WHICH NASH EQUILIBRIUM? SOLVER-DEPENDENT SELECTION ON ZERO-SUM NASH POLYTOPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many two-player zero-sum games admit not a unique Nash equilibrium but a convex *polytope* of them: profiles that share the minimax value $V^\star$ yet prescribe materially different behavior. Standard solvers are guaranteed to find *some* equilibrium and are treated as interchangeable; we ask whether they instead systematically *select* different members. Using a tabular, exactly solvable testbed of six games with analytically known Nash sets—including a two-dimensional Nash polytope and Kuhn poker—and a randomized ensemble of 180 asymmetric safe-action matrix games, we find that the selected member is a deterministic property of the update rule, and that, on matrix games, solver families separate only on geometrically *asymmetric* Nash sets. Regularized Nash Dynamics (R-NaD), started from a uniform reference, selects the *maximum-entropy* equilibrium—the information projection of its reference onto the Nash set—to three decimals on the 2-D polytope and in 100% of converged ensemble games, while CFR$^+$ lands on lower-entropy members (94% of games; paired Wilcoxon $p < 10^{-27}$), and additional CFR$^+$ iterations widen the gap rather than closing it. A fixed-reference baseline (magnetic mirror descent with a small magnet) does not converge on these games, so the evidence concerns moving-reference regularization specifically. Two further tests qualify natural readings of these results: replacing regret matching's $\max(R, 0)$ projection with a softmax does not remove the drift of regret-averaging methods, and R-NaD's selection is *anchor-following*—it moves with the initial reference—rather than initialization-independent. Kuhn poker marks the limits of the picture. There, selection has real off-path consequences—against mis-calling opponents the value of an equilibrium grows linearly with its bluffing frequency, by up to 30% of $|V^\star|$ between returned members—but the characterization does not carry over: R-NaD converges exactly to a member near, not at, maximum entropy, which shifts with the regularization temperature, and the member returned by CFR depends on how counterfactual values are normalized. We therefore state the information-projection characterization as a conjecture for matrix games only, and report its naive extensive-form analogue as refuted.

1 INTRODUCTION

A central simplification in computational game theory is the treatment of a Nash equilibrium as a unique target: one solves a game, obtains *the* equilibrium, and deploys it. For two-player zero-sum (2p0s) games this is justified at the level of *value*—the minimax theorem guarantees a unique game value $V^\star$—but it is frequently false at the level of *behavior*. Degenerate games possess an entire convex set of Nash equilibria; every member achieves $V^\star$ against a best-responding opponent, yet members can differ in the support they use and in their value against imperfect opponents.

When the equilibrium set is not a singleton, a solver does not merely *find* an equilibrium; it *selects* one. The solvers behind modern game-playing systems span two broad families with very different update structure: *regret-averaging* methods such as counterfactual regret minimization (CFR) and CFR$^+$ (Zinkevich et al., 2007; Tammelin, 2014), which power superhuman poker agents (Bowling et al., 2015; Moravčík et al., 2017; Brown & Sandholm, 2019) and deploy a time-averaged strategy; and *regularized last-iterate* methods such as Regularized Nash Dynamics (R-NaD), the engine behind DeepNash (Pérolat et al., 2021; 2022), and Magnetic Mirror Descent (MMD) (Sokota et al., 2023), which deploy the final iterate of a dynamics anchored to a reference policy. Because a learned agent's

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

behavior against real, imperfect opponents is determined by *which* equilibrium it carries, the choice is of practical as well as conceptual interest, yet the two families are routinely used interchangeably. We ask:

> *When a 2p0s game has a set of Nash equilibria, do different solvers select different members of that set, and is the choice a systematic property of the algorithm?*

We study this in a deliberately minimal, fully transparent setting: a tabular extensive-form engine using exact counterfactual values (no sampling or function-approximation variance), applied to games whose Nash sets we know analytically, so that every solver output is located *on the known Nash set* and compared against ground truth rather than against another solver's output.

Contributions.

1. **Selection is algorithmic and geometry-gated** (§4.1, §4.4). On matrix games, across-algorithm differences are large on *asymmetric* Nash sets and vanish on symmetric ones. The separation is robust to the iteration budget: over two orders of magnitude R-NaD's selection does not move, while CFR$^+$'s moves *farther* from the maximum-entropy member.
2. **A maximum-entropy / information-projection characterization of R-NaD** (§4.1–4.5). With a uniform reference, R-NaD selects the maximum-entropy equilibrium on every matrix game tested—including the interior of a 2-D polytope and 100% of the converged games of a randomized 180-game ensemble (median coordinate error 2×10^{-4})—while CFR$^+$ selects a lower-entropy member (mean entropy gap $+0.121$, Wilcoxon $p < 10^{-27}$). The selection is *anchor-following* (§4.5): it moves with the initial reference, so maximum entropy is the uniform-reference special case. We state the characterization as a conjecture for matrix games (§5).
3. **A test of a natural mechanism for regret-averaging drift** (§4.6). A natural hypothesis is that the drift of CFR-family averages toward low-entropy members is caused by the $\max(R, 0)$ projection of regret matching. Replacing the projection with a softmax over the same regrets does not remove the drift; at the step size tested it increases it.
4. **The limits of the picture in an extensive-form game** (§4.2, §4.7, Appendix F). In Kuhn poker R-NaD converges to machine precision at 99.7% of the maximum entropy but at a different coordinate (0.180 vs. 0.201) that is budget-independent and shifts with the regularization temperature, so the naive extensive-form analogue of the conjecture fails; and the member returned by CFR/CFR$^+$ depends on whether counterfactual values are reach-normalized. Kuhn is also where selection has consequences: on our matrix games every equilibrium is an equalizer, so members are payoff-equivalent against *every* opponent, whereas in Kuhn the value against mis-calling opponents grows linearly with the selected bluffing frequency.

Section 2 defines the objects we measure, §3 the testbed, §4 the results, and §5 separates what is established from what is conjectured; supporting tables, figures and implementation details are in the appendix.

2 PRELIMINARIES

Games and equilibria. We consider finite two-player zero-sum games in extensive form with perfect recall; normal-form (matrix) games are the single-decision special case. Player i has a behavioral strategy σ_i assigning a distribution over actions to each of its information sets, and $\sigma = (\sigma_0, \sigma_1)$ denotes a profile with $u(\sigma) = u_0(\sigma) = -u_1(\sigma)$. We write $\mathcal{N}(G)$ for the set of Nash equilibria of G. By the minimax theorem all $\sigma \in \mathcal{N}(G)$ share the value $u(\sigma) = V^\star$, and $\mathcal{N}(G)$ is convex in sequence-form coordinates; we refer to it as the *Nash polytope*.

Exploitability. Convergence is measured by exploitability,

$$\textsc{NashConv}(\sigma) \;=\; \Big[\max_{\sigma_0'} u(\sigma_0', \sigma_1) - V^\star\Big] \;+\; \Big[V^\star - \min_{\sigma_1'} u(\sigma_0, \sigma_1')\Big], \tag{1}$$

non-negative and zero exactly at Nash; best responses are computed exactly by enumeration.

