WHAT DOES THE AVERAGE STRATEGY OF MONTE CARLO CFR ACTUALLY AVERAGE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual regret minimization solves two-player zero-sum games with hidden information. Its guarantee holds for the exact average, which weights each iteration at an information set by the player's own reach probability. However, outcome-sampling solvers update an information set only when a sampled traversal visits it. The original sampling algorithm adds all iterations since the previous visit at the current weight; OpenSpiel adds only the current iteration divided by its sampling probability. We follow OpenSpiel's default sampling: one traversal per player, where only the traversing player explores at a fixed rate. On three of four games, both rules are on average ten to just over fifty percent more exploitable than the exact average after a hundred thousand iterations. After ten million, both are on average two to eight times as exploitable. Since the opponent never explores on the player's traversal, an information set can stay unreached while its weight grows. Adding on the opponent's traversal instead leaves OpenSpiel's rule on average at most two percent worse than the exact average. We prove that an information set's weight between visits is the growth of the exact average's sum at the player's preceding decision, for the action leading there. Reading that growth on the player's traversal at every visit and at the end returns the exact average for at most five percent more training time in our implementation.

1 INTRODUCTION

Counterfactual regret minimization, CFR for short, is the standard method for solving two-player zero-sum games with hidden information (Zinkevich et al., 2007). Its variants are behind the programs that solved heads-up limit poker and beat professionals at no-limit (Bowling et al., 2015; Moravčík et al., 2017; Brown & Sandholm, 2018; 2019b). CFR runs regret matching (Hart & Mas-Colell, 2000) at every decision point, an information set that groups the histories the player cannot tell apart. Its strategies need not converge, but their average approaches an equilibrium. In that average each iteration counts at a set in proportion to the player's own reach: the probability of the player's own earlier actions on the path there, leaving out the opponent's actions and chance. We call it the exact average, because the convergence guarantees of CFR and of its sampled variants are statements about it (Zinkevich et al., 2007; Farina et al., 2020).

Monte Carlo CFR (Lanctot et al., 2009) replaces the pass over the whole game tree by a sampled one. In the outcome-sampling form that OpenSpiel implements (Lanctot et al., 2019), each iteration samples one history, a playthrough of the game, for each player in turn, called that player's traversal. A set is updated only on the iterations whose sampled history passes through it, which we call its visits. A sampled solver therefore cannot add each iteration to the average as it happens. At a visit it has to decide what to add for the iterations since the previous visit, a stretch we call a window, and on whose traversal to add it. The original pseudocode (Lanctot et al., 2009) and OpenSpiel each use one rule, described below, and we call these two the standard rules. By default OpenSpiel lets only the traversing player explore, by mixing uniform play into its strategy with probability 0.6, and we call this one-sided exploration. In this setting we find that both standard rules return strategies 10.8 to 50.1% more exploitable than the exact average of the same run after 10^5 iterations on Leduc poker, Liar's Dice and Goofspiel, the three larger of our four games. On all three the gap grows with training, and after ten million iterations both rules are in the mean two to eight times as exploitable as the exact average. A solver that runs either rule therefore returns a strategy that its convergence guarantee does not describe.

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.

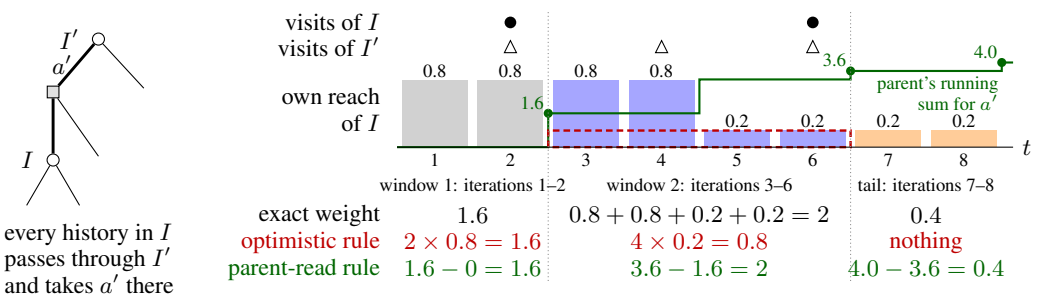


Figure 1: Left: the parent (I', a') of I, the player's most recent earlier decision; circles are the player's decisions, the square a move of the opponent or chance. Right, with illustrative values: dots mark visits of I, triangles visits of I'. Bars: the own reach of I in each iteration, shaded by window and orange in the tail. Green line, on its own scale: the parent's sum for a', updated at the triangles and at the end. Rows, by window and tail: the exact weight, what the optimistic rule adds (dashed rectangle) and what the parent-read rule adds.

One source of the gap is the weight a visit adds. The own reach of a set can change whenever one of the player's earlier decisions is updated, and those updates happen at the visits of the earlier decisions, which need not reach the set. In Figure 1 a set I is visited at iterations 2 and 6, and its own reach is 0.8 until an earlier decision of the player changes after iteration 4, then 0.2. Because a strategy changes only at an update, the one produced by the visit at iteration 2 is in force during iterations 3 to 6. In the exact average that strategy has the weight $0.8 + 0.8 + 0.2 + 0.2 = 2$. The optimistic rule, published with the algorithm, sees only the reach at the visit and gives it $4 \times 0.2 = 0.8$. It also adds nothing after the last visit, a stretch we call the tail. Its authors call it "not technically a correct average" (Lanctot et al., 2009).

A second source is the traversal on which the average is accumulated. On the player's own traversal the opponent is sampled from its current strategy, which can give probability zero to every history of a set for long stretches. Such a set then goes unvisited while its own reach stays positive, and the exact average keeps giving it weight. Stochastic rules such as OpenSpiel's per-history rule add only the iteration of a visit, divided by its sampling probability, and therefore never add those iterations (Theorem 2). On the opponent's traversal the opponent explores instead. Every iteration at which the player reaches the set then has a positive probability of being added, and the same rule is unbiased, up to a constant factor, for the strategies it adds (Theorem 3).

The exact weight is nevertheless available at the visit. Where the player has an earlier decision, perfect recall singles out the most recent one (I', a'), the parent of I. The own reach of I is the own reach of that parent times the probability of a'. Hence the sum that the parent keeps for a' in the exact average is the cumulative own reach of I. It is up to date whenever I is visited, because the sampled history passes through I' first and we update the sums from the root down. The parent-read rule reads this sum at each visit and subtracts the value it read at the previous visit. Their difference is the exact weight of the window ($3.6 - 1.6 = 2$ in Figure 1), and one more read when the average is read out adds the tail.

The parent-read rule pushes no weight down the tree: it reads the weight from a sum the solver keeps anyway, and no sampled path can lose it. The supplement of Lanctot et al. (2009) suggests keeping a weight at each parent information set and pushing it down to the sets below, and it sketches the stochastic update that divides by the probability of reaching the set. The lazy rule of Lanctot (2013) implements the push, and it loses the pushed weight whenever a sampled path ends before the player's next decision. Gibson et al. (2012b) add to the average on the opponent's traversal, where we prove the stochastic rules unbiased for the strategies they add but not exact (Theorem 3). We answer three questions. How far are the standard rules from the exact average? Does the rule or the traversal matter more? Can the exact average be kept for the same computation?

Our contributions are as follows. (1) We prove what the rules return. The optimistic rule averages the right strategies with the wrong weights, although the weights are exact wherever the player has no earlier decision, and it adds nothing for the tail (Theorem 1). On the player's own traversal under one-sided exploration the stochastic rules omit every iteration at which the set cannot be sampled (Theorem 2), and without the tail these rules and the optimistic rule can fail to converge while the exact average of the same run does (Proposition 2). On the opponent's traversal the stochastic sums

are unbiased, up to a constant factor, for the strategies of the exact average, each offset by at most one iteration (Theorem 3). (2) We prove that the parent-read rule returns the exact average on the player's own traversal, for one number and one parent pointer per information set, one read per visit and one pass over the sets when the average is read out (Proposition 1). (3) We measure the rules on the same sampled histories on Kuhn poker, the three larger games and four other instances. On the three larger games under one-sided exploration at $\varepsilon = 0.6$ after 10^5 iterations, the traversal matters more than the choice of standard rule: moving the per-history rule to the opponent's traversal leaves it at most 2% worse than the exact average, while switching between the two standard rules moves the gap by at most 4.8 points. Under CFR+ (Tammelin, 2014) and discounted CFR (Brown & Sandholm, 2019a), which change both the regret update and the weights of the average, the gaps on the player's own traversal shrink but remain the largest (Section 5.3). In our implementation the parent-read rule takes at most 5% more training time than the optimistic rule.

Moving the per-history rule to the opponent's traversal removes most of the gap, and the parent-read rule removes all of it. The first is the smaller change to an existing solver. A single run still lands off the exact average, with a standard deviation over seeds of 0.9 to 3.0 points on the three larger games (Table 2). The parent-read rule returns the exact average itself on every run and under any deterministic iteration weights (Remark 1), and we recommend it wherever the returned strategy should be the one that the convergence guarantee describes.

2 Background

Extensive-form games. We consider two-player zero-sum games with chance in extensive form (von Stengel, 2007). A history h is a sequence of actions from the root of the game tree. At a terminal history z player i receives the payoff $u_i(z)$. Its range across terminal histories has width Δ_u. An information set I of player i is a set of histories at which i moves and that i cannot tell apart, with actions $A(I)$. A behavioral strategy σ_i assigns a distribution $\sigma_i(I, \cdot)$ over $A(I)$ to every such I. Under a profile $\sigma = (\sigma_1, \sigma_2)$, the reach probability of h is the product of the probabilities of the actions on h. It factors into player i's own actions and those of the opponent and chance: $\pi^\sigma(h) = \pi_i^\sigma(h)\,\pi_{-i}^\sigma(h)$. Under perfect recall all histories in an information set share the same sequence of the player's own earlier information sets and actions (von Stengel, 2007). Hence the own reach $\pi_i^\sigma(h)$ is the same for every $h \in I$, and we write it $\pi_i^\sigma(I)$. The sequence-form strategy of σ_i assigns to each sequence of own actions the product of their probabilities (Koller et al., 1996; von Stengel, 1996; 2007).

CFR and the exact average. Counterfactual regret minimization (Zinkevich et al., 2007) runs regret matching (Hart & Mas-Colell, 2000) at every information set. After iteration t, CFR adds to the cumulative regret of each action at I its counterfactual value (its expected payoff weighted by the opponent's and chance's reach of I) minus that of σ_t. It then sets $\sigma_{t+1}(I, \cdot)$ proportional to the positive part of the cumulative regret, or uniform if no part is positive. The iterates $\sigma_1, \dots, \sigma_T$ need not converge, but the reach-weighted average

$$\bar{\sigma}^T(I, a) = \frac{S_I^T(a)}{\sum_{b \in A(I)} S_I^T(b)}, \qquad S_I^T(a) = \sum_{t=1}^{T} \pi_i^{\sigma_t}(I)\,\sigma_t(I, a), \tag{1}$$

which we call the exact average, is a 2η-equilibrium whenever both players' average overall regrets are at most η (Zinkevich et al., 2007). In a 2η-equilibrium neither player gains more than 2η by deviating. We call S_I^T the running sum at I. Lemma 2 in Appendix A shows that (1) realizes the uniform average $S_I^T(a)/T$ of the sequence-form iterates. The experiments report exploitability, half the sum over both players of the gain of an exact best response against the other player's strategy. It is zero at an equilibrium.

Outcome-sampling MCCFR. Monte Carlo CFR (Lanctot et al., 2009) replaces the full traversal of the tree by a sampled one. In outcome sampling each player in turn is the traversing player i, whose regrets are updated. One terminal history z is sampled from a sampling profile σ'. The sampled counterfactual value of an action at a set on z is the payoff of z times the opponent's and chance's probabilities of z and the player's probability of the rest of z after that action. The cumulative regret of that action grows by this value minus that of σ_t, as in CFR. Dividing each such value by the sampling probability $q(z)$ makes the sampled regrets unbiased. The traversing player explores with $\sigma' = \varepsilon\,\mathrm{unif} + (1 - \varepsilon)\sigma$ for an exploration parameter $\varepsilon \in (0, 1]$. Under one-sided exploration, the

convention of OpenSpiel (Lanctot et al., 2019) and of Lanctot (2013), the opponent is sampled from its current strategy. Regret matching gives probability exactly zero to an action with non-positive cumulative regret whenever another action's regret is positive. The opponent's current strategy can therefore give probability zero to every history in a set for stretches of iterations while the own reach of that set, the weight of the exact average, stays positive. Under two-sided exploration both players mix with the uniform distribution. Every terminal history then has sampling probability at least a floor $\delta > 0$ in every iteration (Appendix B gives a valid floor), so the sampled history passes through every information set with probability at least δ. The regret bound of Lanctot et al. (2009) assumes such a floor and carries a factor $1/\delta$. External sampling, the other variant we use, expands all of the traversing player's actions and samples those of the opponent and chance.

3 Averaging between visits

Fix an information set I of player i. Each iteration contains one sampled traversal for each player. Set I is visited at iteration t when the history sampled on player i's traversal passes through I. We write σ_t for a player's strategy before its own update in iteration t. Let $t_1 < \cdots < t_K \le T$ be the visits of I and $t_0 = 0$. We call the iterations $(t_{k-1}, t_k]$ window k and the iterations after t_K the tail. Only visited sets have their regrets updated, so the strategy at I equals $\sigma_{t_k}(I, \cdot)$ throughout window k (Appendix A, Lemma 1). Then (1) becomes a sum over windows,

$$S_I^T = \sum_{k=1}^{K} w_k \, \sigma_{t_k}(I, \cdot) + w_{\text{tail}} \, \sigma_{t_K+1}(I, \cdot), \qquad w_k = \sum_{\tau \in (t_{k-1}, t_k]} \pi_i^{\sigma_\tau}(I), \tag{2}$$

where the tail weight w_{tail} is the same sum over the tail. We write $W_I(t) = \sum_{\tau \le t} \pi_i^{\sigma_\tau}(I)$ for the cumulative own reach of I through iteration t, so that $w_k = W_I(t_k) - W_I(t_{k-1})$. To reproduce (2) with one addition per visit, a rule must therefore add the window weight w_k at visit k; in Figure 1 the weights are $w_1 = 1.6$, $w_2 = 2$ and $w_{\text{tail}} = 0.4$.

This weight is hard to know because it depends on the player's earlier decisions. If player i has an earlier decision on the histories of I, perfect recall singles out the most recent one. We call that pair (I', a') the parent of I. Every $h \in I$ passes through I' and takes a' there (von Stengel, 2007); hence $\pi_i^\sigma(I) = \pi_i^\sigma(I')\,\sigma_i(I', a')$. The depth $d(I)$ is the number of the player's earlier decisions on the path to I. At depth zero $\pi_i^\sigma(I) \equiv 1$. At depth one or more the own reach can change whenever one of these decisions is updated. Those updates happen at visits of the earlier decisions, which need not reach I. Then a visit of I alone does not reveal w_k.

