant under a two-sided F-test.

**Takeaways** | **IDEA.** The fixed CFR budget is a variance channel in the recursion. | **METHOD.** Draw it from a truncated Gaussian; one line, same expected cost. | **RESULT.** 34-61% narrower across-seed interval; the gain grows with game size. | **PRACTICAL.** Drops into a ReBeL-style trainer.


Contact: **lyh010428@alu.cau.edu.cn**