The maximum-entropy member and the I-projection. To say *which* equilibrium a solver returns we need a reference point on the Nash set. A canonical one is the member that commits to as little as the equilibrium constraints allow (Jaynes, 1957), the maximum-entropy equilibrium

$$\sigma^{\mathrm{ME}} \;=\; \operatorname*{arg\,max}_{\sigma \in \mathcal{N}(G)} \mathcal{H}(\sigma), \qquad \mathcal{H}(\sigma) = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} H\big(\sigma(\cdot \mid I)\big), \tag{2}$$

where $\mathcal{I}$ is the set of information sets and H is Shannon entropy. Informally, σ^{ME} spreads probability as evenly as equilibrium allows; low-entropy members concentrate on a face or vertex of the Nash set. Note that $\mathcal{H}$ is the *unweighted mean behavioral entropy*. In matrix games it is the usual strategy entropy; in extensive-form games it differs from the reach-weighted (dilated) entropies used to regularize sequence-form solvers (Hoda et al., 2010; Kroer et al., 2020), a distinction that matters in §5. Since $H(p) = \log|A| - \mathrm{KL}(p \,\|\, \mathrm{unif})$, maximizing entropy is equivalent to minimizing relative entropy to the uniform policy: σ^{ME} is the *information projection* (I-projection; Csiszár, 1975) of the uniform policy onto $\mathcal{N}(G)$, i.e. the equilibrium closest to uniform in KL divergence. More generally, for a reference policy ρ,

$$\Pi_{\mathcal{N}}(\rho) \;=\; \operatorname*{arg\,min}_{\sigma \in \mathcal{N}(G)} \mathrm{KL}\big(\sigma \,\|\, \rho\big), \tag{3}$$

the equilibrium closest to ρ, with $\Pi_{\mathcal{N}}(\mathrm{unif}) = \sigma^{\mathrm{ME}}$. This object is relevant because R-NaD and MMD regularize toward a reference policy with exactly this KL term, so (3) is the natural candidate for what they select. That entropic regularization selects the maximum-entropy member of an optimal set is well established outside game solving—for entropically penalized linear programs (Weed, 2018) and entropic optimal transport (Cuturi, 2013)—and the maximum-entropy Nash equilibrium has been singled out as canonical before (Balduzzi et al., 2018). Our question is not whether such a selection principle exists, but which of the solvers in standard use realize it.

Solvers. A single tabular engine instantiates all update rules. At each iteration it computes, for every information set I, the exact counterfactual action values divided by the counterfactual reach of I, denoted $q = q_t(I, \cdot)$; let $s = \sigma_t(\cdot \mid I)$.

- **CFR** (Zinkevich et al., 2007): cumulative regret $R \mathrel{+}= q - s^\top q$; current policy $\sigma_{t+1} \propto \max(R, 0)$; deploys the time-average of the current policies. **CFR⁺** (Tammelin, 2014): regret-matching⁺, $R \leftarrow \max(R + (q - s^\top q), 0)$; deploys the linearly ($t$-)weighted average. **Hedge** (ablation only, §4.6): identical regrets but $\sigma_{t+1} \propto \exp(\eta R)$, a softmax with no positive-orthant projection; deploys the average. *Implementation note.* Averages are taken per information set over the behavioral policies, without weighting by the player's own reach, and regrets use the reach-normalized values q. In matrix games (one information set per player, reach one) both coincide exactly with textbook CFR/CFR⁺, which covers five of the six games and the whole ensemble. In Kuhn they define a *reach-normalized variant* that differs from textbook CFR, and the difference matters: Appendix F shows that textbook CFR/CFR⁺ return different Kuhn members than this variant. Kuhn entries for CFR and CFR⁺ are marked accordingly and carry no weight in our conclusions.
- **Fictitious play (FP)** (Brown, 1951; Robinson, 1951): realization-weighted averaging of exact best responses; deploys the average.
- **Magnetic Mirror Descent (MMD)** (Sokota et al., 2023): with reference (magnet) ρ, regularization temperature λ, step size η, and $c = 1/(1 + \eta\lambda)$,
$$\log \sigma_{t+1} \;\propto\; c \,\log s + c\,\eta\, q + (1 - c)\,\log \rho, \tag{4}$$
run here with a *fixed* uniform reference and a small temperature ($\lambda = 0.1$); deploys the final iterate.
- **R-NaD** (Pérolat et al., 2021; 2022): the same update (4) with a *moving* reference, reset to the current policy every 60 iterations ($\lambda = 0.5$). This traces a sequence of quantal-response equilibria (McKelvey & Palfrey, 1995) converging to an unregularized Nash equilibrium, the algorithmic analogue of the logit-QRE homotopy (Turocy, 2005; Gemp et al., 2022). Deploys the final iterate.
- **MWU / NeuRD** (Hennes et al., 2020): the unregularized last iterate $\sigma_{t+1} \propto \exp(\log s + \eta q)$.

The families differ in what is deployed: regret-averaging methods (CFR, CFR⁺, Hedge, FP) deploy a *time-average*; (un)regularized last-iterate methods (MMD, R-NaD, MWU) deploy the *final iterate*. All hyperparameters are listed in Appendix E.

3 EXPERIMENTAL SETUP

Testbed. We use six games with analytically characterized Nash sets (payoff matrices in Appendix E). Each game's Nash set is parameterized by a scalar *selection coordinate* c, except `polytope4` whose Nash set is genuinely two-dimensional. `pennies_safe` and `two_safe` are matching-pennies cores augmented with one or two value-preserving "safe" actions, giving *symmetric* 1-D Nash segments (max-entropy coordinates 0.333 and 0.500); `dup_action` is a duplicated-action degeneracy (max-entropy 0.250); `asym_safe` is an *asymmetric* safe-action game whose Nash face is $P_0 = (p_0,\, 2p_0,\, 1 - 3p_0)$, $p_0 \in [0, 1/3]$ (max-entropy 0.218); `polytope4` has a two-dimensional Nash polytope (max-entropy point $(0.161, 0.256)$); and `kuhn` is Kuhn poker, whose equilibria form a one-parameter family indexed by the first player's bluff frequency (max-entropy bluff 0.201; $V^\star = -1/18$). The symmetric games are controls. Section 4.3 adds a randomized ensemble of 180 asymmetric matrix games with analytically known Nash faces.

How we measure which member is selected. Because each Nash set is known in closed form, a converged solver's output can be located on it directly. Our primary measurement is the *selection coordinate* of the returned profile (e.g. the mass on the free action, or the bluff frequency), compared with the analytic max-entropy coordinate. We complement it with two coordinate-free measures of set-member (dis)similarity: the mean policy entropy (2), and the mean Jensen–Shannon divergence between the profiles returned by two solvers. We also report exploitability (1) and support size (actions with probability $> 10^{-4}$). A solver is called *converged* if its exploitability is below 0.02; this operational threshold is loose, so we report the actual exploitabilities in Table 2 (Appendix A) and discuss what it does and does not license in §4.7 and §7.

Protocol. All solvers start from the uniform policy unless stated otherwise; iteration budgets are fixed per solver and not tuned per game. The fixed budget is not why the fixed-magnet MMD and MWU baselines fail on the asymmetric games: their non-convergence is the expected cycling of insufficiently regularized and unregularized learning in zero-sum games (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018), which larger budgets do not remove.

4 RESULTS

Ground truth. We first verify that the analytic families are Nash and *exhaust* the Nash set: every member has exploitability $< 10^{-15}$ ($< 10^{-6}$ across the swept Kuhn family), profiles breaking the defining ratio constraints are exploitable (e.g. 0.08 on `polytope4`), and in each matrix game player 1's equilibrium strategy is unique while player 0's indifference pins the active-row ratio, leaving only the safe-mass (or bluff) parameter free.