The rules. The rules we compare see the same visits, and Table 1 lists what each adds. In the table σ_{t_K+1} is the strategy in force from the iteration after the last visit onwards, $q_t(h)$ is the probability that iteration t samples $h \in I$ on player i's traversal, and $q_t(I) = \sum_{h \in I} q_t(h)$ is the probability that it visits I. Adding $w_k = W_I(t_k) - W_I(t_{k-1})$ times the strategy, the parent-read rule reproduces (2). In contrast, the optimistic rule multiplies the length of the window by the own reach at the visit and adds nothing for the tail. The optimistic rule plus tail also adds the tail at the reach in force at the end, the one final update that Lanctot (2013) requires when the average is read. Comparing the two separates the error of the optimistic rule into the tail and the weights inside windows.

The stochastic rules. These divide the increment by a sampling probability instead of reconstructing the weight. The per-set rule divides the own reach times the strategy by $q_t(I)$. Because $q_t(I)$ sums over every history of I, a sampled history does not supply it, and we compute it from the game tree. That makes the per-set rule an idealized baseline with an exact visit probability. It is the stochastic update that the supplement of Lanctot et al. (2009) sketches. Before normalization the per-set sum is unbiased for S_I^T whenever $q_t(I) > 0$ for all t. OpenSpiel's outcome-sampling solver uses the per-history rule, which divides by $q_t(h)$ instead and needs only the sampled history. It adds this only on the player's own traversal, as Appendix B quotes from its source. Appendix C.12 relates the per-history rule to the update of Gibson et al. (2012b). The per-history sum is unbiased for $|I|\,S_I^T$ whenever every $q_t(h) > 0$. Here $|I|$ is the number of histories in I, a property of the game and not of the iteration (Appendix B). Dividing the per-history sum by the constant $|I|$ leaves its normalized average unchanged, and that normalized average still carries sampling error.

The lazy rule and edge cases. The lazy rule of Lanctot (2013) targets the same average but hands each pending weight down the sampled path and nowhere else. At a visit it adds

Table 1: What each rule adds to the running sum of I at visit k and at the end of the run. Here $W_I(t) = \sum_{\tau \le t} \pi_i^{\sigma_\tau}(I)$, h is the history sampled at visit k and $w_{\text{tail}} = W_I(T) - W_I(t_K)$. $*$: the standard rules.

Rule	Added at visit k	Added at the end of the run
Parent-read	$\sigma_{t_k}(I,\cdot)\,\big(W_I(t_k) - W_I(t_{k-1})\big)$	$w_{\text{tail}}\,\sigma_{t_K+1}(I,\cdot)$
Optimistic*	$(t_k - t_{k-1})\,\pi_i^{\sigma_{t_k}}(I)\,\sigma_{t_k}(I,\cdot)$	nothing
Optimistic plus tail	as the optimistic rule	$(T - t_K)\,\pi_i^{\sigma_{T+1}}(I)\,\sigma_{t_K+1}(I,\cdot)$
Stochastic, per set	$\pi_i^{\sigma_{t_k}}(I)\,\sigma_{t_k}(I,\cdot)/q_{t_k}(I)$	nothing
Stochastic, per history*	$\pi_i^{\sigma_{t_k}}(I)\,\sigma_{t_k}(I,\cdot)/q_{t_k}(h)$	nothing

$\big(r_k + \pi_i^{\sigma_{t_k}}(I)\big)\,\sigma_{t_k}(I,\cdot)$, where r_k is the weight handed down to it. It leaves the bracketed weight times $\sigma_{t_k}(I,a)$ pending at each action a it did not sample (Algorithm 2 in Appendix C.14). As written it therefore drops the weight of paths that end before the player's next decision. The optimistic rule adds zero but still starts a new window when exploration visits I through an own action that σ gives probability zero. Then the own reach is zero while the sampling probability is not. This loses the reach accumulated earlier in that window. It widens the distance, measured in Appendix C.5, between the optimistic average and the exact one. Every rule returns the uniform initial strategy at a never-visited set, and at any set where its running sum is zero.

Two traversals. A rule can add to player i's running sums on player i's own traversal, as the pseudocode of Lanctot et al. (2009) and OpenSpiel's outcome-sampling solver do, or on the opponent's traversal. A visit of I is then an iteration whose sampled history on the opponent's traversal passes through I. There $q_t(h)$ is the probability of sampling h on that traversal and $q_t(I)$ its sum over I. A rule then adds at the visits on the opponent's traversal, while the strategy at I changes at the visits on player i's own traversal, which fall in between. Equation (2) therefore no longer holds, and even the parent-read rule does not reproduce (1) there. What a stochastic rule keeps on that traversal is unbiasedness for the strategies it adds (Theorem 3).

3.1 The parent-read rule is exact for the cost of a sampled update

Proposition 1 (exact averaging at sampled cost)**.** *Assume perfect recall. Assume that each player's strategy is updated once per iteration after that player's traversal and changes at an information set only at its visits, so that the strategy is constant on each window (Lemma 1). Assume also that the running sum of the parent of a visited set is updated before that of the set, which a root-down traversal gives because the sampled history reaches I' before I. Then the parent-read rule adding on the player's own traversal, with the tail added at the end of the run, returns S_I^T exactly for every I. Each visit costs $O(|A(I)|)$ arithmetic and one read of the running sum of the parent, each read-out adds the tails to a copy of the running sums in $O(\sum_I |A(I)|)$, and the extra state kept between read-outs is one number and one parent pointer per information set.*

Its proof rests on the identity

$$W_I(t) = \sum_{\tau \le t} \pi_i^{\sigma_\tau}(I) = \sum_{\tau \le t} \pi_i^{\sigma_\tau}(I')\,\sigma_\tau(I', a') = S_{I'}^t(a'). \tag{3}$$

In words, the sum that the parent keeps for a' is the cumulative own reach of I. At depth zero there is no parent and $W_I(t) = t$. Since the identity holds with any deterministic iteration weight inside its sums (Remark 1 in Appendix A), the rule also maintains the linear and discounted averages of Brown & Sandholm (2019a). Because the output is the uniform sequence-form average, the high-probability bound of Farina et al. (2020) applies to it under simultaneous updates. Their outcome-sampling estimator draws the opponent from its current strategy, as one-sided exploration does. The bound's constant grows with Δ_u times the largest inverse probability that the sampling strategy gives a sequence of the player's own actions, and mixing in the uniform strategy keeps that finite.

Timed on the three larger games and on the two larger instances of Appendix C.10, the parent-read training loop takes at most 5% longer than the optimistic one. It stays within 3% of the per-history rule adding on the opponent's traversal. On those larger instances it costs 20 to 63% less time than the per-set rule on that traversal, which pays for the exact visit probability.

4 WHAT THE OPTIMISTIC AND STOCHASTIC RULES RETURN

All of the rules we compare keep the strategies of the exact average and change only the weights, possibly to zero: the optimistic rule misweights windows and drops the tail, while the stochastic rules replace each weight by a random estimate. Lemma 3 in Appendix A turns such differences at sets into a difference in exploitability, weighting each set by the own reach there. Theorems 1 and 2 concern the player's own traversal and Theorem 3 concerns the opponent's. All three theorems are proved for outcome sampling.

4.1 WITHOUT A TAIL A RULE ADDING ON THE PLAYER'S OWN TRAVERSAL NEED NOT CONVERGE

The iterations a rule never adds can cost accuracy and even convergence. In the smallest game in which a set can be lost, with constant probability the returned profile never improves while the exact average of the same run approaches equilibrium.

Proposition 2 (without a tail, a rule adding on the player's own traversal need not converge)**.** *Let the opponent choose x or y, where y ends the game at payoff zero, and after x let player i choose at the information set I between L, which pays player i one, and R, which pays $-G$ with $G > 1$. Run outcome-sampling MCCFR with regret matching and one-sided exploration with any $\varepsilon \in (0, 1]$, player i traversing first in every iteration, each player's regrets updated at the end of its own traversal rather than at the end of the iteration. With probability at least $1/4$ the following holds for every $T \geq 1$. The optimistic rule, the per-set rule and the per-history rule, adding on player i's traversal, return the uniform strategy at I, and the profile they return has exploitability at least $(G-1)/4$. The exact average has exploitability $1/(4T)$ once $T \geq (1+G)/2$.*

By Theorem 1(b) below, the optimistic rule plus tail equals the exact average here. What fails is the missing stretch and not the weighting. The set I lies off the equilibrium path: once the opponent's strategy gives x probability zero, a traversal that samples the opponent from its strategy never returns to I, while a best response to the returned profile does. The same state is reached with probability at least $1/8$ if the opponent traverses first.

4.2 THE OPTIMISTIC RULE: THE RIGHT STRATEGIES WITH THE WRONG WEIGHTS

At visit k the optimistic rule gives the strategy $\sigma_{t_k}(I, \cdot)$ the weight $\tilde{w}_k = (t_k - t_{k-1})\, \pi_i^{\sigma_{t_k}}(I)$, the length of the window times the own reach at its last iteration. When $w_k > 0$ we call $\rho_k = \tilde{w}_k / w_k$ the weight ratio of window k and $|\rho_k - 1|$ the weight error. The ratio divides the own reach at the end of the window by its mean over the window. When the window weight is zero so is the reach at the visit. Neither the optimistic rule nor the exact average gives weight to that window. We write $\bar{\sigma}^{T,\mathrm{op}}(I, \cdot)$ for the normalized running sum of the optimistic rule.

Theorem 1 (what the optimistic rule returns)**.** *(a) If the running sum of the optimistic rule at I is not zero, the optimistic average and the exact average at I are convex combinations of the same per-window strategies. The sums below run over the windows with $w_k > 0$, and the exact one also gives weight to the strategy after the last visit:*

$$\bar{\sigma}^{T,\mathrm{op}}(I, \cdot) = \sum_k \frac{\rho_k w_k}{\sum_j \rho_j w_j}\, \sigma_{t_k}(I, \cdot), \qquad \bar{\sigma}^{T}(I, \cdot) = \sum_k \frac{w_k}{W_I(T)}\, \sigma_{t_k}(I, \cdot) + \frac{w_{\mathrm{tail}}}{W_I(T)}\, \sigma_{t_K+1}(I, \cdot).$$

(b) If the own reach of I is constant on every window and through iteration $T+1$, in particular at depth zero, then $\rho_k \equiv 1$, the optimistic rule plus tail equals the exact average, and the optimistic rule alone differs from it by at most $2\, w_{\mathrm{tail}}/W_I(T)$ in ℓ_1.

The own reach of I is the product $\prod_{j \leq d(I)} \sigma(I_j, a_j)$ over the player's earlier decisions (I_j, a_j) on the path to I. Each factor changes only at the visits of its own set I_j, so the optimistic weights can be wrong only where one changes between two visits of I. By Lemma 3 the exploitability of the optimistic averages exceeds that of the exact averages by at most $\frac{\Delta_u}{2T} \sum_I \big(\sum_k |\rho_k - 1|\, w_k + w_{\mathrm{tail}} \big)$, a share of the exact weight that is small when windows are short or the own reach moves little. Yet ρ_k itself can be as large as the window is long (Theorem 5 in Appendix A). Section 5.1 measures the misweighting and the tail separately.

4.3 THE STOCHASTIC RULES: UNBIASED WHERE EVERY REACHED SET CAN BE VISITED, MISSED ITERATIONS ELSEWHERE

Fix I and abbreviate $\pi_t = \pi_i^{\sigma_t}(I)$, $\sigma_t = \sigma_t(I, \cdot)$ and $q_t = q_t(I)$. Let $B_t \in \{0, 1\}$ indicate a visit at iteration t, and write $S(a) = \sum_t \pi_t \sigma_t(a)$ for the running sum of the exact average, $N = \sum_t \pi_t$ for its total and $\nu^2 = \mathbb{E} \sum_t \pi_t^2 (1/q_t - 1)$ for the variance scale. In the per-set stochastic rule, $S(a)$ becomes $\hat{S}(a) = \sum_t B_t \pi_t \sigma_t(a)/q_t$, with total $\hat{N} = \sum_b \hat{S}(b)$ and normalized average $\hat{\sigma} = \hat{S}/\hat{N}$.

Theorem 2 (what the stochastic rules return)**.** *(a) Under two-sided exploration, $q_t > 0$ for all t, $\mathbb{E}[\hat{S}(a) - S(a)] = 0$ and $\mathrm{Var}(\hat{S}(a) - S(a)) = \mathbb{E} \sum_t \pi_t^2 \sigma_t(a)^2 \, (1/q_t - 1)$. For every deterministic N_0 with $N \geq N_0$ almost surely, $\mathbb{E}\big[\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1 \, \mathbb{1}\{\hat{N} \geq N/2\}\big] \leq 4\sqrt{|A(I)|}\, \nu/N_0$ and $\Pr(\hat{N} < N/2) \leq 4\nu^2/N_0^2$; at depth zero $N_0 = T$. (b) Under one-sided exploration, on every iteration with $q_t = 0$ and $\pi_t > 0$ both the per-set and the per-history rule omit the term $\pi_t \sigma_t$ from their running sums, so they are biased for S.*

Part (a) concerns the per-set rule: dividing by the visit probability makes its running sum unbiased. Its normalized average, a self-normalized importance-sampling estimate, keeps a sampling error. The first bound holds when the estimated total $\hat{N}$ is at least half the true one, and the second bounds the probability that it is not. Under two-sided exploration, at a set whose own reach stays above a positive level on a positive fraction of the iterations, that error is at most of order $T^{-1/2}$, the rate of the exact average's exploitability bound (Corollary 1 in Appendix A). That order is attained by the per-set rule (Remark 4). The expected exploitability needs no condition on the own reach, and its excess is of the same order (Corollary 2). Under two-sided exploration the per-history increment divided by $|I|$ has the mean of the per-set increment and at least its variance (Lemma 4 in Appendix A). The own-traversal per-history column of Table 2 is measured in the setting of part (b).

Theorem 3 (what the stochastic rules return on the opponent's traversal)**.** *Assume perfect recall and one-sided exploration with $\varepsilon > 0$, and let a stochastic rule add on the opponent's traversal. Then every iteration at which the own reach is positive has a positive probability of being added, the per-set sum is unbiased for the strategies it adds, and the per-history sum is unbiased for them up to the factor $|I|$, whatever the own reach. Those strategies are the exact average's, each offset by at most one iteration, and against any strategy of the opponent the average of the offset strategies and the exact average differ in payoff by at most Δ_u/T.*

Theorem 4 in Appendix A contains this theorem as its parts (a), (b) and (d) and adds as part (c) error bounds for the normalized averages, which keep a sampling error.

5 EXPERIMENTS

Setup. We implement outcome-sampling and external-sampling MCCFR with regret matching in Python on the game definitions and the exact best-response evaluator of OpenSpiel (version 1.6.11). Within a run every rule reads the same sampled histories and strategies, and we keep the sums for both traversals from those samples; any difference between two rules then comes from the rule and the traversal alone. On all four games our implementation of the per-history rule and OpenSpiel's own outcome-sampling solver reach the same exploitability within the spread over seeds (Appendix B). We use Kuhn poker (Kuhn, 1950) with 12 information sets and three larger games with 936, 1,024 and 3,608: Leduc poker (Southey et al., 2005), Liar's Dice with one four-sided die per player and Goofspiel (Ross, 1971) with four cards and a random point order. Unless stated otherwise, every result below is a mean over thirty seeds after 10^5 iterations of one-sided exploration at $\varepsilon = 0.6$, the value OpenSpiel's solver uses. Appendix C repeats the measurements for seven values of ε from 0.1 to 1, under two-sided exploration and on four other instances with up to 236,450 information sets. Appendix B gives the full setup and the cost tables.

Metric. We compare every rule with the exact average of the same run, which we take from the parent-read rule on the player's own traversal and check against a brute-force sum of (1). We report the relative gap of a rule in per cent: the mean over seeds of its exploitability divided by that of the exact average of the same run, minus one. Appendix C.1 gives the same cells in exploitability. A positive gap means that the rule is more exploitable. A negative gap, which can occur at a finite horizon, means that it is less. We give the mean and sample standard deviation over seeds in the tables, and 95% Student-t intervals of paired per-seed differences in the text.

Table 2: Relative gap (%) to the exact average of the same run; positive is more exploitable. One-sided exploration, $\varepsilon = 0.6$, 10^5 iterations, mean $\pm$ sample s.d. over thirty seeds. The lazy rule adds on both traversals. $*$: the standard rules, those of the published pseudocode and of OpenSpiel.

Game	own traversal			both	opponent's traversal		exact average (absolute expl.)
	optimistic*	optimistic + tail	per-history*	lazy (Lanctot)	optimistic + tail	per-history	
Kuhn	$+0.9_{\pm 4.2}$	$+0.1_{\pm 0.3}$	$+11.8_{\pm 27.0}$	$+9.9_{\pm 25.0}$	$-0.0_{\pm 0.9}$	$-0.2_{\pm 9.1}$	$0.007_{\pm 0.002}$
Leduc	$+13.2_{\pm 7.2}$	$+4.3_{\pm 3.0}$	$+10.8_{\pm 8.8}$	$-0.6_{\pm 3.1}$	$+2.8_{\pm 1.7}$	$-0.0_{\pm 0.9}$	$0.467_{\pm 0.033}$
Liar's Dice	$+16.4_{\pm 10.5}$	$+9.9_{\pm 5.8}$	$+18.0_{\pm 14.0}$	$+2.7_{\pm 7.7}$	$+3.6_{\pm 3.0}$	$+0.1_{\pm 3.0}$	$0.098_{\pm 0.011}$
Goofspiel	$+45.3_{\pm 14.0}$	$+19.6_{\pm 8.7}$	$+50.1_{\pm 13.8}$	$+2.2_{\pm 4.7}$	$+5.6_{\pm 3.6}$	$-1.9_{\pm 3.0}$	$0.098_{\pm 0.008}$

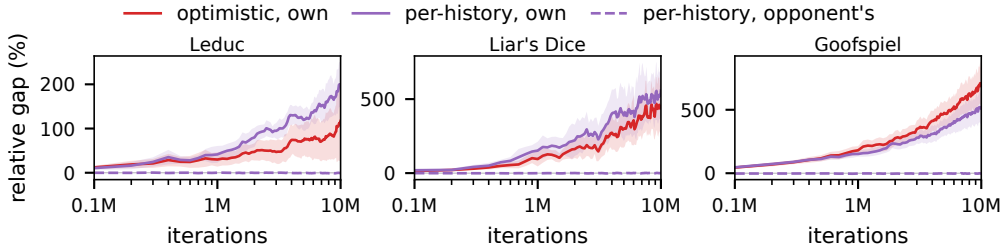


Figure 2: The relative gap of the standard rules grows with training; that of the per-history rule on the opponent's traversal stays near zero. Gaps in per cent, $\varepsilon = 0.6$, one-sided exploration; mean over twelve seeds, shaded 95% interval; the vertical scales differ.

5.1 On the three larger games the standard rules are more exploitable than the exact average, and the gap grows with training

Both standard rules are 10.8 to 50.1% more exploitable than the exact average on the three larger games (Table 2, starred columns), and the optimistic rule is above it on all but one of those ninety seeds. Kuhn poker is the exception for the optimistic rule: nine of its twelve sets have no earlier decision of the player, and there the weights are exact and only the tail is missing (Theorem 1(b)). The lazy rule of Lanctot (2013) ends within 3% of the exact average on the three larger games, but rerun confined to the player's own traversal its gap grows there by 13.0 to 51.9 points (Appendix C.16).

At its exact weight $W_I(T) - W_I(t_K)$ read from the parent, the tail removes 56 to 82% of the optimistic rule's mean gap on the three larger games (Appendix C.5), and the rest is the misweighting of the windows (Theorem 1). On Liar's Dice with six-sided dice the exact-weight tail again removes over half of the gap, but on fixed-order Goofspiel with five and with six cards under a third (Appendix C.8).

Training longer widens the standard rules' gaps (Figure 2). On each of the three larger games, twelve seeds carried to 10^7 iterations end with the standard rules in the mean two to eight times as exploitable as the exact average, while the per-history rule on the opponent's traversal ends within 3.2% of it (Appendix C.4). To reach the exact average's exploitability at 10^6 iterations, both standard rules on Leduc poker and the optimistic rule on Liar's Dice need a median of 1.44 to 3.25 times the exact average's work. The per-history rule on Liar's Dice and both standard rules on Goofspiel miss that exploitability within 10^7 iterations in half or more of the runs. On all three games the per-history rule on the opponent's traversal needs a median of 0.95 to 0.99 times the work (Appendix C.15).

5.2 On the opponent's traversal the missing iterations come back

In three runs of 2×10^4 iterations, the iterations that sampling cannot reach on the player's own traversal account for 20 to 50% of the player's own reach on the three larger games and under 10% on Kuhn poker (Appendix C.13). On the opponent's traversal every iteration at which the player reaches a set has a positive probability of being added (Theorem 3). There the per-history rule is within 2% of the exact average on all four games (Table 2). On the three larger games, moving either standard rule to the opponent's traversal changes its gap by at least 10 points, while the two rules differ from each other by at most 4.8 points. Paired by seed, the change for the per-history rule exceeds the per-seed difference between the rules by 7 to 40 points in the mean, and every

95% interval excludes zero (Appendix C.12). On Goofspiel the per-history rule on the opponent's traversal ends 1.9% below the exact average and is below it on twenty-four of thirty seeds.

Two-sided exploration also closes the gaps, and no mean gap on the three larger games exceeds 2.9% for $\varepsilon \geq 0.5$. Its price on those games is that the exact average itself is 1.4 to 3.3 times as exploitable for $\varepsilon \geq 0.6$ as under one-sided exploration at equal iterations (Appendices C.2 and C.3).

5.3 The largest gaps stay on the player's own traversal at scale, under external sampling and under CFR+

On the largest instance, Goofspiel with five cards and a random point order, the optimistic and per-history rules on the player's own traversal end 10.9 and 14.9% above the exact average at 10^6 iterations and are above it on every seed. On the opponent's traversal the optimistic rule plus tail and both stochastic rules are at most 1.6% above it (Appendix C.7).

Under external sampling OpenSpiel adds on the opponent's traversal at weight one, and on the three larger games its rule is up to 7.9% below the exact average. In contrast, the optimistic and per-history rules adding at the traversing player's own sets are 150 to 284% above it (Appendix C.9).

CFR+ and discounted CFR (Tammelin, 2014; Brown & Sandholm, 2019a) change the regret update and the weights of the average, and the parent-read rule stays exact under any deterministic weights (Remark 1 in Appendix A). Under both, the largest own-traversal gap falls from 50.1% to 13.5 and 10.6%, while on the opponent's traversal the optimistic rule plus tail and both stochastic rules stay within 0.6% (Appendix C.6).

6 Related work

Sampled CFR and its average. Other sampled forms of CFR (Johanson et al., 2012; Schmid et al., 2019; Steinberger et al., 2020; McAleer et al., 2023) change how the tree is sampled or how the sampled values are estimated. Gibson et al. (2012a) and Li et al. (2020) change which of the player's own actions are sampled, and so which sets a traversal reaches. Lanctot (2013) calls the optimistic rule incorrect and the lazy rule a correct way. Gibson et al. (2012b) add to the average during the opponent's traversal so that every information set with positive weight can be reached. Deep CFR (Brown et al., 2019) fills its strategy memory at the same place, weighting each sample by its iteration and not by the own reach.

Guarantees for averages. Farina et al. (2020) bound the saddle-point gap of the uniform average of the sequence-form iterates (the exact average, Lemma 2) with high probability, for any deterministic regret minimizer fed by unbiased gradient estimates. CFR+ (Tammelin, 2014), predictive CFR+ (Farina et al., 2021) and discounted CFR (Brown & Sandholm, 2019a) change the regret update and the weights of the average but still return a weighted average of the iterates; the parent-read rule maintains that average.

7 Conclusion

CFR's guarantee is about the exact average, but the two standard rules of outcome-sampling MCCFR average the same per-iteration strategies with different weights (Section 4). Under one-sided exploration at $\varepsilon = 0.6$ the returned strategy is 10.8 to 50.1% more exploitable than the exact average on the three larger games after 10^5 iterations, and the gap grows with training (Section 5.1). Most of that gap comes from where the average is accumulated: moving the per-history rule to the opponent's traversal leaves that rule at most 2% worse than the exact average (Section 5.2). The parent-read rule returns the exact average on the player's own traversal for one number and one pointer per information set and at most 5% more training time in our implementation. We recommend it wherever a solver's output should be the strategy that the guarantee describes. A solver that keeps the per-history rule should at least accumulate it on the opponent's traversal. The measurements cover eight tabular two-player zero-sum instances, and we plan to take the parent-read rule to networks that fit the average next.

REPRODUCIBILITY STATEMENT

The code, the per-run result files and the scripts that generate every table and figure are in the supplementary material. Every run is seeded. Re-running the first five seeds of the one-sided grid reproduced all 700 recorded exploitabilities bit for bit. Appendices B and C give the OpenSpiel game strings, the payoff ranges of the four main games, the iteration counts and evaluation schedules of every experiment including the external-sampling protocol, the two exploration settings, the seeds, the checkpoint conventions and the controls.

USE OF LARGE LANGUAGE MODELS

The authors take responsibility for the research questions, the theorems and their proofs, the experimental design and the interpretation of the results. Large language models assisted with code and writing: they drafted the experiment drivers and the scripts that produce the tables and figures, drafted and edited prose, and searched the proofs for errors. The authors checked every such output before use.

REFERENCES

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.

Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. doi: 10.1126/science.aao1733.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1829–1836, 2019a. doi: 10.1609/aaai.v33i01.33011829.

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

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 793–802, 2019.

Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting CFR+ and alternating updates. *Journal of Artificial Intelligence Research*, 64:429–443, 2019. doi: 10.1613/jair.1.11370.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Stochastic regret minimization in extensive-form games. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3018–3028. PMLR, 2020. arXiv:2002.08493.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive Blackwell approachability: Connecting regret matching and mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5363–5371, 2021. doi: 10.1609/aaai.v35i6.16676.

Richard Gibson, Neil Burch, Marc Lanctot, and Duane Szafron. Efficient Monte Carlo counterfactual regret minimization in games with many player actions. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1880–1888, 2012a.

Richard Gibson, Marc Lanctot, Neil Burch, Duane Szafron, and Michael Bowling. Generalized sampling and variance in counterfactual regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 1355–1361, 2012b. doi: 10.1609/aaai.v26i1.8241.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Michael Johanson, Nolan Bard, Marc Lanctot, Richard Gibson, and Michael Bowling. Efficient Nash equilibrium approximation through Monte Carlo counterfactual regret minimization. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pp. 837–846, 2012.

Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and Economic Behavior*, 14(2):247–259, 1996. doi: 10.1006/game.1996.0051.

Harold W. Kuhn. A simplified two-person poker. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume I*, number 24 in Annals of Mathematics Studies, pp. 97–103. Princeton University Press, 1950.

Marc Lanctot. *Monte Carlo Sampling and Regret Minimization for Equilibrium Computation and Decision-Making in Large Extensive Form Games*. PhD thesis, University of Alberta, Edmonton, Alberta, 2013.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte Carlo sampling for regret minimization in extensive games. In *Advances in Neural Information Processing Systems 22*, 2009.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019. version of 2020-09-26.

Hui Li, Kailiang Hu, Shaohua Zhang, Yuan Qi, and Le Song. Double neural counterfactual regret minimization. In *International Conference on Learning Representations*, 2020.

Stephen McAleer, Gabriele Farina, Marc Lanctot, and Tuomas Sandholm. ESCHER: Eschewing importance sampling in games by computing a history value function to estimate regret. In *International Conference on Learning Representations*, 2023.

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.

Sheldon M. Ross. Goofspiel—the game of pure strategy. *Journal of Applied Probability*, 8(3): 621–625, 1971. doi: 10.1017/S0021900200035725.

Martin Schmid, Neil Burch, Marc Lanctot, Matej Moravcik, Rudolf Kadlec, and Michael Bowling. Variance reduction in Monte Carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2157–2164, 2019. doi: 10.1609/aaai.v33i01.33012157.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, 2005.

Eric Steinberger, Adam Lerer, and Noam Brown. DREAM: Deep regret minimization with advantage baselines and model-free learning. *arXiv preprint arXiv:2006.10410*, 2020.

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

Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996. doi: 10.1006/game.1996.0050.

Bernhard von Stengel. Equilibrium computation for two-player games in strategic and extensive form. In Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani (eds.), *Algorithmic Game Theory*, chapter 3, pp. 53–78. Cambridge University Press, 2007.

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

A PROOFS

Table 3: What is proved for which setting, and where it is measured.