4.1 CORE RESULT: SELECTION ACROSS GAMES

Table 1 reports the selection coordinate of every solver on every game; Figure 3 (Appendix B) shows the corresponding policy entropies. On the *symmetric* games every converging solver *except fictitious play* agrees on the max-entropy coordinate; FP drifts to the boundary even here (e.g. 0.000 on `pennies_safe`, where the others select 0.333). On the *asymmetric* matrix games the families separate: R-NaD sits at the max-entropy coordinate, while CFR and CFR$^+$ land on a distinctly lower-entropy point. On `asym_safe`, R-NaD selects 0.218 (the max-entropy value) versus CFR$^+$'s 0.275; on `polytope4`, R-NaD selects 0.162 against the analytic 0.161 while CFR$^+$ selects 0.246. Kuhn is less clean. R-NaD selects bluff 0.180 (entropy 0.261 versus the maximal 0.262). The reach-normalized CFR$^+$ variant selects 0.068, but this is an artifact of the normalization: textbook CFR and CFR$^+$ select 0.203 and 0.242 (Appendix F), and FP selects 0.203. In Kuhn, then, the averaging methods do not drift to low entropy, and the separation we report is a matrix-game result. The Jensen–Shannon divergences to the R-NaD profile corroborate this coordinate-free: essentially zero on symmetric games, growing on asymmetric ones (CFR$^+$–R-NaD JS 0.009 on `asym_safe`, 0.016 on `polytope4`), with MWU an order of magnitude larger and collapsed support ($\approx$1.1–1.5 actions versus 2–3).

Selection is meaningful only for solvers that reach a Nash equilibrium. The regret-averaging methods and R-NaD converge on the matrix games (vanilla CFR misses the threshold on Kuhn within budget;

Table 1: Selection coordinate by solver and game. ME = analytic max-entropy coordinate. Bold marks agreement with ME to within 0.02; a dagger (†) marks solvers that did not reach exploitability < 0.02 (coordinates reflect non-convergence, not selection). `polytope4` coordinates are the first component p_0 of its 2-D selection. ‡CFR and CFR⁺ on `kuhn` are the reach-normalized variant of §2; textbook CFR/CFR⁺ give 0.203/0.242 (Appendix F). Entropies and exploitabilities for every cell are in Table 2.

Game	ME	CFR	CFR⁺	FP	MMD	R-NaD	MWU
`pennies_safe`	0.333	**0.333**	**0.333**	0.000	**0.333**	**0.333**	**0.333**
`dup_action`	0.250	**0.248**	**0.250**	0.504	0.008†	**0.250**	0.000†
`two_safe`	0.500	**0.500**	**0.500**	0.000	**0.500**	**0.500**	**0.500**
`asym_safe`	0.218	0.293	0.275	0.332	0.015†	**0.218**	0.997†
`polytope4`	0.161	0.264	0.246	0.332	0.001†	**0.162**	1.000†
`kuhn`	0.201	0.005†‡	0.068‡	**0.203**	0.055†	0.180	0.000†

CFR⁺ reaches it), whereas fixed-magnet MMD and MWU cycle on `dup_action` and all asymmetric games (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018); their daggered coordinates in Table 1 reflect non-convergence, not selection (Figure 6). MMD therefore contributes evidence only on the two symmetric controls, where every solver agrees. *Wherever selection is at issue, our evidence about regularized last-iterate methods is evidence about R-NaD's moving reference*, and we phrase all claims accordingly.

4.2 Geometry of the selection

Figure 1 draws the landing points of the converged solvers directly on the Nash set: Kuhn's 1-D family and the 2-D `polytope4` face. On `polytope4` the max-entropy member is a non-trivial interior point $(0.161, 0.256)$ rather than a segment midpoint. R-NaD lands at $(0.162, 0.257)$—max-entropy to three decimals in *both* coordinates, with exploitability 0.00—whereas CFR⁺ lands at $(0.246, 0.135)$ with exploitability 0.01. This is the cleanest single piece of evidence that the regularized fixed point computes the I-projection (3) of the uniform reference: on a 2-D set the claim has two independent degrees of freedom to satisfy, ruling out a 1-D midpoint artifact.

A fixed magnet cannot achieve the same. Sweeping the temperature λ of a *fixed* uniform magnet on Kuhn (Figure 5), the selected coordinate varies monotonically with λ and crosses the max-entropy value near $\lambda \approx 0.2$, but no setting reaches the convergence threshold: for large λ the fixed point is a biased, regularized equilibrium, and once $\lambda \lesssim 0.2$ exploitability rises again as the dynamics begin to cycle. The regularization that biases the solution is also what stabilizes it. R-NaD's moving reference avoids this bias–stability trade-off, reaching exploitability 0.0000 (four decimals) at 99.7% of the maximum entropy. Its coordinate, however, is 0.180 rather than the max-entropy 0.201. This is not slow motion along a flat entropy landscape: the coordinate is 0.180226 at $3{,}000$, $12{,}000$ and $48{,}000$ iterations, with exploitability 2×10^{-12}, so it is an exact equilibrium and a genuine fixed point. It does move with the regularization temperature (0.178 at $\lambda = 1$, 0.188 at $\lambda = 0.2$; Appendix F). In Kuhn, R-NaD thus selects a member that is near, but not at, the maximum of $\mathcal{H}$, and which member depends on a hyperparameter; we return to this in §5.

4.3 A randomized ensemble of safe-action games

The six games above are hand-built, so the split could be a property of those constructions. We test this on a randomized population of $N = 180$ *random asymmetric safe-action* games: a random 2×2 zero-sum core with a verified unique interior mixed equilibrium, augmented with $k \in \{1, 2, 3\}$ safe rows worth $V^\star$ (60 games each). Each game has an analytically known, *skewed* Nash face whose max-entropy member is computed in closed form, and every face member is Nash to machine precision ($< 10^{-15}$). We run the two solvers of interest, CFR⁺ and R-NaD. R-NaD's step size is set per game to $\eta = 1/\mathrm{range}(M)$, which uses only the payoff range and no equilibrium information; it is needed for *stability* (at fixed $\eta = 1$ R-NaD cycles on roughly half the ensemble), whereas regret matching is scale-invariant and needs no adjustment (Appendix C). R-NaD then converges on 100% of the games and CFR⁺ on 90% at its fixed budget; the comparison uses the 162 games on which both converge.

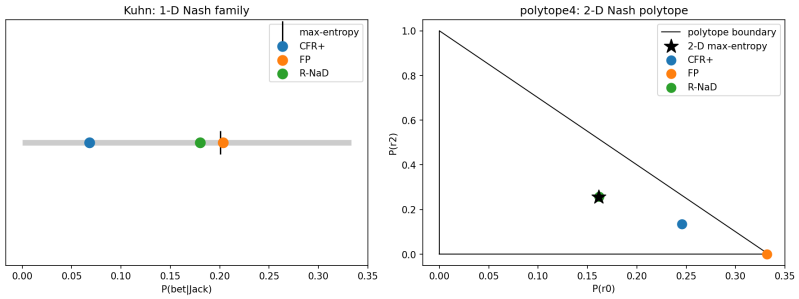

Figure 1: Where each converged solver lands on the actual Nash set. Left: Kuhn's 1-D family (bluff frequency). Right: the 2-D Nash polytope of `polytope4`. On `polytope4` R-NaD coincides with the max-entropy point and CFR$^+$ does not; on Kuhn R-NaD is near, but not at, the max-entropy member, and the CFR$^+$ point is the reach-normalized variant (textbook CFR$^+$ lands at 0.242).

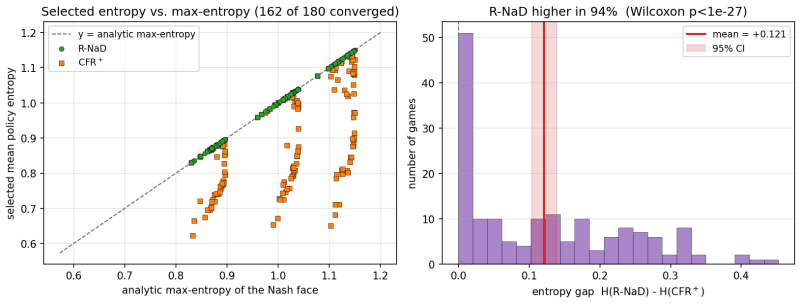

Figure 2: The 180-game random ensemble of asymmetric safe-action games. **Left:** selected mean policy entropy versus the analytic max-entropy of each game's Nash face; R-NaD (green) lies on the diagonal, CFR$^+$ (orange) lies below; bands correspond to $k \in \{1, 2, 3\}$ safe rows. **Right:** distribution of the entropy gap over the 162 games where both converge; mean $+0.121$, paired Wilcoxon $p < 10^{-27}$.

The split is stark and consistent (Figure 2). On all 162 games R-NaD lands on the analytic max-entropy member: its entropy is within 0.01 of the face maximum and its selection coordinate within 6×10^{-4} (median 2×10^{-4}) of the analytic one, on faces that are two- and three-dimensional for $k \geq 2$. CFR$^+$ has lower entropy than R-NaD in 153 games (94%), with mean entropy gap $H(\text{R-NaD}) - H(\text{CFR}^+) = +0.121$ (95% bootstrap CI $[+0.103, +0.139]$). A paired one-sided Wilcoxon signed-rank test rejects the null of no gap at $p < 10^{-27}$; the sign test over the 161 non-tied pairs gives $153/161$, $p < 10^{-35}$. Of the remaining nine games one is a tie and in eight CFR$^+$'s entropy is marginally *higher* than the face maximum, which can occur only for a profile slightly off the Nash set, i.e. within the residual exploitability our threshold permits. The phenomenon of §4.1 is therefore robust across this construction family. All ensemble games share one generator, however, so the result does not yet cover structurally different degeneracies (duplicated or affine-dependent actions, degeneracy for both players, random higher-dimensional games conditioned on non-uniqueness).

4.4 Robustness to seed and iteration budget

The tabular solvers are *deterministic*: they iterate on exact counterfactual values from a fixed uniform start, with no sampling, so selection is seed-invariant by construction (re-running across seeds gives a standard deviation of exactly zero); differences between solvers are therefore not noise. Nor is the separation a symptom of incomplete convergence that more compute would erase. Sweeping the iteration budget over two orders of magnitude on `asym_safe`, R-NaD stays at the max-entropy value 0.218 at every budget, while CFR$^+$'s selected coordinate is budget-*dependent* and moves *farther* from max-entropy as the budget grows ($0.272 \to 0.283$ from $2{,}000$ to $200{,}000$ iterations) even as its exploitability continues to fall (Figure 8). Additional computation does not close the gap between the

two solvers; it widens it. Whether the CFR$^+$ average eventually settles, and where, is not determined by this sweep.

4.5 Anchor-following: the selection depends on the reference

Figure 4 sweeps the *initial reference* on Kuhn, biasing the starting policy from $b = 0.1$ to 0.9 toward the betting action. The selected coordinate is *not* invariant: it moves over the range 0.08–0.18, all at exploitability 0.0000. R-NaD is therefore *anchor-following*—regularized toward its initial reference, which it then tracks. With a uniform reference (the standard default) it selects the (near-)max-entropy member; biasing the reference moves the selection in the direction of the I-projection of the reference for moderate biases, though the tracking is approximate—and non-monotone at extreme biases—plausibly because the reference moves during training, so the limit depends on the *trajectory* of references rather than on a single projection of ρ_0 (Appendix D). Max-entropy selection is thus the special case of a uniform reference, not an initialization-independent property.

4.6 Is the $\max(R, 0)$ projection the cause of the drift?

A natural hypothesis for why CFR-family averages land on low-entropy members is that regret matching's positive-orthant projection $\max(R, 0)$ zeroes out actions and so structurally favors sparse policies. We test this with Hedge, which uses the *same* regrets but replaces the hard projection with a softmax ($\eta = 1$), retaining strategy averaging. If the projection were the cause, removing it should move the selection toward max-entropy. It does the opposite: on `asym_safe`, CFR selects coordinate 0.293 (entropy 0.796) while Hedge selects 0.334 (entropy 0.649); the same holds on `polytope4` (0.264 vs. 0.334). The $\max(R, 0)$ projection is therefore not *necessary* for the drift, which appears to be a property of regret-*averaging* dynamics more broadly. One caveat: we ran a single step size. With fixed η and growing cumulative regrets Hedge approaches a best-response dynamic—its coordinates nearly coincide with fictitious play's (0.332)—so the *magnitude* of its drift is likely step-size dependent, and a sweep over η is needed to quantify it. The mechanism of the drift remains open.

4.7 Where downstream consequences of selection appear

Against a best-responding opponent selection cannot matter—all members share $V^\star$—so any effect must be off-path, against sub-optimal opponents.

Matrix games: none. On `asym_safe` and `polytope4` (and every safe-action game of the ensemble) each equilibrium strategy of player 0 is an *equalizer*: for $P_0 = (p_0, 2p_0, 1 - 3p_0)$ both columns pay exactly $1/3$. Every member therefore earns $V^\star$ against *every* opponent, optimal or not, and the members are behaviorally distinct but payoff-equivalent. Evaluating the solver outputs against perturbed opponents does show small differences (0.003–0.008), but these measure the outputs' residual distance from the Nash set, not selection.

Kuhn: linear in the bluffing frequency. We evaluate player 0's equilibrium strategies against opponents whose calling probability when facing a bet is shifted by δ (over-folding for $\delta < 0$, over-calling for $\delta > 0$; Appendix E). For the exact equilibrium with bluffing frequency c the value is

$$u(c, \delta) \;=\; V^\star + \tfrac{1}{2}\, c\, |\delta|, \tag{5}$$

which we verified numerically on the analytic family. Off-path value thus increases linearly with the selected coordinate: members that bluff (and value-bet) more profit more from either kind of calling error, and the best hedge against this opponent class is the boundary member $c = 1/3$, not the max-entropy member. Figure 7 (Appendix B) shows the measured curves for two solver outputs, R-NaD ($c = 0.180$) and the reach-normalized CFR$^+$ variant ($c = 0.068$): the former is weakly better at every δ and ties at $\delta = 0$. The measured maximal gap is 0.0181; by (5) the exact equilibria at those coordinates differ by 0.0168 at $|\delta| = 0.3$, i.e. 30% of $|V^\star| = 1/18$, so about 93% of the measured gap is selection and the rest is residual exploitability of the CFR$^+$ output. By the same formula the members returned by FP and textbook CFR (0.203) and textbook CFR$^+$ (0.242) are slightly better hedges than R-NaD's. The lesson is not that maximum entropy is the robust choice, but that in a sequential game *which* member a solver returns has first-order off-path consequences, while in our

matrix games it has none. With a single extensive-form game we make no claim about how this scales with game structure.

5 DISCUSSION

What is established. (i) In 2p0s games with non-singleton Nash sets, the returned member is a deterministic function of the solver. On matrix games the separation between R-NaD and regret averaging is gated by the geometry of the Nash set—absent on symmetric sets, present on asymmetric ones (§4.1)—and persists, indeed widens, with the iteration budget (§4.4). (ii) With a uniform reference, R-NaD selects the maximum-entropy member on every matrix game tested: the five hand-built games, including the interior of a 2-D polytope, and all 162 converged ensemble games, where CFR$^+$ lies below ($p < 10^{-27}$; §4.3). (iii) That selection is anchor-following, not unconditional (§4.5). (iv) On matrix games regret-averaging methods land on lower-entropy members, FP even on the symmetric games, and the $\max(R, 0)$ projection is not necessary for this (§4.6). (v) In Kuhn the picture does not carry over: FP and textbook CFR land at the max-entropy member, textbook CFR$^+$ above it, R-NaD below it at a temperature-dependent point, and reach-normalizing the counterfactual values moves CFR/CFR$^+$ to the other end of the family (Appendix F). (vi) Selection is payoff-irrelevant on our matrix games and first-order off-path in Kuhn (§4.7). Fixed-magnet MMD does not converge where selection is at issue, so none of these statements extends to regularized last-iterate methods as a class.