Statement	Proved for	Measured in
Parent-read rule returns the exact average (Proposition 1)	any game; own traversal; one update per iteration; the parent's sum written first	Section 3.1, checked in Section 5
Optimistic rule: right strategies, wrong weights, no tail (Theorem 1)	any game; own traversal; ratio one where the own reach is constant	Section 5.1
Optimistic rule: bounded distance to the exact average; without the iterations after the last visit, no convergence (Theorem 5, Proposition 2)	any run, own traversal; a two-set game under one-sided exploration, constant probability	Appendix C.5; the counterexample is a construction
Stochastic rules: unbiased with bounded error, biased otherwise (Theorem 2)	(a) two-sided exploration; (b) one-sided, on the player's own traversal	Sections 5.1 and 5.2
Per-set rule: rate in the horizon, and expected exploitability (Corollaries 1 and 2)	two-sided exploration; the rate also needs the own reach above a level often enough	Appendix C.11, Section 5.2
Stochastic rules adding on the opponent's traversal: unbiased for the iterates added, expected exploitability within $O(T^{-1/2})$ of the exact average (Theorem 3, Corollary 3)	one-sided exploration, two-sided by Theorem 2(a); no condition on the own reach; the player updated first targets the iterates one update later, at most $\Delta_u/(2T)$	Section 5.2, Appendix C.12

This appendix proves the statements of Sections 3 and 4. Table 3 lists what is proved for which setting and where it is measured. Throughout, I is an information set of player i with visit iterations $t_1 < \cdots < t_K \le T$ and $t_0 = 0$. A never-visited information set has $K = 0$, its tail is the whole run, and σ_{t_K+1} is the initial strategy. The window weights w_k, the tail weight w_{tail} and the cumulative own reach $W_I(t)$ are those of Section 3.

Remark 1 (weighted iterations)**.** *Every statement of Sections 3 and 4 is written for an average that gives each iteration weight one. A solver that gives iteration τ the weight $\alpha_\tau > 0$, as CFR+ does with $\alpha_\tau = \tau$ and discounted CFR with $\alpha_\tau = \tau^\gamma$ (Tammelin, 2014; Brown & Sandholm, 2019a), targets the average with $w_k = \sum_{\tau\in(t_{k-1},t_k]} \alpha_\tau\, \pi_i^{\sigma_\tau}(I)$ and $W_I(t) = \sum_{\tau\le t} \alpha_\tau\, \pi_i^{\sigma_\tau}(I)$ in place of the sums of Section 3. With those two substitutions the window decomposition, Theorem 1 and the exactness of the parent-read rule carry over unchanged, since they use only that $w_k = W_I(t_k) - W_I(t_{k-1})$, that $\sigma_\tau(I,\cdot)$ is constant between visits and that W_I factors through the parent. At depth zero $W_I(t) = \sum_{\tau\le t} \alpha_\tau$ rather than t. The optimistic rule then adds $\big(\sum_{\tau\in(t_{k-1},t_k]} \alpha_\tau\big)\pi_i^{\sigma_{t_k}}(I)$ at visit k, the stochastic rules add α_{t_k} times their increment, and the parent-read rule is unchanged. The variance and offset statements on the opponent's traversal change as Remark 3 gives.*

Lemma 1 (piecewise constancy)**.** *$\sigma_\tau(I,\cdot) = \sigma_{t_k}(I,\cdot)$ for all $\tau \in (t_{k-1}, t_k]$, and $\sigma_\tau(I,\cdot) = \sigma_{t_K+1}(I,\cdot)$ for $\tau \in (t_K, T]$.*

Proof. Let $R_\tau(I,\cdot)$ denote the cumulative regret at I from which the strategy of iteration τ is computed. The regret at I changes only at visits of I, so for every $\tau \in (t_{k-1}, t_k]$

$$\begin{aligned} R_\tau(I,\cdot) &= R_{t_{k-1}+1}(I,\cdot) && \text{(no visit of } I \text{ in } (t_{k-1},\tau)) \\ &= R_{t_k}(I,\cdot), && \text{(the same, for } \tau = t_k) \end{aligned}$$

and therefore

$$\begin{aligned} \sigma_\tau(I,\cdot) &= \mathrm{RM}\big(R_\tau(I,\cdot)\big) && \text{(regret matching)} \\ &= \mathrm{RM}\big(R_{t_k}(I,\cdot)\big) = \sigma_{t_k}(I,\cdot). \end{aligned}$$

The same two lines with t_K in place of t_{k-1} and T in place of t_k give $\sigma_\tau(I,\cdot) = \sigma_{t_K+1}(I,\cdot)$ for $\tau \in (t_K, T]$. □

Grouping the sum in (1) by windows and applying Lemma 1 gives the window form (2):

$$\begin{aligned} S_I^T &= \sum_{k=1}^{K} \sum_{\tau \in (t_{k-1}, t_k]} \pi_i^{\sigma_\tau}(I)\, \sigma_\tau(I, \cdot) + \sum_{\tau \in (t_K, T]} \pi_i^{\sigma_\tau}(I)\, \sigma_\tau(I, \cdot) \\ &= \sum_{k=1}^{K} \sigma_{t_k}(I, \cdot) \sum_{\tau \in (t_{k-1}, t_k]} \pi_i^{\sigma_\tau}(I) + \sigma_{t_K+1}(I, \cdot) \sum_{\tau \in (t_K, T]} \pi_i^{\sigma_\tau}(I) \qquad \text{(Lemma 1)} \\ &= \sum_{k=1}^{K} w_k\, \sigma_{t_k}(I, \cdot) + w_{\text{tail}}\, \sigma_{t_K+1}(I, \cdot). \end{aligned}$$

Lemma 2 (sequence-form average)**.** *Let $x_t(I, a) = \pi_i^{\sigma_t}(I)\, \sigma_t(I, a)$ be the sequence-form strategy of σ_t and $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$ its uniform average. Then $\bar{x}$ is a sequence-form strategy, and the behavioral strategy that realizes it is* (1)*.*

Proof. Write $\text{par}(I)$ for the parent sequence of I, the sequence of own actions leading to I; it is unique by perfect recall. Its sequence-form weight is $x_t(\text{par}(I)) = \pi_i^{\sigma_t}(I)$, with $x_t(\emptyset) = 1$ for the empty sequence. A sequence-form strategy of player i assigns to each own sequence the product of the behavioral probabilities along it. Hence x_t satisfies the root constraint $x_t(\emptyset) = 1$ and at every I the flow constraint

$$\sum_{a \in A(I)} x_t(I, a) = \pi_i^{\sigma_t}(I) \sum_{a \in A(I)} \sigma_t(I, a) = \pi_i^{\sigma_t}(I) = x_t(\text{par}(I)).$$

Both constraints are linear in x_t, so they hold for the average:

$$\bar{x}(\emptyset) = \frac{1}{T} \sum_{t=1}^{T} x_t(\emptyset) = 1,$$

$$\sum_{a \in A(I)} \bar{x}(I, a) = \frac{1}{T} \sum_{t=1}^{T} \sum_{a \in A(I)} x_t(I, a) = \frac{1}{T} \sum_{t=1}^{T} x_t(\text{par}(I)) = \bar{x}(\text{par}(I)).$$

Hence $\bar{x}$ is a sequence-form strategy. At every I with $x(\text{par}(I)) > 0$ the behavioral strategy realizing x divides the weight of each continuation by the weight of the parent sequence. For $x = \bar{x}$ this gives

$$\begin{aligned} \sigma(I, a) &= \frac{\bar{x}(I, a)}{\bar{x}(\text{par}(I))} = \frac{\bar{x}(I, a)}{\sum_{b \in A(I)} \bar{x}(I, b)} \qquad \text{(flow constraint)} \\ &= \frac{\frac{1}{T} \sum_{t=1}^{T} \pi_i^{\sigma_t}(I)\, \sigma_t(I, a)}{\frac{1}{T} \sum_{t=1}^{T} \sum_{b \in A(I)} \pi_i^{\sigma_t}(I)\, \sigma_t(I, b)} = \frac{S_I^T(a)}{\sum_{b \in A(I)} S_I^T(b)}, \qquad \text{(by (1))} \end{aligned}$$

which is (1). Information sets with $\sum_b \bar{x}(I, b) = 0$ are never reached under $\bar{x}$ and receive the uniform strategy by convention. □

Restatement (Proposition 1, exact averaging at sampled cost)**.** *Assume perfect recall. Assume that each player's strategy is updated once per iteration after that player's traversal and changes at an information set only at its visits, so that the strategy is constant on each window (Lemma 1). Assume also that the running sum of the parent of a visited set is updated before that of the set, which a root-down traversal gives because the sampled history reaches I' before I. Then the parent-read rule adding on the player's own traversal, with the tail added at the end of the run, returns S_I^T exactly for every I. Each visit costs $O(|A(I)|)$ arithmetic and one read of the running sum of the parent, each read-out adds the tails to a copy of the running sums in $O(\sum_I |A(I)|)$, and the extra state kept between read-outs is one number and one parent pointer per information set.*

Proof. We first identify $W_I(t)$. At depth zero no own decision precedes I, so $\pi_i^{\sigma_\tau}(I) = 1$ for every τ and

$$W_I(t) = \sum_{\tau \le t} \pi_i^{\sigma_\tau}(I) = \sum_{\tau \le t} 1 = t.$$

Otherwise let (I', a') be the parent of I; then

$$\begin{aligned} W_I(t) &= \sum_{\tau \le t} \pi_i^{\sigma_\tau}(I) \\ &= \sum_{\tau \le t} \pi_i^{\sigma_\tau}(I')\, \sigma_\tau(I', a') && \text{(perfect recall, Section 3)} \\ &= S_{I'}^t(a'). && \text{(by (1))} \end{aligned} \tag{4}$$

We now show by induction on $d(I)$, along the forward traversal of the sampled history, that the running sum of I after its k-th visit holds $\sum_{j \le k} w_j\, \sigma_{t_j}(I, \cdot)$. At depth zero the rule reads $W_I(t_k) = t_k$, which is exact since $W_I(t) = t$. At depth $d > 0$ the parent I' has depth $d - 1$ and lies on the same sampled history before I, so its increment is added before that of I at this visit. The parent is itself visited at t_k, so its window closes there and its running sum has no pending tail. By the induction hypothesis its entry for a' equals $S_{I'}^{t_k}(a')$, which is $W_I(t_k)$ by (4). The increment at visit k is therefore

$$\begin{aligned} \sigma_{t_k}(I, \cdot)\,\big(W_I(t_k) - W_I(t_{k-1})\big) &= \sigma_{t_k}(I, \cdot)\Big(\sum_{\tau \le t_k} \pi_i^{\sigma_\tau}(I) - \sum_{\tau \le t_{k-1}} \pi_i^{\sigma_\tau}(I)\Big) \\ &= \sigma_{t_k}(I, \cdot) \sum_{\tau \in (t_{k-1}, t_k]} \pi_i^{\sigma_\tau}(I) = w_k\, \sigma_{t_k}(I, \cdot), \quad \text{(definition of } w_k\text{)} \end{aligned}$$

and after K visits the running sum holds $\sum_{k=1}^{K} w_k\, \sigma_{t_k}(I, \cdot)$. If a sampling scheme visits I more than once in the same iteration, every increment after the first is

$$\sigma_{t_k}(I, \cdot)\,\big(W_I(t_k) - W_I(t_k)\big) = 0,$$

because strategies are not recomputed within the iteration. At the end of the run the tail increment is

$$\sigma_{t_K+1}(I, \cdot)\,\big(W_I(T) - W_I(t_K)\big) = \sigma_{t_K+1}(I, \cdot) \sum_{\tau \in (t_K, T]} \pi_i^{\sigma_\tau}(I) = w_{\text{tail}}\, \sigma_{t_K+1}(I, \cdot),$$

where $W_I(T)$ is read recursively and top-down from the final value of the parent. Combining these results,

$$\text{running sum of } I \text{ after the tail increment} = \sum_{k=1}^{K} w_k\, \sigma_{t_k}(I, \cdot) + w_{\text{tail}}\, \sigma_{t_K+1}(I, \cdot) = S_I^T \quad \text{(by (2))}$$

for every I. Each visit reads the running sum of the parent once and does $O(|A(I)|)$ arithmetic. The memoized final recursion visits each information set once and does $O(|A(I)|)$ arithmetic there, on a copy of the running sums when the run continues afterwards (Algorithm 1), giving the stated cost. The same argument applies to $W_I^\alpha(t) = \sum_{\tau \le t} \alpha_\tau\, \pi_i^{\sigma_\tau}(I)$ for deterministic weights $\alpha_\tau > 0$. The α-weighted running sum of the parent then replaces its plain one, since (4) becomes $W_I^\alpha(t) = \sum_{\tau \le t} \alpha_\tau\, \pi_i^{\sigma_\tau}(I')\, \sigma_\tau(I', a')$. □

Lemma 3 (from per-set differences to exploitability)**.** *For any two profiles σ and $\tilde{\sigma}$, the exploitability of $\tilde{\sigma}$ exceeds that of σ by at most $\frac{1}{4}\,\Delta_u \sum_I \pi_i^\sigma(I)\, \|\tilde{\sigma}(I, \cdot) - \sigma(I, \cdot)\|_1$, where the sum runs over the information sets of both players and $\pi_i^\sigma(I)$ is the own reach of the player who acts at I.*

Proof. We first show that for any two strategies $\sigma, \tilde{\sigma}$ of player i and any opponent strategy, $|u_i(\sigma, \sigma_{-i}) - u_i(\tilde{\sigma}, \sigma_{-i})| \le \frac{1}{2}\Delta_u \sum_I \pi_i^\sigma(I)\, \|\sigma(I, \cdot) - \tilde{\sigma}(I, \cdot)\|_1$. Fix the opponent strategy σ_{-i} and order the information sets $I_1, \dots, I_M$ of player i bottom-up, descendants before ancestors. Let $\sigma^{(m)}$ be the hybrid strategy that plays $\tilde{\sigma}$ at $I_1, \dots, I_m$ and σ elsewhere, so that $\sigma^{(0)} = \sigma$ and $\sigma^{(M)} = \tilde{\sigma}$. Write $V_m(a)$ for the value to player i of taking action a at I_m and continuing with $\sigma^{(m)}$, weighted by the opponent and chance reach of I_m:

$$V_m(a) = \sum_{h \in I_m} \pi_{-i}^\sigma(h) \sum_{z \sqsupseteq ha} \pi^{\sigma^{(m)}}(ha, z)\, u_i(z).$$

Every ancestor of I_m still plays σ at step m, so the own reach of I_m under both $\sigma^{(m-1)}$ and $\sigma^{(m)}$ is $\pi_i^\sigma(I_m)$. The two hybrids differ only at I_m, so

$$u_i(\sigma^{(m)}, \sigma_{-i}) - u_i(\sigma^{(m-1)}, \sigma_{-i}) = \pi_i^\sigma(I_m) \sum_{a \in A(I_m)} \big(\tilde{\sigma}(I_m, a) - \sigma(I_m, a)\big)\, V_m(a).$$

The histories of I_m are prefix-free and share the same own sequence, so $\sum_{h \in I_m} \pi_{-i}^\sigma(h) \le 1$. Every $V_m(a)$ therefore lies in an interval $[v_m, v_m + \Delta_u]$. With $\bar{v}_m = v_m + \Delta_u/2$ and $\sum_a(\tilde{\sigma}(I_m, a) - \sigma(I_m, a)) = 0$,

$$\begin{aligned}\Big|\sum_a \big(\tilde{\sigma}(I_m, a) - \sigma(I_m, a)\big) V_m(a)\Big| &= \Big|\sum_a \big(\tilde{\sigma}(I_m, a) - \sigma(I_m, a)\big)\big(V_m(a) - \bar{v}_m\big)\Big| \\ &\le \frac{\Delta_u}{2}\, \|\tilde{\sigma}(I_m) - \sigma(I_m)\|_1.\end{aligned}$$

Summing the telescoping differences,

$$\begin{aligned}|u_i(\tilde{\sigma}, \sigma_{-i}) - u_i(\sigma, \sigma_{-i})| &= \Big|\sum_{m=1}^{M} \big(u_i(\sigma^{(m)}, \sigma_{-i}) - u_i(\sigma^{(m-1)}, \sigma_{-i})\big)\Big| \\ &\le \sum_{m=1}^{M} \pi_i^\sigma(I_m)\, \frac{\Delta_u}{2}\, \|\tilde{\sigma}(I_m) - \sigma(I_m)\|_1 \\ &= \frac{\Delta_u}{2} \sum_I \pi_i^\sigma(I)\, \|\sigma(I) - \tilde{\sigma}(I)\|_1.\end{aligned}$$

We now pass from payoffs to exploitability. The best-response gain of player $-i$ against a strategy of player i is a maximum over σ_{-i} of $u_{-i}(\cdot, \sigma_{-i}) = -u_i(\cdot, \sigma_{-i})$. A maximum over σ_{-i} changes by at most the largest pointwise change, so

$$\begin{aligned}\max_{\sigma_{-i}} u_{-i}(\tilde{\sigma}, \sigma_{-i}) - \max_{\sigma_{-i}} u_{-i}(\sigma, \sigma_{-i}) &\le \max_{\sigma_{-i}} \big|u_i(\tilde{\sigma}, \sigma_{-i}) - u_i(\sigma, \sigma_{-i})\big| \\ &\le \frac{\Delta_u}{2} \sum_{I \text{ of player } i} \pi_i^\sigma(I)\, \|\tilde{\sigma}(I) - \sigma(I)\|_1.\end{aligned}$$

The same bound with the roles of the players exchanged controls the gain of player i by a sum over the information sets of player $-i$. Adding the two covers every information set exactly once and bounds the increase of the sum of the two best-response gains by $\frac{1}{2}\Delta_u \sum_I \pi_i^\sigma(I)\, \|\tilde{\sigma}(I, \cdot) - \sigma(I, \cdot)\|_1$. That sum is twice the exploitability, so the stated bound follows. □

Restatement (Theorem 1, what the optimistic rule returns)**.** *(a) If the running sum of the optimistic rule at I is not zero, the optimistic average and the exact average at I are convex combinations of the same per-window strategies. The sums below run over the windows with $w_k > 0$, and the exact one also gives weight to the strategy after the last visit:*

$$\bar{\sigma}^{T,\mathrm{op}}(I, \cdot) = \sum_k \frac{\rho_k w_k}{\sum_j \rho_j w_j}\, \sigma_{t_k}(I, \cdot), \qquad \bar{\sigma}^{T}(I, \cdot) = \sum_k \frac{w_k}{W_I(T)}\, \sigma_{t_k}(I, \cdot) + \frac{w_{\mathrm{tail}}}{W_I(T)}\, \sigma_{t_K+1}(I, \cdot).$$

(b) If the own reach of I is constant on every window and through iteration $T+1$, in particular at depth zero, then $\rho_k \equiv 1$, the optimistic rule plus tail equals the exact average, and the optimistic rule alone differs from it by at most $2\, w_{\mathrm{tail}}/W_I(T)$ in ℓ_1.

Proof. (a) By the update rule the running sum of the optimistic rule after K visits is

$$\begin{aligned}\sum_{k=1}^{K} (t_k - t_{k-1})\, \pi_i^{\sigma_{t_k}}(I)\, \sigma_{t_k}(I, \cdot) &= \sum_{k=1}^{K} \tilde{w}_k\, \sigma_{t_k}(I, \cdot) && \text{(definition of } \tilde{w}_k) \\ &= \sum_{k=1}^{K} \rho_k\, w_k\, \sigma_{t_k}(I, \cdot), && \text{(definition of } \rho_k)\end{aligned}$$

and dividing by $\sum_j \rho_j w_j$ gives the first formula. The second formula is (2) divided by

$$W_I(T) = \sum_{\tau \le T} \pi_i^{\sigma_\tau}(I) = \sum_{k=1}^{K} w_k + w_{\text{tail}}.$$

(b) If the own reach equals a constant c_k on the window $(t_{k-1}, t_k]$, then by the definition of ρ_k

$$\rho_k = \frac{\pi_i^{\sigma_{t_k}}(I)}{\frac{1}{t_k - t_{k-1}} \sum_{\tau \in (t_{k-1}, t_k]} \pi_i^{\sigma_\tau}(I)} = \frac{c_k}{\frac{1}{t_k - t_{k-1}} (t_k - t_{k-1})\, c_k} = 1;$$

at depth zero the own reach is identically one, so this holds on every window. With $\rho_k \equiv 1$ the optimistic weights are $\tilde{w}_k = w_k$. If the reach is also constant on the tail then the tail increment equals

$$(T - t_K)\, \pi_i^{\sigma_{T+1}}(I)\, \sigma_{t_K+1}(I, \cdot) = \Big(\sum_{\tau \in (t_K, T]} \pi_i^{\sigma_\tau}(I) \Big) \sigma_{t_K+1}(I, \cdot) = w_{\text{tail}}\, \sigma_{t_K+1}(I, \cdot),$$

so the optimistic rule plus tail and the exact average are the same convex combination. Without the tail increment the two weight vectors are $\tilde{p}_k = w_k / \sum_j w_j$ on the visit points and $p_k = w_k / W_I(T)$ on the visit points together with $p_{\text{tail}} = w_{\text{tail}} / W_I(T)$ on the tail point. Their total variation distance TV is

$$\begin{aligned} \mathrm{TV}(\tilde{p}, p) &= \frac{1}{2} \sum_{k=1}^{K} w_k \Big(\frac{1}{\sum_j w_j} - \frac{1}{W_I(T)} \Big) + \frac{1}{2} \frac{w_{\text{tail}}}{W_I(T)} \\ &= \frac{1}{2} \Big(1 - \frac{\sum_j w_j}{W_I(T)} \Big) + \frac{1}{2} \frac{w_{\text{tail}}}{W_I(T)} = \frac{w_{\text{tail}}}{W_I(T)}, \qquad (\text{since } W_I(T) = \sum_j w_j + w_{\text{tail}}) \end{aligned}$$

and the ℓ_1 bound $2\, w_{\text{tail}} / W_I(T)$ follows from the inequality below. For two convex combinations $\sum_k p_k x_k$ and $\sum_k \tilde{p}_k x_k$ of points x_k in the probability simplex,

$$\begin{aligned} \Big\| \sum_k (\tilde{p}_k - p_k)\, x_k \Big\|_1 &\le \sum_k |\tilde{p}_k - p_k|\, \|x_k\|_1 && \text{(triangle inequality)} \\ &= \sum_k |\tilde{p}_k - p_k| && (\|x_k\|_1 = 1) \\ &= 2\, \mathrm{TV}(\tilde{p}, p), \end{aligned}$$

which with $x_k = \sigma_{t_k}(I, \cdot)$ gives $\|\bar{\sigma}^{T,\text{op}}(I, \cdot) - \bar{\sigma}^T(I, \cdot)\|_1 \le 2\, \mathrm{TV}(\tilde{p}, p)$ for every set whose optimistic running sum is not zero. The optimistic weight on the tail is zero.

The weight ratio can also be read off the factorization of the own reach. By perfect recall the own reach is the product over the player's earlier decisions $(I_1, a_1), \ldots, (I_{d(I)}, a_{d(I)})$ on the path to I, so substituting into the definition of ρ_k,

$$\rho_k - 1 = \frac{\prod_{j \le d(I)} \sigma_{t_k}(I_j, a_j)}{\frac{1}{t_k - t_{k-1}} \sum_{\tau \in (t_{k-1}, t_k]} \prod_{j \le d(I)} \sigma_\tau(I_j, a_j)} - 1,$$

the relative deviation of the end-of-window product from its window mean. Each factor $\sigma_\tau(I_j, a_j)$ changes only at visits of I_j, by Lemma 1 applied to I_j. Under two-sided exploration every terminal history has sampling probability at least δ, so I is visited in each iteration with probability at least δ. Window lengths are then dominated by geometric variables with mean at most $1/\delta$. Under one-sided exploration the opponent is sampled from its current strategy. That strategy assigns probability exactly zero to an action with non-positive cumulative regret whenever another action's regret is positive, so the sampling probability of I can be zero for stretches of iterations while $\pi_i^{\sigma_t}(I)$ stays positive. □

In the notation of Section 4.3, $\pi_t = \pi_i^{\sigma_t}(I)$, $\sigma_t = \sigma_t(I, \cdot)$, $q_t = q_t(I)$, B_t indicates a visit at iteration t, $S(a) = \sum_t \pi_t \sigma_t(a)$, $N = \sum_t \pi_t$, $\nu^2 = \mathbb{E} \sum_t \pi_t^2 (1/q_t - 1)$, $\hat{S}(a) = \sum_t B_t\, \pi_t \sigma_t(a) / q_t$, $\hat{N} = \sum_b \hat{S}(b)$ and $\hat{\sigma} = \hat{S} / \hat{N}$.

Restatement (Theorem 2, what the stochastic rules return)**.** *(a) Under two-sided exploration, $q_t > 0$ for all t, $\mathbb{E}[\hat{S}(a) - S(a)] = 0$ and $\mathrm{Var}(\hat{S}(a) - S(a)) = \mathbb{E}\sum_t \pi_t^2 \sigma_t(a)^2 \,(1/q_t - 1)$. For every deterministic N_0 with $N \geq N_0$ almost surely, $\mathbb{E}\big[\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1 \mathbb{1}\{\hat{N} \geq N/2\}\big] \leq 4\sqrt{|A(I)|}\,\nu/N_0$ and $\Pr(\hat{N} < N/2) \leq 4\nu^2/N_0^2$; at depth zero $N_0 = T$. (b) Under one-sided exploration, on every iteration with $q_t = 0$ and $\pi_t > 0$ both the per-set and the per-history rule omit the term $\pi_t\sigma_t$ from their running sums, so they are biased for S.*

Proof. Let $\mathcal{G}_t$ be the sigma-field generated by the history before player i's traversal of iteration t. For the second player of an iteration $\mathcal{G}_t$ includes the first player's traversal. Write $q_t = q_t(I) = \sum_{h \in I} q_t(h)$ as in Section 3. Then π_t, σ_t and q_t are $\mathcal{G}_t$-measurable, because the opponent's traversal does not change the regrets of player i. Also $\Pr(B_t = 1 \mid \mathcal{G}_t) = q_t$. With $D_t = (B_t/q_t - 1)\,\pi_t\sigma_t(a)$,

$$\mathbb{E}[D_t \mid \mathcal{G}_t] = \Big(\frac{\mathbb{E}[B_t \mid \mathcal{G}_t]}{q_t} - 1\Big)\pi_t\sigma_t(a) = \Big(\frac{q_t}{q_t} - 1\Big)\pi_t\sigma_t(a) = 0,$$
$$\mathbb{E}[D_t^2 \mid \mathcal{G}_t] = \pi_t^2\sigma_t(a)^2\ \mathrm{Var}\left(\frac{B_t}{q_t}\,\middle|\,\mathcal{G}_t\right) = \pi_t^2\sigma_t(a)^2\,\frac{q_t(1-q_t)}{q_t^2} = \pi_t^2\sigma_t(a)^2\Big(\frac{1}{q_t} - 1\Big),$$

so the D_t are square-integrable martingale differences, as $q_t \geq \delta$. Since $\hat{S}(a) - S(a) = \sum_t D_t$, the tower property gives

$$\mathbb{E}\big[\hat{S}(a) - S(a)\big] = \sum_t \mathbb{E}\big[\mathbb{E}[D_t \mid \mathcal{G}_t]\big] = 0,$$
$$\mathrm{Var}\,\big(\hat{S}(a) - S(a)\big) = \mathbb{E}\Big[\sum_t D_t^2\Big] + 2\sum_{s<t}\mathbb{E}\big[D_s\,\mathbb{E}[D_t \mid \mathcal{G}_t]\big] = \mathbb{E}\sum_t \pi_t^2\sigma_t(a)^2\Big(\frac{1}{q_t} - 1\Big),$$

where the cross terms vanish because D_s is $\mathcal{G}_t$-measurable for $s < t$.