What is conjectured, and what is open. The explanation most consistent with the matrix-game data is that the moving-reference dynamics compute the I-projection (3) of the initial reference onto the Nash polytope.

Conjecture 1 (R-NaD selection in matrix games)**.** *Let G be a 2p0s matrix game with Nash polytope $\mathcal{N}(G)$ and let R-NaD be run from the uniform reference $\rho_0 = \mathrm{unif}$ with a step size for which it converges. Then its limit is $\Pi_{\mathcal{N}}(\mathrm{unif}) = \sigma^{\mathrm{ME}}$, the maximum-entropy equilibrium.*

Matrix games: strong evidence. The conjecture matches every matrix game we ran, to the precision of the analytic maximizer. It is also what theory leads one to expect: a 2p0s matrix game is a linear program, max-entropy selection is the vanishing-regularization limit of an entropic penalty (Weed, 2018; Cuturi, 2013), and the moving reference is the algorithmic counterpart of the logit-QRE homotopy (Turocy, 2005; Gemp et al., 2022). What is new is not the principle but the verification that a deployed solver realizes it, and that the regret-averaging solvers do not. *General references: partial evidence.* For $\rho_0 \neq \mathrm{unif}$ the natural extension is that the limit is $\Pi_{\mathcal{N}}(\rho_0)$. Our only test (Kuhn, §4.5) shows the correct direction for moderate biases but not equality, so we do not conjecture it. *Extensive-form games: the naive analogue fails.* In Kuhn, the uniform-reference limit is an exact equilibrium at bluff 0.180, whereas σ^{ME} under (2) is at 0.201; the coordinate is unchanged by a $16\times$ larger budget and moves with λ (§4.2). This refutes the extensive-form analogue of Conjecture 1 with $\mathcal{H}$ as the entropy, for our per-information-set implementation of R-NaD. The dependence on λ further suggests that no *fixed* divergence makes the limit an I-projection of ρ_0 alone: as under biased references, the limit appears to depend on the trajectory of references. Whether a reach-weighted (dilated) formulation of the dynamics (Hoda et al., 2010; Kroer et al., 2020) restores an exact characterization, and whether the limit as $\lambda \to 0$ is the max-entropy member, are open. Appendix D discusses the relation between the uniform- and biased-reference regimes.

6 RELATED WORK

Regret minimization underlies CFR (Zinkevich et al., 2007) and CFR$^+$ (Tammelin, 2014; Bowling et al., 2015), which deploy time-averaged strategies; fictitious play (Brown, 1951) converges in value in 2p0s games (Robinson, 1951). A separate line regularizes the dynamics toward a reference to obtain last-iterate convergence: NeuRD (Hennes et al., 2020), the regularization analysis of Pérolat et al. (2021), R-NaD/DeepNash (Pérolat et al., 2022), and MMD (Sokota et al., 2023), which unifies regularized RL, quantal-response equilibria (McKelvey & Palfrey, 1995), and 2p0s solving; see also Cen et al. (2021) for entropy-regularized matrix and Markov games and Liu et al. (2023) for regularization in extensive-form games. Periodically updating the anchor of a perturbed game, as R-NaD does, is also the mechanism behind recent perturbed-mirror-descent and

mutation-driven methods (Abe et al., 2022; 2024). These works establish *that* the regularized or perturbed dynamics converge to an equilibrium; we ask *which* equilibrium, when there are many. Closest to our question, Moon (2026) shows that when a player's equilibrium set is non-singleton (positive measure, fully mixed), the empirical strategy of fictitious play cannot converge to any single equilibrium point, clarifying the gap between set and point convergence for one averaging method; and Cai et al. (2025) study whether the last iterates of regret-matching algorithms converge at all. Neither characterizes which member is approached. The instability of unregularized dynamics is well documented (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018), and no-regret/FTRL last iterates are unstable at mixed Nash equilibria (Vlatakis-Gkaragkounis et al., 2020). Optimistic methods obtain last-iterate convergence without regularization (Daskalakis & Panageas, 2019; Wei et al., 2021); which member they select on a Nash polytope is a natural complementary experiment, as it would separate "regularization implies I-projection" from "last-iterate convergence implies max-entropy". Equilibrium selection as a refinement question dates to Harsanyi & Selten (1988), and refinements such as perfection (van Damme, 1991) select members by robustness to trembles; our selection principle is informational—the I-projection / maximum-entropy member (Jaynes, 1957; Csiszár, 1975)—established as a regularization limit outside game solving (Weed, 2018; Cuturi, 2013) and singled out as canonical for games before (Balduzzi et al., 2018). The logit-QRE homotopy literature (McKelvey & Palfrey, 1995; Turocy, 2005; Gemp et al., 2022) generically concerns games with a unique limiting endpoint; our degenerate testbed is the non-generic case it sets aside. Our contribution is the ground-truth-verified comparison of *which* Nash-polytope member each standard solver returns.

7 Limitations

Our testbed is tabular and small, chosen so that ground truth is exact; we do not study function-approximation or sampling effects, which could interact with selection. Conjecture 1 is unproven, and its naive extensive-form analogue fails in Kuhn (§5). Our CFR, CFR$^+$ and Hedge implementations use reach-normalized counterfactual values and average behavioral policies per information set; this is exactly standard in matrix games, which carry the core and ensemble results, but in Kuhn it changes the selected member (Appendix F), so the Kuhn entries for these solvers in Table 1 and Figures 1 and 3 describe a non-standard variant. The checks of Appendix F are sanity checks, not a systematic study. The convergence threshold (0.02) is loose: matrix-game selection coordinates are separated by much more than the residual exploitability, but quantities of that order (off-path values, the entropy of eight ensemble games) can be contaminated by it, which is why §4.7 relies on exact equilibria. The random ensemble covers one construction family of matrix games, only R-NaD and CFR$^+$ were run on it, and Kuhn is the sole extensive-form instance, so we make no general claim about sequential games. The Hedge ablation uses one step size, and R-NaD's convergence (not its selected member) is sensitive to payoff scale (Appendix C). Finally, the max-entropy results are conditional on uniform initialization; warm-started deployments will select a different member.

8 Conclusion

Solvers that are guaranteed to find *a* Nash equilibrium do not find *the same* one. On asymmetric Nash polytopes of matrix games, R-NaD from a uniform reference selects the maximum-entropy member (the I-projection of its reference), regret-averaging methods land on lower-entropy members, and more computation widens the gap rather than closing it. The selection follows the reference, and the $\max(R, 0)$ projection is not necessary for regret-averaging drift. Kuhn poker shows both why this matters and how much is open: there the returned member has first-order consequences against imperfect opponents, yet R-NaD's member is near, not at, maximum entropy and depends on the regularization temperature, and CFR's member depends on an implementation detail. Next steps are to prove Conjecture 1, characterize R-NaD's selection in extensive-form games, explain regret-averaging drift, test optimistic and annealed-regularization methods on structurally different degenerate games, and extend to general-sum games and function approximation.

AI USE STATEMENT

In this work, we used generative AI tools (a large-language-model coding assistant) to assist in implementing methods (the tabular engine, the experiment notebook, and the sanity-check script), to provide feedback on research methodology and experiments, and to assist in interpreting results. Specifically, while revising the paper in response to reviewer feedback, the assistant audited the code against the text, identified that the engine's CFR family is a reach-normalized variant that differs from textbook CFR in Kuhn poker, identified that equilibria of the safe-action matrix games are payoff-equivalent against every opponent, and proposed and ran the sanity checks of Appendix F and the closed-form robustness expression of §4.7; the authors checked these findings and decided how to report them. The research question, the testbed, the experimental design of the main study, and Conjecture 1 originate with the authors; we have not used generative AI tools to provide ingredients for or write proofs (the paper contains none), and generating synthetic data sets, translation, data-set cleaning, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit software code, to draft and edit parts of the paper for readability and to tighten the scope of claims, to help identify and check relevant literature, and to format references and LaTeX for this template. We have reviewed all AI-assisted work: code was tested against analytically known equilibria (every analytic Nash family is checked to have exploitability below 10^{-6}, and solver outputs are compared against closed-form maximizers), the sanity checks are reproducible from the supplementary code, every bibliographic entry was checked against its source, and all text was read and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies equilibrium-finding algorithms on small abstract games with synthetic payoffs. It involves no human subjects, personal data, or deployed systems, and we are not aware of ethical concerns specific to it.

REPRODUCIBILITY STATEMENT

All games, solvers, metrics, and the random-ensemble generator are implemented in a single self-contained NumPy engine and an executable notebook, provided as anonymized supplementary material; together they regenerate every number and figure in the paper deterministically on a single CPU in minutes. Payoff matrices, hyperparameters, iteration budgets, the implementation details of the CFR variants, and the downstream evaluation protocol are given in Appendix E; the Kuhn sanity checks in Appendix F; the ensemble generator and statistical methodology in Appendix C; and per-cell entropies and exploitabilities in Appendix A.

REFERENCES

Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1–10. PMLR, 2022.

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31–80. PMLR, 2024.

James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 321–338. ACM, 2018. doi: 10.1145/3219166.3219235.

David Balduzzi, Karl Tuyls, Julien Pérolat, and Thore Graepel. Re-evaluating evaluation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015. doi: 10.1126/science.1259433.

George W. Brown. Iterative solution of games by fictitious play. In Tjalling C. Koopmans (ed.), *Activity Analysis of Production and Allocation*, pp. 374–376. Wiley, New York, 1951.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890, 2019. doi: 10.1126/science.aay2400.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Last-iterate convergence properties of regret-matching algorithms in games. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27952–27964. Curran Associates, Inc., 2021.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, volume 124 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 27:1–27:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPIcs.ITCS.2019.27.

Ian Gemp, Rahul Savani, Marc Lanctot, Yoram Bachrach, Thomas Anthony, Richard Everett, Andrea Tacchetti, Tom Eccles, and János Kramár. Sample-based approximation of Nash in large many-player games via gradient descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 507–515. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA, 1988. ISBN 9780262081733.

Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Pérolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, and Karl Tuyls. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 492–501. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2): 494–512, 2010. doi: 10.1287/moor.1100.0452.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi: 10.1103/PhysRev.106.620.

Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179 (1–2):385–417, 2020. doi: 10.1007/s10107-018-1336-7.

Mingyang Liu, Asuman Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=bPiHuNUNv_R.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995. doi: 10.1006/game.1995.1023.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2703–2717. SIAM, 2018. doi: 10.1137/1.9781611975031.172.

Jaehong Moon. Zero-sum fictitious play cannot converge to a point. arXiv preprint arXiv:2604.07544, 2026.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960.

Julien Pérolat, Rémi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8525–8535. PMLR, 2021.

Julien Pérolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Rémi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022. doi: 10.1126/science.add4679.

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, 1951. doi: 10.2307/1969530.

Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=DpE5UYUQzZH.

Oskari Tammelin. Solving large imperfect information games using CFR+. arXiv preprint arXiv:1407.5042, 2014.

Theodore L. Turocy. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior*, 51(2):243–263, 2005. doi: 10.1016/j.geb.2004.04.003.

Eric van Damme. *Stability and Perfection of Nash Equilibria*. Springer, Berlin, Heidelberg, 2 edition, 1991. doi: 10.1007/978-3-642-58242-4.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1380–1391. Curran Associates, Inc., 2020.

Jonathan Weed. An explicit analysis of the entropic penalty in linear programming. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1841–1855. PMLR, 2018.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *The Ninth International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=dx11_7vm5_r.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

A FULL METRIC TABLE

Table 2 reports, for every (game, solver), the selection coordinate, mean policy entropy, and exploitability. Daggers in Table 1 correspond to the exploitability values here that are not below 0.02.

Table 2: Full metrics: coordinate / entropy / exploitability.

Game	metric	CFR	CFR$^+$	FP	MMD	R-NaD	MWU
`pennies_safe`	coord/H/expl	0.333/0.90/0.00	0.333/0.90/0.00	0.000/0.69/0.01	0.333/0.90/0.00	0.333/0.90/0.00	0.333/0.90/0.00
`dup_action`	coord/H/expl	0.248/0.87/0.01	0.250/0.87/0.01	0.504/0.69/0.01	0.008/0.24/1.72	0.250/0.87/0.00	0.000/0.00/2.00
`two_safe`	coord/H/expl	0.500/1.04/0.00	0.500/1.04/0.00	0.000/0.69/0.01	0.500/1.04/0.00	0.500/1.04/0.00	0.500/1.04/0.00
`asym_safe`	coord/H/expl	0.293/0.80/0.01	0.275/0.81/0.01	0.332/0.64/0.01	0.015/0.39/3.47	0.218/0.85/0.00	0.997/0.01/1.99
`polytope4`	coord/H/expl	0.264/0.91/0.02	0.246/0.94/0.01	0.332/0.64/0.01	0.001/0.52/2.99	0.162/1.00/0.00	1.000/0.00/2.00
`kuhn`	coord/H/expl	0.005/0.18/0.04	0.068/0.23/0.01	0.203/0.26/0.00	0.055/0.25/0.25	0.180/0.26/0.00	0.000/0.00/0.33

B ADDITIONAL FIGURES

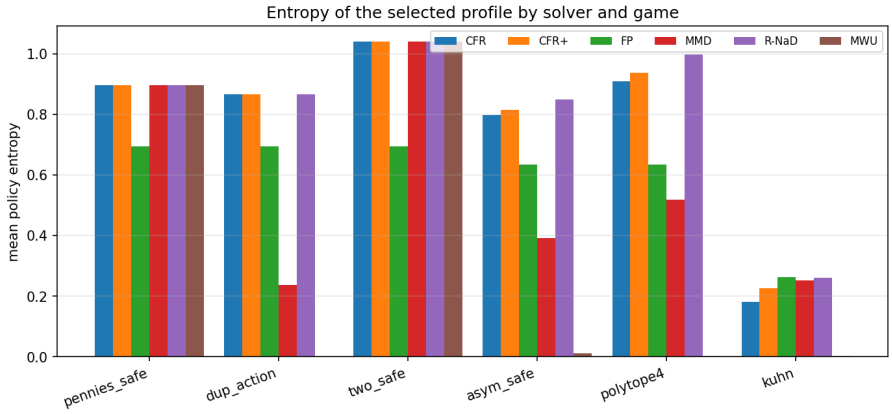

Figure 3: Mean policy entropy of the selected profile, by solver and game. On symmetric games all converging solvers except FP coincide; on the asymmetric matrix games (`asym_safe`, `polytope4`) R-NaD attains the highest entropy while CFR/CFR$^+$ sit lower and MWU collapses. The `kuhn` bars for CFR/CFR$^+$ are the reach-normalized variant (see Table 1).