For the normalized error write $\hat{N} = \sum_b \hat{S}(b)$ and $N = \sum_b S(b) = \sum_t \pi_t$. Then

$$\begin{aligned}\hat{\sigma}(a) - \bar{\sigma}^T(I,a) &= \frac{\hat{S}(a)}{\hat{N}} - \frac{S(a)}{N} \\ &= \frac{\hat{S}(a) - S(a)}{\hat{N}} + S(a)\Big(\frac{1}{\hat{N}} - \frac{1}{N}\Big) \\ &= \frac{\hat{S}(a) - S(a)}{\hat{N}} + S(a)\,\frac{N - \hat{N}}{N\hat{N}}.\end{aligned}$$

On the event $\hat{N} \geq N/2$, summing absolute values over a and using $\sum_a S(a) = N$,

$$\begin{aligned}\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1 &\leq \frac{1}{\hat{N}}\sum_a |\hat{S}(a) - S(a)| + \frac{|N - \hat{N}|}{N\hat{N}}\sum_a S(a) \\ &\leq \frac{2}{N}\Big(\sum_a |\hat{S}(a) - S(a)| + |\hat{N} - N|\Big) && (\hat{N} \geq N/2) \\ &\leq \frac{4}{N}\sum_a |\hat{S}(a) - S(a)|. && (|\hat{N} - N| \leq \sum_a |\hat{S}(a) - S(a)|)\end{aligned}$$

Since $N \geq N_0$ almost surely with N_0 deterministic,

$$\begin{aligned}
\mathbb{E}\big[\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1 \, \mathbb{1}\{\hat{N} \geq N/2\}\big] &\leq \frac{4}{N_0} \sum_a \mathbb{E}|\hat{S}(a) - S(a)| \\
&\leq \frac{4}{N_0} \sum_a \sqrt{\mathbb{E}\big[(\hat{S}(a) - S(a))^2\big]} && \text{(Jensen)} \\
&\leq \frac{4\sqrt{|A(I)|}}{N_0} \sqrt{\sum_a \mathbb{E} \sum_t \pi_t^2 \sigma_t(a)^2 \Big(\frac{1}{q_t} - 1\Big)} \\
&\leq \frac{4\sqrt{|A(I)|}}{N_0} \sqrt{\mathbb{E} \sum_t \pi_t^2 \Big(\frac{1}{q_t} - 1\Big)} && (\textstyle\sum_a \sigma_t(a)^2 \leq 1) \\
&= \frac{4\sqrt{|A(I)|}\,\nu}{N_0},
\end{aligned}$$

where the third inequality is Cauchy–Schwarz over a. For $\Pr(\hat{N} < N/2)$, write $\hat{N} - N = \sum_a (\hat{S}(a) - S(a)) = \sum_t (B_t/q_t - 1)\,\pi_t \sum_a \sigma_t(a) = \sum_t (B_t/q_t - 1)\,\pi_t$. By the same martingale computation it has mean zero and variance $\mathbb{E}\sum_t \pi_t^2(1/q_t - 1) = \nu^2$, so

$$\Pr(\hat{N} < N/2) \leq \Pr\big(|\hat{N} - N| \geq N/2\big) \leq \Pr\big(|\hat{N} - N| \geq N_0/2\big) \leq \frac{4\,\nu^2}{N_0^2}. \qquad \text{(Chebyshev)}$$

At depth zero $\pi_t = 1$ for every t, so $N = T$ and $N_0 = T$ suffices. In the stationary case $\pi_t \equiv \bar{\pi}$ and $q_t \equiv q$ both sums are explicit:

$$\frac{\nu}{N} = \frac{\sqrt{T\bar{\pi}^2(1/q - 1)}}{T\bar{\pi}} = \sqrt{\frac{1-q}{q\,T}}.$$

For part (b), if $q_t(I) = 0$ then $B_t = 0$ surely and no weight can restore the term $\pi_t\sigma_t$. By the tower property over the adaptive trajectory,

$$\mathbb{E}\big[\hat{S}(a)\big] = \mathbb{E} \sum_{t:\, q_t > 0} \pi_t \sigma_t(a) \neq \mathbb{E} \sum_t \pi_t \sigma_t(a) = \mathbb{E}\, S(a)$$

whenever some iteration has $q_t = 0$ and $\pi_t\sigma_t(a) > 0$ with positive probability, so the per-set rule estimates $\sum_{t:q_t>0} \pi_t\sigma_t$ rather than S. For the per-history rule the conditional mean of the increment is $|I_t|\,\pi_t\sigma_t(a)$ with $|I_t|$ the number of histories of I that have positive sampling probability at iteration t. Hence its running sum estimates $\sum_{t:q_t>0} |I_t|\,\pi_t\sigma_t$ and the same conclusion holds. □

Lemma 4 (per-history increment)**.** *Under two-sided exploration the increment of the per-history rule divided by $|I|$ has the same conditional mean as that of the per-set rule and at least its conditional variance. When $\pi_t\sigma_t(a) > 0$ the two variances are equal if and only if $q_t(h)$ is constant across $h \in I$.*

Proof. For the per-history rule, the increment at a visit through the sampled history $h \in I$ is $\pi_t\sigma_t(a)/q_t(h)$, so its conditional second moment and the per-set quantities compare as

$$\begin{aligned}
\mathbb{E}\Big[\Big(\frac{\pi_t\sigma_t(a)}{q_t(h)}\Big)^2 B_t \,\Big|\, \mathcal{G}_t\Big] &= \sum_{h \in I} q_t(h) \Big(\frac{\pi_t\sigma_t(a)}{q_t(h)}\Big)^2 = \pi_t^2\sigma_t(a)^2 \sum_{h \in I} \frac{1}{q_t(h)} \\
&\geq \pi_t^2\sigma_t(a)^2\, \frac{|I|^2}{\sum_{h \in I} q_t(h)} = |I|^2\, \pi_t^2\sigma_t(a)^2\, \frac{1}{q_t(I)}, \quad \text{(AM–HM inequality)}
\end{aligned}$$

with equality if and only if $q_t(h)$ is constant on I or $\pi_t\sigma_t(a) = 0$. Its conditional mean is

$$\mathbb{E}\Big[\frac{\pi_t\sigma_t(a)}{q_t(h)} B_t \,\Big|\, \mathcal{G}_t\Big] = \sum_{h \in I} q_t(h)\, \frac{\pi_t\sigma_t(a)}{q_t(h)} = |I|\,\pi_t\sigma_t(a).$$

Hence the per-history increment divided by $|I|$ has the same conditional mean as the per-set increment, namely $\pi_t\sigma_t(a)$. Its conditional second moment is at least $\pi_t^2\sigma_t(a)^2/q_t(I)$, that of the per-set

increment. Comparing second moments therefore compares conditional variances, with the same equality condition. Under two-sided exploration the factor $|I|$ is the same at every iteration and cancels in the normalization, whereas under one-sided exploration the number of histories of I with positive sampling probability can vary with t. □

Corollary 1 (two-sided rate)**.** *In the setting of Theorem 2, assume almost surely $\pi_t \geq \underline{\pi}$ and $q_t \leq \bar{q}$ on a fraction θ of the iterations and $q_t \geq \delta$ throughout. Then $N \geq N_0 := \theta T \underline{\pi}$ almost surely, N_0 is admissible in Theorem 2(a) and $\sqrt{\theta}\,\underline{\pi}\sqrt{(1-\bar{q})/\bar{q}}\;T^{-1/2} \leq \nu/N \leq \nu/N_0 \leq (\theta\underline{\pi})^{-1}\sqrt{(1-\delta)/\delta}\;T^{-1/2}$.*

Proof. Restricting the sum in ν^2 to the fraction θ of iterations on which $\pi_t \geq \underline{\pi}$ and $q_t \leq \bar{q}$ gives the lower bound, and $\pi_t \leq 1$ with $q_t \geq \delta$ gives the upper bound:

$$\frac{\nu}{N} \geq \frac{\sqrt{\theta T\,\underline{\pi}^2\,(1/\bar{q}-1)}}{T} = \sqrt{\theta}\,\underline{\pi}\sqrt{\frac{1-\bar{q}}{\bar{q}}}\;T^{-1/2}, \qquad (N \leq T)$$

$$\frac{\nu}{N} \leq \frac{\sqrt{T\,(1/\delta-1)}}{\theta T\underline{\pi}} = (\theta\underline{\pi})^{-1}\sqrt{\frac{1-\delta}{\delta}}\;T^{-1/2}. \qquad (N \geq \theta T\underline{\pi})$$

□

Remark 2 (the reach condition of Corollary 1 is needed)**.** *Both halves of the bound fail without it. For the upper half, take a set whose own reach is one at the first iteration and zero afterwards. Regret matching produces it as soon as the action leading to it has non-positive cumulative regret and another action's regret is positive. Then $N = 1$ and $\nu^2 = 1/q_1 - 1$ for every T, so ν/N does not decrease with the horizon. The estimate is that bad, because with probability $1 - q_1$ the set is never visited and $\hat{\sigma}$ is the uniform strategy of the convention. For the lower half, a first decision of a game without a chance move has $q_t = 1$ and therefore $\nu = 0$. Corollary 2 below needs neither condition.*

Corollary 2 (exploitability of the per-set rule)**.** *Write $\nu(I)$ and $N(I)$ for the quantities of Theorem 2 at the information set I, and let $\hat{\sigma}$ be the profile that the per-set rule returns at every set. Under two-sided exploration, and with no condition on the own reach, the expected exploitability of $\hat{\sigma}$ exceeds that of the exact average by at most*

$$\frac{\Delta_u}{T}\sum_I \left(\sqrt{|A(I)|}+2\right)\nu(I)$$

$$\leq\ \Delta_u\sqrt{\frac{1-\delta}{\delta}}\;\frac{1}{\sqrt{T}}\sum_I \left(\sqrt{|A(I)|}+2\right)\sqrt{\mathbb{E}\big[\pi_i^{\bar{\sigma}}(I)\big]},$$

the sums running over the information sets of both players.

Proof. Fix I and write π_t, σ_t, q_t, S, N, $\hat{S}$, $\hat{N}$ and $\hat{\sigma}$ for its quantities. By Lemma 2 the own reach of the exact average at I is the average of the own reaches of the iterates,

$$\pi_i^{\bar{\sigma}}(I) = \bar{x}(\mathrm{par}(I)) = \frac{1}{T}\sum_{t\leq T}\pi_i^{\sigma_t}(I) = \frac{N}{T}. \tag{5}$$

If $N = 0$ then $\pi_t = 0$ for every t and $\hat{S} = 0$. Both averages are the uniform strategy of the convention of Section 3 and the set contributes nothing. Assume $N > 0$. On the event $\{\hat{N} \geq N/2\}$ the proof of Theorem 2(a) gives $\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1 \leq \frac{4}{N}\sum_a |\hat{S}(a) - S(a)|$, and multiplying by (5) cancels N:

$$\pi_i^{\bar{\sigma}}(I)\,\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1\,\mathbb{1}\{\hat{N} \geq N/2\} \leq \frac{4}{T}\sum_a |\hat{S}(a) - S(a)|.$$

Taking expectations and repeating the Jensen and Cauchy–Schwarz steps of that proof with T in place of N_0,

$$\mathbb{E}\big[\pi_i^{\bar{\sigma}}(I)\,\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1\,\mathbb{1}\{\hat{N} \geq N/2\}\big] \leq \frac{4}{T}\sum_a \sqrt{\mathbb{E}\big[(\hat{S}(a) - S(a))^2\big]} \leq \frac{4\sqrt{|A(I)|}\,\nu(I)}{T}.$$

On the complementary event the ℓ_1 distance is at most 2, so that part contributes at most $\frac{2}{T}\mathbb{E}[N\,\mathbb{1}\{\hat{N} < N/2\}]$. The difference $\hat{N} - N$ has mean zero and variance $\nu(I)^2$, so for every deterministic $c > 0$

$$\mathbb{E}\big[N\,\mathbb{1}\{\hat{N} < N/2\}\big] \le \mathbb{E}\big[N\,\mathbb{1}\{N \le c\}\big] + \mathbb{E}\Big[N\,\frac{4(\hat{N}-N)^2}{N^2}\,\mathbb{1}\{N > c\}\Big] \qquad \text{(on that event)}$$
$$\le c + \frac{4}{c}\,\mathbb{E}\big[(\hat{N}-N)^2\big] = c + \frac{4\,\nu(I)^2}{c} = 4\,\nu(I), \qquad (c = 2\nu(I))$$

and if $\nu(I) = 0$ then $\hat{S} = S$ almost surely and this term is zero. Adding the two parts gives $\mathbb{E}[\pi_i^{\bar{\sigma}}(I)\,\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1] \le 4(\sqrt{|A(I)|} + 2)\,\nu(I)/T$ at every information set. Lemma 3 with σ the exact average and $\tilde{\sigma} = \hat{\sigma}$ holds for every realization, so taking expectations bounds the expected excess by $\frac{1}{4}\Delta_u \sum_I \mathbb{E}[\pi_i^{\bar{\sigma}}(I)\,\|\hat{\sigma} - \bar{\sigma}^T(I)\|_1]$. That is the first bound. The second follows from $\pi_t \le 1$ and $q_t \ge \delta$:

$$\nu(I)^2 = \mathbb{E}\sum_t \pi_t^2\Big(\frac{1}{q_t} - 1\Big) \le \Big(\frac{1}{\delta} - 1\Big)\mathbb{E}\sum_t \pi_t = \Big(\frac{1}{\delta} - 1\Big)\,T\,\mathbb{E}\big[\pi_i^{\bar{\sigma}}(I)\big]. \qquad \text{(by (5))}$$

□

The per-set relative error $\nu(I)/N(I)$ of Corollary 1 does need a condition on the own reach, and Remark 2 shows that it cannot be dropped there. The exploitability does not, because Lemma 3 weights a set by the own reach of the exact average at it. That weight is $N(I)/T$ and cancels the $N(I)$ that Theorem 2(a) divides by. A set counts in the constant only through the square root of its expected own reach.

For rules that add on the opponent's traversal, fix an information set I of player i, and let $s = 1$ if player i's traversal comes first in every iteration and $s = 0$ if it comes second. When the opponent traverses in iteration t, the strategy of player i in force is σ_{t+s}, because for $s = 1$ player i has already been updated in that iteration. Abbreviate $\tilde{\pi}_t = \pi_i^{\sigma_{t+s}}(I)$ and $\tilde{\sigma}_t = \sigma_{t+s}(I,\cdot)$. On that traversal the opponent samples from its mixture $\varepsilon\,\mathrm{unif} + (1-\varepsilon)\sigma_t$. Let $q_t^{-i}(h)$ be the probability that chance and the opponent give to h, and $q_t^{-i}(I) = \sum_{h\in I} q_t^{-i}(h)$. Let B_t indicate that the opponent's traversal of iteration t visits I through the history h_t, and let $q_t(h)$ and $q_t(I)$ be the probabilities of Section 3 on that traversal. The per-set and the per-history rule then hold

$$\hat{S}(a) = \sum_{t=1}^{T} B_t\,\frac{\tilde{\pi}_t\tilde{\sigma}_t(a)}{q_t(I)}, \qquad\qquad \hat{S}^{\mathrm{h}}(a) = \sum_{t=1}^{T} B_t\,\frac{\tilde{\pi}_t\tilde{\sigma}_t(a)}{q_t(h_t)},$$

and we compare them with

$$S^+(a) = \sum_{t=1}^{T} \tilde{\pi}_t\tilde{\sigma}_t(a) = \sum_{t=1+s}^{T+s} \pi_i^{\sigma_t}(I)\,\sigma_t(I,a), \qquad\qquad N^+ = \sum_{t=1}^{T} \tilde{\pi}_t.$$

We write $\bar{\sigma}^+(I,\cdot) = S^+/N^+$, the exact average (1) of the iterates $\sigma_{1+s},\dots,\sigma_{T+s}$. The number δ is the one defined in Section 2. Here it acts only as a constant, since under one-sided exploration it does not bound the sampling probability of a terminal history.

Theorem 4 (the stochastic rules on the opponent's traversal)**.** *Theorem 3 is parts (a), (b) and (d) below, and (c) adds the error bounds. Assume perfect recall and one-sided exploration with $\varepsilon > 0$, and let the stochastic rules add on the opponent's traversal. (a) For every t and every $h \in I$, $q_t(h) = \tilde{\pi}_t\,q_t^{-i}(h)$ with $q_t^{-i}(h) \ge \delta$. Hence $q_t(I) = \tilde{\pi}_t\,q_t^{-i}(I)$ is positive on exactly the iterations on which the own reach $\tilde{\pi}_t$ is, and neither rule omits a term to which S^+ gives weight. (b) The per-set sum satisfies $\mathbb{E}[\hat{S}(a) - S^+(a)] = 0$ and $\mathrm{Var}(\hat{S}(a) - S^+(a)) = \mathbb{E}\sum_t \tilde{\pi}_t\tilde{\sigma}_t(a)^2\big(1/q_t^{-i}(I) - \tilde{\pi}_t\big)$. The per-history sum satisfies $\mathbb{E}[\hat{S}^{\mathrm{h}}(a)] = |I|\,\mathbb{E}[S^+(a)]$ and $\mathrm{Var}(\hat{S}^{\mathrm{h}}(a)/|I| - S^+(a)) = \mathbb{E}\sum_t \tilde{\pi}_t\tilde{\sigma}_t(a)^2\big(|I|^{-2}\sum_{h\in I} 1/q_t^{-i}(h) - \tilde{\pi}_t\big)$, which is at least the variance of the per-set sum. (c) Let $\nu_+^2 = \mathbb{E}\sum_t \tilde{\pi}_t\big(1/q_t^{-i}(I) - \tilde{\pi}_t\big)$ and $\nu_{\mathrm{h}}^2 = \mathbb{E}\sum_t \tilde{\pi}_t\big(|I|^{-2}\sum_{h\in I} 1/q_t^{-i}(h) - \tilde{\pi}_t\big)$; both are at most $\mathbb{E}[N^+]/\delta$. The two bounds of Theorem 2(a) hold for the normalized per-set sum with ν_+, N^+ and $\bar{\sigma}^+(I,\cdot)$ in place of ν, N and $\bar{\sigma}^T(I)$, and for the normalized per-history sum with ν_{h}. (d) If $s = 0$ then $S^+ = S_I^T$. If $s = 1$ then $S^+(a) - S_I^T(a) = \pi_i^{\sigma_{T+1}}(I)\,\sigma_{T+1}(I,a) - \pi_i^{\sigma_1}(I)\,\sigma_1(I,a)$ for every run, and against every strategy of the opponent the payoffs of $\bar{\sigma}^+$ and of the exact average differ by at most Δ_u/T.*

Proof. (a) On the opponent's traversal player i is sampled from σ_{t+s} and does not explore, so the probability of sampling $h \in I$ is $\pi_i^{\sigma_{t+s}}(h)\, q_t^{-i}(h)$. By perfect recall $\pi_i^{\sigma_{t+s}}(h) = \tilde{\pi}_t$ for every $h \in I$ (Section 2), which gives the factorization. Summing over h gives $q_t(I) = \tilde{\pi}_t\, q_t^{-i}(I)$. The history h is a prefix of a terminal history z, carries at most L decisions (the largest number of decisions on a terminal history, Appendix B.3), and has chance probability at least that of z. At each of its decisions the opponent gives every action probability at least $\varepsilon / \max_{I'} |A(I')|$, a number that is at most 1. Raising that number to the power L can therefore only lower the bound. Hence

$$q_t^{-i}(h) \geq \Big(\frac{\varepsilon}{\max_{I'} |A(I')|}\Big)^{L} \min_z \pi_c(z) = \delta,$$

where $\pi_c(z)$ is the chance probability of z. Hence $q_t(I) > 0$ if and only if $\tilde{\pi}_t > 0$.

(b) Let $\mathcal{H}_t$ be the sigma-field generated by the history before the opponent's traversal of iteration t; for $s = 1$ it includes player i's traversal and update of iteration t. Then $\tilde{\pi}_t$, $\tilde{\sigma}_t$ and every $q_t^{-i}(h)$ are $\mathcal{H}_t$-measurable, and $\Pr(B_t = 1,\ h_t = h \mid \mathcal{H}_t) = q_t(h)$. Let $D_t = B_t\, \tilde{\pi}_t \tilde{\sigma}_t(a)/q_t(I) - \tilde{\pi}_t \tilde{\sigma}_t(a)$, the error of the per-set rule in iteration t. If $\tilde{\pi}_t = 0$ then $q_t(I) = 0$ by (a), so $B_t = 0$ surely and $D_t = 0$. If $\tilde{\pi}_t > 0$ then

$$\begin{aligned}
\mathbb{E}[D_t \mid \mathcal{H}_t] &= \Big(\frac{q_t(I)}{q_t(I)} - 1\Big)\tilde{\pi}_t \tilde{\sigma}_t(a) = 0, \\
\mathbb{E}[D_t^2 \mid \mathcal{H}_t] &= \tilde{\pi}_t^2 \tilde{\sigma}_t(a)^2 \Big(\frac{1}{q_t(I)} - 1\Big) = \tilde{\pi}_t \tilde{\sigma}_t(a)^2 \Big(\frac{1}{q_t^{-i}(I)} - \tilde{\pi}_t\Big), \qquad \text{(by (a))}
\end{aligned}$$

and the last expression is zero, the correct value, when $\tilde{\pi}_t = 0$. It is at most $1/\delta$, so the D_t are square-integrable martingale differences. D_r is $\mathcal{H}_t$-measurable for $r < t$. Since $\hat{S}(a) - S^+(a) = \sum_t D_t$, the tower-property computation in the proof of Theorem 2(a) gives the mean and the variance. For the per-history rule let $D_t^{\mathrm{h}} = B_t\, \tilde{\pi}_t \tilde{\sigma}_t(a) / \big(|I|\, q_t(h_t)\big) - \tilde{\pi}_t \tilde{\sigma}_t(a)$, which is again zero when $\tilde{\pi}_t = 0$. When $\tilde{\pi}_t > 0$ every $h \in I$ has $q_t(h) > 0$ by (a), so all $|I|$ histories enter the conditional mean:

$$\begin{aligned}
\mathbb{E}\Big[\frac{B_t\, \tilde{\pi}_t \tilde{\sigma}_t(a)}{|I|\, q_t(h_t)} \,\Big|\, \mathcal{H}_t\Big] &= \sum_{h \in I} q_t(h)\, \frac{\tilde{\pi}_t \tilde{\sigma}_t(a)}{|I|\, q_t(h)} = \tilde{\pi}_t \tilde{\sigma}_t(a), \\
\mathbb{E}\Big[\Big(\frac{B_t\, \tilde{\pi}_t \tilde{\sigma}_t(a)}{|I|\, q_t(h_t)}\Big)^2 \,\Big|\, \mathcal{H}_t\Big] &= \frac{\tilde{\pi}_t^2 \tilde{\sigma}_t(a)^2}{|I|^2} \sum_{h \in I} \frac{1}{q_t(h)} = \frac{\tilde{\pi}_t \tilde{\sigma}_t(a)^2}{|I|^2} \sum_{h \in I} \frac{1}{q_t^{-i}(h)}. \qquad \text{(by (a))}
\end{aligned}$$

Subtracting the squared mean gives the conditional variance, and the same martingale computation gives the stated mean and variance. In contrast to Theorem 2(b), the number of histories of I with positive sampling probability is $|I|$ on every iteration with $\tilde{\pi}_t > 0$, so $|I|$ is a constant of the game and cancels in the normalization. By Cauchy–Schwarz, $|I|^2 \leq \big(\sum_{h \in I} q_t^{-i}(h)\big)\big(\sum_{h \in I} 1/q_t^{-i}(h)\big)$, so $|I|^{-2} \sum_{h \in I} 1/q_t^{-i}(h) \geq 1/q_t^{-i}(I)$ and the variance is at least that of the per-set sum.

(c) Summing the conditional variances in (b) over a with $\sum_a \tilde{\sigma}_t(a)^2 \leq 1$ bounds them by ν_+^2 and ν_{h}^2. For the total, $\sum_a D_t$ equals $(B_t/q_t(I) - 1)\, \tilde{\pi}_t$ when $\tilde{\pi}_t > 0$ and zero otherwise, so $\hat{N} - N^+ = \sum_t \sum_a D_t$ has mean zero and variance ν_+^2. In the same way the error of the per-history total divided by $|I|$ has mean zero and variance ν_{h}^2. The second half of the proof of Theorem 2(a) uses only these properties of the running sum, so both of its bounds follow. Finally $q_t^{-i}(I) \geq \delta$ gives $\nu_+^2 \leq \mathbb{E} \sum_t \tilde{\pi}_t/\delta$, and $|I|^{-2} \sum_{h \in I} 1/q_t^{-i}(h) \leq 1/(|I|\,\delta) \leq 1/\delta$ gives the same bound for ν_{h}^2.

(d) The case $s = 0$ is the definition of S^+. For $s = 1$, write $x_t(I, a) = \pi_i^{\sigma_t}(I)\, \sigma_t(I, a)$ as in Lemma 2; then

$$S^+(a) - S_I^T(a) = \sum_{t=2}^{T+1} x_t(I, a) - \sum_{t=1}^{T} x_t(I, a) = x_{T+1}(I, a) - x_1(I, a).$$

By Lemma 2 applied to $\sigma_2, \ldots, \sigma_{T+1}$, the strategy $\bar{\sigma}^+$ realizes the sequence-form average $\frac{1}{T} \sum_{t=2}^{T+1} x_t$. The payoff of player i against a fixed strategy σ_{-i} is linear in the sequence-form

strategy of player i, so

$$u_i(\bar{\sigma}^+, \sigma_{-i}) - u_i(\bar{\sigma}^T, \sigma_{-i}) = \frac{1}{T}\sum_{t=2}^{T+1} u_i(\sigma_t, \sigma_{-i}) - \frac{1}{T}\sum_{t=1}^{T} u_i(\sigma_t, \sigma_{-i}) = \frac{u_i(\sigma_{T+1}, \sigma_{-i}) - u_i(\sigma_1, \sigma_{-i})}{T},$$

whose absolute value is at most Δ_u/T. □

Section 3 notes that (2) fails on the opponent's traversal, because the strategy at I changes between the visits at which a rule adds. That concerns the rules that give a visit the weight of a window. The stochastic rules use no window: each increment is divided by its own visit probability. The proof above uses only that the quantities of iteration t are known before the opponent's traversal of iteration t. Under two-sided exploration player i explores on the opponent's traversal as well, and every $q_t(h)$ is at least δ. Theorem 2(a) and Lemma 4 apply verbatim with $\tilde{\pi}_t$, $\tilde{\sigma}_t$, $\mathcal{H}_t$ and S^+.

Corollary 3 (exploitability on the opponent's traversal)**.** *Let both players' stochastic rules add on the opponent's traversal under one-sided exploration, let $\hat{\sigma}$ be the profile that the per-set rule returns, and write $\nu_+(I)$ and $\bar{\sigma}^+$ for the quantities of Theorem 4 at I. With no condition on the own reach, the expected exploitability of $\hat{\sigma}$ exceeds that of the exact average by at most*

$$\frac{\Delta_u}{T}\sum_I \left(\sqrt{|A(I)|}+2\right)\nu_+(I) + \frac{\Delta_u}{2T} \le \frac{\Delta_u}{\sqrt{\delta\, T}}\sum_I \left(\sqrt{|A(I)|}+2\right)\sqrt{\mathbb{E}\left[\pi_i^{\bar{\sigma}^+}(I)\right]} + \frac{\Delta_u}{2T},$$

the sums running over the information sets of both players. The same holds for the per-history rule with $\nu_{\mathrm{h}}(I)$ in place of $\nu_+(I)$.

Proof. Let $\bar{\sigma}^+$ also denote the profile that plays $\bar{\sigma}^+$ for the player with $s = 1$ and the exact average for the player with $s = 0$. The proof of Corollary 2 uses four properties of the per-set sum at a set. Its error is a sum of martingale differences with total variance at most $\nu(I)^2$. The error of its total has variance $\nu(I)^2$, the sum is zero when its target is zero, and the own reach of the target average at I is its total divided by T. Theorem 4(b) and (c) give the first two for $\hat{S}$ with target S^+ and $\nu_+(I)$, and for $\hat{S}^{\mathrm{h}}/|I|$ with $\nu_{\mathrm{h}}(I)$. If $N^+ = 0$ then every $\tilde{\pi}_t = 0$, no visit occurs by Theorem 4(a), and both sums are zero. Lemma 2 applied to $\sigma_{1+s}, \ldots, \sigma_{T+s}$ gives $\pi_i^{\bar{\sigma}^+}(I) = N^+/T$. That proof therefore bounds the expected excess of $\hat{\sigma}$ over the profile $\bar{\sigma}^+$ by the first term, and $\nu_+(I)^2 \le \mathbb{E}[N^+]/\delta = T\,\mathbb{E}[\pi_i^{\bar{\sigma}^+}(I)]/\delta$ gives the first term of the second line. The profiles $\bar{\sigma}^+$ and $\bar{\sigma}^T$ differ only in the strategy of the player with $s = 1$. By Theorem 4(d) the best-response value of the other player against that strategy changes by at most Δ_u/T, as in the last step of the proof of Lemma 3. The exploitability is half the sum of the two best-response gains, so it changes by at most $\Delta_u/(2T)$. □

The offset $\Delta_u/(2T)$ is of lower order than the sampling error and not random. No condition on the opponent's strategy enters either: on the player's own traversal an on-policy opponent can give every history of I probability zero. Here the opponent explores, and the exact average weights player i's own play.

Remark 3 (weighted iterations on the opponent's traversal)**.** *With deterministic iteration weights $\alpha_t > 0$ the stochastic rules add α_t times their increment, and their target on the opponent's traversal is $\sum_t \alpha_t \tilde{\pi}_t \tilde{\sigma}_t$. Theorem 4(a) to (c) hold with α_t^2 inside every variance sum and α_t inside N^+, by the same proof. In Corollary 3 the horizon T becomes the total weight $A_T = \sum_{t\le T}\alpha_t$, and $\sum_t \alpha_t^2 \tilde{\pi}_t \le \max_t \alpha_t\, A_T\, \pi_i^{\bar{\sigma}^+}(I)$ turns the first term into $\Delta_u \sqrt{\max_t \alpha_t/(\delta A_T)}\sum_I(\sqrt{|A(I)|}+2)\sqrt{\mathbb{E}[\pi_i^{\bar{\sigma}^+}(I)]}$. The offset no longer telescopes, but summation by parts gives*

$$\sum_{t=1}^{T}\alpha_t\,(x_{t+1}-x_t) = \alpha_T\, x_{T+1} - \alpha_1\, x_1 - \sum_{t=2}^{T}(\alpha_t - \alpha_{t-1})\, x_t,$$

a combination of sequence-form strategies whose coefficients sum to zero and have absolute sum $\alpha_1 + \alpha_T + \sum_{t=2}^{T} |\alpha_t - \alpha_{t-1}|$. *The offset term of Corollary 3 becomes* $\Delta_u\big(\alpha_1 + \alpha_T + \sum_{t=2}^{T} |\alpha_t - \alpha_{t-1}|\big)/(4A_T)$, *which is* $\Delta_u\, \alpha_T/(2A_T)$ *for non-decreasing weights and* $\Delta_u/(T+1)$ *for* $\alpha_t = t$.