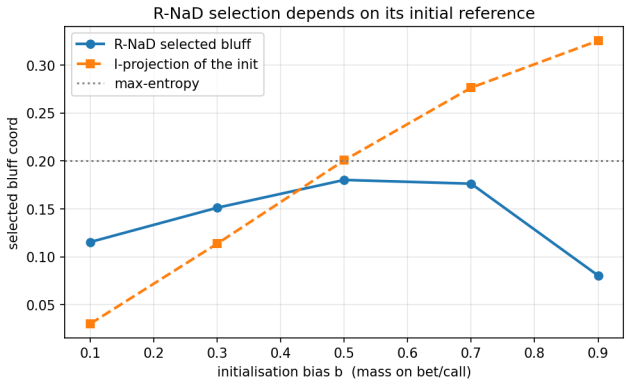

Figure 4: R-NaD on Kuhn is anchor-following (§4.5). As the initial reference is biased (x-axis), the selected equilibrium (solid) shifts over 0.08–0.18 (all at exploitability 0.0000), tracking the I-projection of the reference (dashed) in direction only. The near-max-entropy selection (dotted line: max-entropy) is obtained at a uniform reference, not unconditionally.

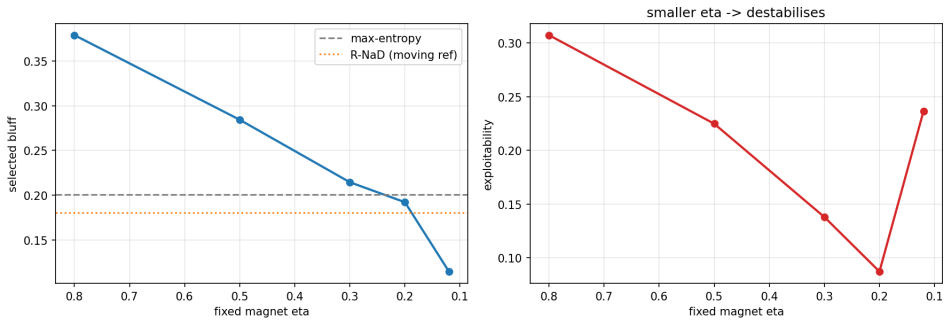

Figure 5: Bias versus convergence on Kuhn for a *fixed* uniform magnet (§4.2). The x-axis, labeled "eta" in the plot, is the regularization temperature λ of Eq. (4) (step size fixed to 1). Left: as λ decreases (right to left), the selected coordinate moves monotonically and crosses max-entropy near $\lambda \approx 0.2$. Right: exploitability falls with the regularization bias and then rises again below $\lambda \approx 0.2$, never reaching the 0.02 threshold. R-NaD's moving reference (dotted, left) avoids the trade-off, reaching exploitability $\sim 10^{-12}$ at 99.7% of maximum entropy.

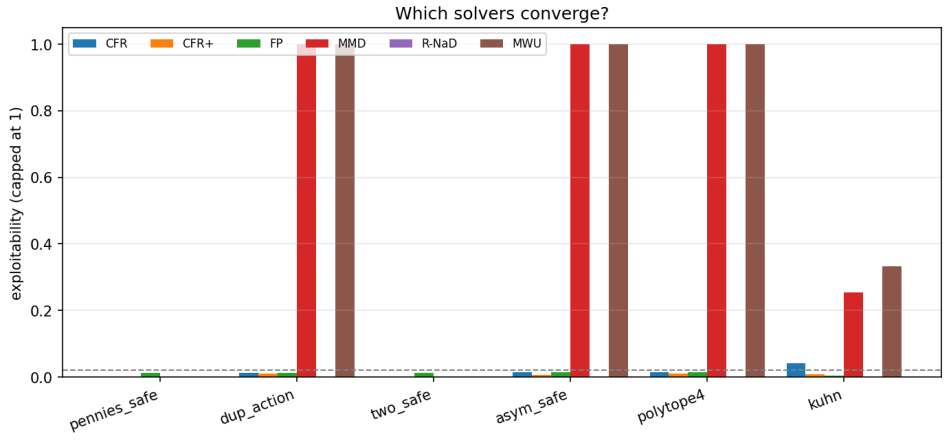

Figure 6: Exploitability (capped at 1) by solver and game; the dashed line is the convergence threshold (§4.1). Fixed-magnet MMD and unregularized MWU fail to converge on `dup_action` and the asymmetric games. Vanilla CFR misses the threshold on Kuhn within budget; CFR$^+$ reaches it.

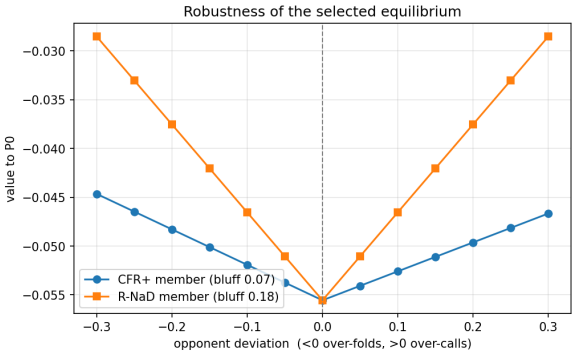

Figure 7: Off-path value of two returned Kuhn profiles against opponents with shifted calling frequency. The R-NaD output (bluff 0.180) is weakly better than the reach-normalized CFR$^+$ variant's output (bluff 0.068) at every δ, tying at $\delta = 0$, as Eq. (5) predicts for any pair of members ordered by bluffing frequency.

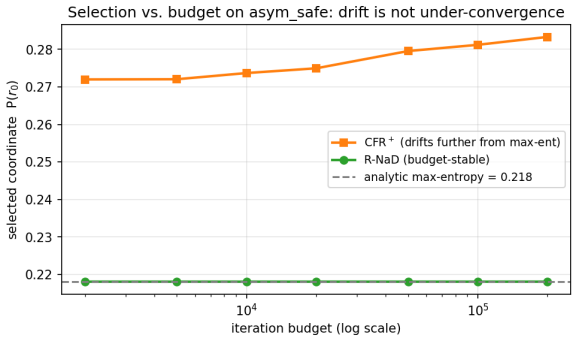

Figure 8: Selection versus iteration budget on `asym_safe` (log x-axis; §4.4). R-NaD stays at the analytic max-entropy coordinate 0.218 at every budget; CFR$^+$ moves *away* from it as the budget grows. The separation between the solvers is not under-convergence.

C Ensemble generation and statistical methodology

Generation. Each of the $N = 180$ games is built from a random 2×2 zero-sum core with entries of magnitude $\mathcal{U}[0.5, 3.0]$ (diagonal positive, off-diagonal negative); a draw is retained only when its unique mixed equilibrium is interior—both equilibrium probabilities in $(0.05, 0.95)$—and the core value $V^\star$ is computed in closed form. The core is augmented with $k \in \{1, 2, 3\}$ safe rows equal to $[V^\star, V^\star]$; the three values of k contribute 60 games each. Because the safe rows are worth $V^\star$ against any opponent, player 1's indifference is governed only by the two active rows; holding them in the equilibrium ratio keeps player 1 at its unique strategy, so each game has an analytically known, skewed Nash face whose max-entropy member—the I-projection of the uniform reference—is computed in closed form. Every face member is Nash to machine precision ($< 10^{-15}$).

Solvers and step-size rescaling. Only CFR$^+$ ($20{,}000$ iterations) and R-NaD ($18{,}000$ iterations, $\lambda = 0.5$, reset period 60) are run on the ensemble. R-NaD places the step size inside the softmax exponent (Eq. 4) and is therefore sensitive to payoff scale; on the ensemble we set $\eta = 1/\mathrm{range}(M)$ per game, a quantity read off the payoff matrix that uses no information about the equilibrium. This is a *stability* requirement: at a fixed $\eta = 1$, R-NaD cycles on games with large payoff range (the instability discussed in §4.2), failing to converge on roughly half the ensemble; with the rescaling it converges on all of it. Regret matching is invariant to payoff scale, so the same rescaling would leave CFR$^+$ unchanged. CFR$^+$ converges on 90% of the games at its fixed budget; the 18 games on which it does not are excluded from the paired comparison.

Choice of statistical tests. We use the Wilcoxon signed-rank test rather than a paired t-test because the per-game entropy gap is strongly right-skewed and floored near zero—R-NaD attains the max-entropy member (the maximum over the face), so the gap is essentially non-negative, with a long right tail—which violates the normality a t-test assumes. The signed-rank test is nonparametric yet, unlike the sign test, still uses the magnitude of each paired difference. The sign test (which discards the single tied pair, leaving 161) and the distribution-free bootstrap confidence interval are reported alongside and reach the same conclusion; the alternative is one-sided because the direction is predicted by the I-projection hypothesis. The comparison is paired (same game, two solvers), which rules out unpaired tests.

D Uniform versus biased references

A careful reader may ask how the anchor-following result (§4.5) bears on Conjecture 1: if R-NaD only *approximately* tracks the I-projection of a *biased* reference, why read the *uniform*-reference matrix-game result as exact rather than approximate-but-close? Two points distinguish the regimes. First, the quality of evidence differs in kind: under a uniform reference the selection coincides with σ^{ME} to the precision of the analytic maximizer on five matrix games, on all 162 converged ensemble games, and at the interior of the two-dimensional polytope—$(0.162, 0.257)$ vs. $(0.161, 0.256)$—a

two-degree-of-freedom target that an approximate mechanism would miss generically. The biased-reference sweep, by contrast, was run only on Kuhn and reports a *trend*. Second, a candidate confound is specific to biased references: with a non-uniform initial reference, the moving-reference schedule traverses a sequence of regularized equilibria anchored to a *changing* reference, so the deployed limit can reflect the *trajectory* of references rather than a single projection of ρ_0. We have not tested biased references on matrix games, where this explanation could be checked against exact ground truth; doing so is the natural next experiment. The Kuhn uniform-reference result (0.180 vs. 0.201) is discussed in §5 and Appendix F: the flat entropy landscape makes a small entropy deficit (0.3%) compatible with a visible coordinate gap, but the coordinate is a budget-independent fixed point that moves with λ, so the gap is real and the uniform-reference limit in Kuhn is also trajectory-dependent.

E Implementation details

Games. Payoffs are to the row player (player 0). `pennies_safe`: $\left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \\ 0 & 0 \end{smallmatrix}\right]$, coordinate $=$ mass on the safe row. `two_safe`: the same with two safe rows, coordinate $=$ total safe mass. `dup_action`: $\left[\begin{smallmatrix} 1 & 1 & -1 \\ -1 & -1 & 1 \end{smallmatrix}\right]$, coordinate $=$ column player's mass on the first duplicate. `asym_safe`: $\left[\begin{smallmatrix} 3 & -1 \\ -1 & 1 \\ 1/3 & 1/3 \end{smallmatrix}\right]$, coordinate $= p_0$, the mass on the first row. `polytope4`: `asym_safe` with a second safe row; Nash set $\{(p_0, 2p_0, p_2, 1 - 3p_0 - p_2)\}$, coordinates (p_0, p_2). `kuhn`: standard three-card Kuhn poker; player 0's equilibria are indexed by the bluff frequency $P(\text{bet} \mid \text{Jack}) \in [0, 1/3]$ and player 1's equilibrium strategy is unique. Analytic max-entropy members are obtained by maximizing (2) over the closed-form family on a fine grid.

Solvers. All solvers start from the uniform profile with step size $\eta = 1$ (except on the ensemble, Appendix C). Budgets: CFR, CFR$^+$, MMD, MWU, Hedge $20{,}000$ iterations; FP $15{,}000$; R-NaD $12{,}000$. MMD: $\lambda = 0.1$, fixed uniform magnet. R-NaD: $\lambda = 0.5$, reference reset to the current policy every 60 iterations. Hedge: $\sigma_{t+1} \propto \exp(R)$. As described in §2, the engine computes exact counterfactual values by full tree traversal and normalizes them by the counterfactual reach of the information set; CFR and Hedge accumulate $\sum_t \sigma_t(\cdot \mid I)$ and CFR$^+$ accumulates $\sum_t t\, \sigma_t(\cdot \mid I)$, normalized per information set, without own-reach weighting. FP averages best responses with own-reach (realization) weighting. Exploitability is exact, by enumeration of pure best responses.

Downstream evaluation. In Kuhn, the opponent (player 1) plays its unique equilibrium strategy except at its three "facing a bet" information sets, where the calling probability is shifted by δ (clipped to $[0, 1]$), for 13 equally spaced $\delta \in [-0.3, 0.3]$; negative δ over-folds and positive δ over-calls. Eq. (5) was checked on the analytic family for $c \in \{0.068, 0.180, 0.201, 1/3\}$ and agrees to numerical precision. The curves in Figure 7 use the frozen player-0 solver outputs of Table 1. For the matrix games, the equalizer property of §4.7 was verified numerically (value $1/3$ to 10^{-12} for every member against arbitrary column strategies).

F Sanity checks in Kuhn poker

Kuhn is the only game in the testbed where the engine's CFR family differs from the textbook algorithm and where R-NaD misses the max-entropy coordinate. Table 3 reports two small checks. *Textbook CFR.* We re-ran CFR and CFR$^+$ with unnormalized counterfactual values and the own-reach-weighted average strategy ($20{,}000$ iterations, simultaneous updates). The selected member moves from the low-bluff end of the family to the max-entropy member (CFR) or above it (CFR$^+$), with lower exploitability than the variant. On `asym_safe` the textbook and variant CFR$^+$ coincide (0.275), as they must. The low-entropy Kuhn members in Table 1 are therefore a consequence of reach normalization, not of regret averaging. *R-NaD.* The uniform-reference R-NaD coordinate is independent of the budget to six digits at exploitability $\sim 10^{-12}$, and varies with the temperature λ. These checks are included in the supplementary code (`sanity_checks.py`).

Table 3: Sanity checks in Kuhn (max-entropy bluff 0.201). Top: reach-normalized variant used in the main experiments versus textbook CFR/CFR$^+$. Bottom: R-NaD from a uniform reference (reset period 60, $\eta = 1$).

Solver	setting	bluff coordinate	$\mathcal{H}$	exploitability
CFR	reach-normalized variant	0.005	0.180	4.1×10^{-2}
CFR$^+$	reach-normalized variant	0.068	0.225	7.4×10^{-3}
CFR	textbook	0.203	0.261	3.3×10^{-3}
CFR$^+$	textbook	0.242	0.258	1.5×10^{-3}
R-NaD	$\lambda = 0.5$, 3,000 iterations	0.180226		1.6×10^{-8}
R-NaD	$\lambda = 0.5$, 12,000 iterations	0.180226		2.4×10^{-12}
R-NaD	$\lambda = 0.5$, 48,000 iterations	0.180226		2.4×10^{-12}
R-NaD	$\lambda = 1.0$, 24,000 iterations	0.178218		3.3×10^{-12}
R-NaD	$\lambda = 0.2$, 24,000 iterations	0.188296		1.8×10^{-12}

G COMPUTE

All experiments run end-to-end on a single commodity CPU (no GPU or cluster) in minutes. Several scoping choices—the tabular, exactly solvable testbed, the absence of function-approximation and sampling experiments, and the restriction of the random ensemble to matrix games—reflect this budget and the requirement of exact ground truth rather than a belief that the phenomena vanish at scale.