We now ask how far the optimistic average is from the exact one, and whether the distance vanishes as T grows. Fix a run, a horizon T and an information set I of player i, and write $\pi_s = \pi_i^{\sigma_s}(I)$. Let $E_I(T) = \sum_{k \le K} |\tilde{w}_k - w_k| = \sum_{k \le K} |\rho_k - 1|\, w_k$ be the weight error of Section 4.2 summed over the windows, each window counted with its exact weight.

Theorem 5 (distance of the optimistic average from the exact average)**.** *(a) At every* I *with* $W_I(T) > 0$, $\|\bar{\sigma}^{T,\mathrm{op}}(I,\cdot) - \bar{\sigma}^{T}(I,\cdot)\|_1 \le 2\,\big(E_I(T) + w_{\mathrm{tail}}\big)/W_I(T)$. *(b) The exploitability of the optimistic averages exceeds that of the exact averages by at most* $\frac{\Delta_u}{2T}\sum_I \big(E_I(T) + w_{\mathrm{tail}}(I)\big)$, *the sum running over the information sets of both players. (c) Every window has* $0 \le \rho_k \le t_k - t_{k-1}$, *with* $\rho_k = 0$ *when the visit finds the own reach at zero and* $\rho_k = t_k - t_{k-1}$ *when the own reach is zero until the last iteration of the window, so no bound on* ρ_k *holds that does not grow with the window. Moreover*

$$E_I(T) \le \sum_{k \le K} \sum_{s=t_{k-1}+1}^{t_k - 1} (s - t_{k-1})\,\big|\pi_{s+1} - \pi_s\big|, \tag{6}$$
$$\big|\pi_{s+1} - \pi_s\big| \le \sum_{j \le d(I)} \big|\sigma_{s+1}(I_j, a_j) - \sigma_s(I_j, a_j)\big|,$$

where (I_j, a_j) *are the player's earlier decisions on the path to* I*: each change of the own reach counts as many times as the window is old when it happens. (d) Under two-sided exploration* $\mathbb{E}[w_{\mathrm{tail}}] \le 1/\delta$ *and* $\mathbb{E}[E_I(T)] \le \frac{\sqrt{2}}{\delta}\sum_{s<T}\sqrt{\mathbb{E}\big[(\pi_{s+1} - \pi_s)^2\big]}$.

Proof. (a) Write $S^{\mathrm{op}} = \sum_{k \le K} \tilde{w}_k\, \sigma_{t_k}(I,\cdot)$ for the running sum of the optimistic rule. By (2),

$$S^{\mathrm{op}} - S_I^T = \sum_{k \le K} (\tilde{w}_k - w_k)\, \sigma_{t_k}(I,\cdot) - w_{\mathrm{tail}}\, \sigma_{t_K+1}(I,\cdot),$$

and every strategy has ℓ_1 norm one, so $\|S^{\mathrm{op}} - S_I^T\|_1 \le E_I(T) + w_{\mathrm{tail}}$. For two non-negative vectors ξ and ζ with $\|\xi\|_1, \|\zeta\|_1 > 0$,

$$\Big\|\frac{\xi}{\|\xi\|_1} - \frac{\zeta}{\|\zeta\|_1}\Big\|_1 \le \frac{\|\xi - \zeta\|_1}{\|\zeta\|_1} + \|\xi\|_1\Big|\frac{1}{\|\xi\|_1} - \frac{1}{\|\zeta\|_1}\Big|$$
$$= \frac{\|\xi - \zeta\|_1}{\|\zeta\|_1} + \frac{\big|\|\zeta\|_1 - \|\xi\|_1\big|}{\|\zeta\|_1} \le \frac{2\,\|\xi - \zeta\|_1}{\|\zeta\|_1}.$$

With $\xi = S^{\mathrm{op}}$ and $\zeta = S_I^T$, whose ℓ_1 norm is $W_I(T)$, this is (a). If $S^{\mathrm{op}} = 0$ the rule returns the uniform strategy; then every $\tilde{w}_k = 0$, so $E_I(T) + w_{\mathrm{tail}} = W_I(T)$, the bound equals 2 and holds trivially.

(b) By Lemma 3 with σ the exact average, the excess is at most $\frac{1}{4}\Delta_u \sum_I \pi_i^{\bar{\sigma}}(I)\, \|\bar{\sigma}^{T,\mathrm{op}}(I,\cdot) - \bar{\sigma}^{T}(I,\cdot)\|_1$. By (5) the weight is $\pi_i^{\bar{\sigma}}(I) = W_I(T)/T$, which cancels the denominator of (a); a set with $W_I(T) = 0$ has weight zero and $E_I(T) = w_{\mathrm{tail}} = 0$.

(c) The window weight contains the term of its last iteration, so $w_k \ge \pi_{t_k}$ and

$$\rho_k = \frac{(t_k - t_{k-1})\,\pi_{t_k}}{w_k} \le t_k - t_{k-1},$$

with equality exactly when $\pi_\tau = 0$ for every τ of the window before t_k, and $\rho_k = 0$ exactly when $\pi_{t_k} = 0 < w_k$. Regret matching produces both cases: it gives probability exactly zero to an action with non-positive cumulative regret while another action's regret is positive, a single update can make that regret positive, and exploration visits I whatever its own reach. For (6), telescope each

difference $\pi_{t_k} - \pi_\tau$ and exchange the two sums:

$$\tilde{w}_k - w_k = \sum_{\tau=t_{k-1}+1}^{t_k} \big(\pi_{t_k} - \pi_\tau\big) = \sum_{\tau=t_{k-1}+1}^{t_k} \sum_{s=\tau}^{t_k-1} \big(\pi_{s+1} - \pi_s\big)$$
$$= \sum_{s=t_{k-1}+1}^{t_k-1} (s - t_{k-1}) \big(\pi_{s+1} - \pi_s\big),$$

since $s - t_{k-1}$ is the number of τ in the window with $\tau \le s$. Taking absolute values and summing over k gives the first inequality. The second holds because π_s is the product of the $d(I)$ numbers $\sigma_s(I_j, a_j) \in [0,1]$, and two such products differ by at most the sum of the differences of their factors.

(d) Under two-sided exploration every iteration visits I with conditional probability at least δ given the past. Hence $\Pr(T - t_K \ge m) \le (1-\delta)^m$, and $w_{\text{tail}} \le T - t_K$ gives $\mathbb{E}[w_{\text{tail}}] \le \sum_{m\ge1}(1-\delta)^m \le 1/\delta$. Let $\kappa_s = s - t_{k-1}$ be the coefficient of (6) at iteration s of window k. Then $\Pr(\kappa_s \ge m) \le (1-\delta)^{m-1}$, so

$$\mathbb{E}[\kappa_s^2] = \sum_{m\ge1} (2m-1) \Pr(\kappa_s \ge m) \le \sum_{m\ge1} (2m-1)(1-\delta)^{m-1} = \frac{2}{\delta^2} - \frac{1}{\delta} \le \frac{2}{\delta^2}.$$

The Cauchy–Schwarz inequality, which needs no independence between κ_s and the movement of the own reach, gives $\mathbb{E}\big[\kappa_s |\pi_{s+1} - \pi_s|\big] \le \frac{\sqrt{2}}{\delta}\sqrt{\mathbb{E}[(\pi_{s+1} - \pi_s)^2]}$, and summing over s is (d). □

Theorem 5 replaces the question of whether ρ_k is bounded by a different one. A bound on ρ_k is not needed: under two-sided exploration at $\varepsilon = 0.6$ the largest ratio we record grows past 10^4 on Goofspiel while the optimistic average stays within 2% of the exact one in mean exploitability. Nor does a bound suffice for an argument that goes through Theorem 1(a). If the windows at I fall into two classes with ratios c and C, exact weights $\lambda_1 W_I(T)$ and $\lambda_2 W_I(T)$ and class averages $\bar\sigma_1$ and $\bar\sigma_2$, then without a tail

$$\bar\sigma^{T,\text{op}}(I,\cdot) - \bar\sigma^T(I,\cdot) = \Big(\frac{c\,\lambda_1}{c\,\lambda_1 + C\lambda_2} - \lambda_1\Big)\big(\bar\sigma_1 - \bar\sigma_2\big),$$

which does not shrink with T unless $c = C$ or the two class averages agree. The bound of Theorem 5(b) vanishes when the share $\sum_I (E_I(T) + w_{\text{tail}}(I))/T$ does, and by (6) that share is small when windows are short or the own reach moves little while a window is open. Two-sided exploration keeps windows short, by (d). The own reach moves when an earlier decision is updated, and a regret-matching update with sampled regret vector r moves the strategy in ℓ_1 by at most $2\|r\|_1$ divided by the sum of the positive cumulative regrets after the update, when that sum is positive. If that sum grows like $\sqrt{s}$ at the earlier decisions of I, the movement is $O(s^{-1/2})$ and (b) and (d) give an excess of order $T^{-1/2}$, the order of the bound on the exact average itself. Regret matching alone does not force that growth. Under one-sided exploration nothing bounds the age of a window, and the next proposition shows that the share need not vanish.

Restatement (Proposition 2, without a tail, a rule adding on the player's own traversal need not converge)**.** *Let the opponent choose x or y, where y ends the game at payoff zero, and after x let player i choose at the information set I between L, which pays player i one, and R, which pays $-G$ with $G > 1$. Run outcome-sampling MCCFR with regret matching and one-sided exploration with any $\varepsilon \in (0,1]$, player i traversing first in every iteration, each player's regrets updated at the end of its own traversal rather than at the end of the iteration. With probability at least $1/4$ the following holds for every $T \ge 1$. The optimistic rule, the per-set rule and the per-history rule, adding on player i's traversal, return the uniform strategy at I, and the profile they return has exploitability at least $(G-1)/4$. The exact average has exploitability $1/(4T)$ once $T \ge (1+G)/2$.*

Proof. At a visited set with strategy σ and sampled action a^*, outcome sampling adds to the cumulative regret of a^* the amount $\omega u\,(1 - \sigma(a^*))$ and to that of every other action $-\omega u\,\sigma(a^*)$, where u is the payoff of the sampled terminal history to the player who acts and $\omega > 0$ is an importance weight. At iteration one all strategies are uniform. On player i's traversal the opponent is sampled from its strategy and plays x with probability $1/2$; then I is visited, and each of the three rules adds

Algorithm 1 The parent-read rule under outcome sampling, one traversal of player i at iteration t.

1: Data: for every information set I of player i, a scalar w_I (the value of W_I at the last increment, initially 0), a row s_I (the running sum, initially 0), the current strategy $\sigma_t(I,\cdot)$ and a pointer to the parent (I', a').
2: **for** each information set I of player i on the sampled history, in order from the root **do**
3: $W \leftarrow t$ if I has no parent, else $W \leftarrow s_{I'}(a')$ ▷ the parent's increment was added earlier in this traversal
4: $s_I \leftarrow s_I + (W - w_I)\,\sigma_t(I,\cdot)$; $w_I \leftarrow W$
5: **end for**
6: Update the regrets of the visited sets and recompute their strategies for iteration $t+1$.
7: At any output time T: copy every s_I; for every I in top-down order set $W_I(T) \leftarrow T$ or the parent's completed copy $s_{I'}(a')$ and add $(W_I(T) - w_I)\,\sigma_{T+1}(I,\cdot)$ to the copy of s_I.
8: Normalize each completed copy for output; the running sums used in training are unchanged.

a positive multiple of the uniform strategy. If player i samples L then $u = 1$, the regret of L grows by $\omega/2$ and that of R falls by $\omega/2$. If it samples R then $u = -G$, the regret of R falls by $\omega G/2$ and that of L grows by $\omega G/2$. Either way $\sigma_2(I,\cdot)$ plays L with probability one. On the opponent's traversal of iteration one the opponent samples x with probability $\varepsilon/2 + (1-\varepsilon)/2 = 1/2$. Player i answers L, the payoff to the opponent is -1, the regret of x falls by $\omega/2$ and that of y grows by $\omega/2$. The opponent then plays y with probability one from iteration two. These two events have probability $1/4$ together. On player i's later traversals the opponent plays y and I is not visited, so $\sigma_t(I,\cdot)$ stays at L. On the opponent's traversals a sampled y has $u = 0$ and changes no regret, and a sampled x has $u = -1$ and $\sigma(x) = 0$, which lowers the regret of x and leaves that of y unchanged. By induction the state persists for ever, I has $K = 1$ and $t_1 = 1$, and the three running sums at I remain the multiple of the uniform strategy added at iteration one. Against the uniform strategy at I the opponent earns $(G-1)/2 > 0$ with x and zero with y, and player i earns at least zero against any strategy of the opponent by playing L. The exploitability is half the sum of these two best-response values, hence at least $(G-1)/4$. The exact average at I gives R the weight $1/(2T)$, against which x earns the opponent $-1 + (1+G)/(2T) \le 0$ once $T \ge (1+G)/2$. The opponent's exact average gives x the weight $1/(2T)$, against which player i earns $1/(2T)$ with L, and the exploitability is $1/(4T)$. □

If the opponent traverses first, the same state is reached by the end of iteration two with probability at least $1/8$: the opponent samples y at iteration one and changes nothing, player i then visits I, and the opponent samples x at iteration two. In the run of the proposition $w_{\mathrm{tail}}/W_I(T) = (T-1)/T$, so the bound $2\,w_{\mathrm{tail}}/W_I(T)$ of Theorem 1(b) is attained up to the factor two and does not vanish. Rules that add on the opponent's traversal visit I whenever the exploring opponent samples x, and Theorem 4 applies to them.

Remark 4 (the rate is attained)**.** *The per-set rule attains the rate of Corollary 1. Consider one information set with two actions, $\pi_t = 1$, visit probability $q \in (0,1)$ and a fixed input sequence of iterates alternating between $(1,0)$ and $(0,1)$. The per-set rule gives $\hat{S}(0) = \frac{1}{q}\sum_{t\ \mathrm{even}} B_t$ and $\hat{S}(1) = \frac{1}{q}\sum_{t\ \mathrm{odd}} B_t$, so*

$$\hat{\sigma}(0) - \frac{1}{2} = \frac{\hat{S}(0)}{\hat{S}(0)+\hat{S}(1)} - \frac{1}{2} = \frac{\hat{S}(0)-\hat{S}(1)}{2\big(\hat{S}(0)+\hat{S}(1)\big)}.$$

The numerator has standard deviation exactly $\sqrt{T(1-q)/q}$. The sum $\hat{S}(0)+\hat{S}(1)$ has mean T and variance $T(1-q)/q$, so by Chebyshev's inequality it lies within a factor two of T with probability at least $1 - 4(1-q)/(qT)$. On that event the ratio is at least $|\hat{S}(0)-\hat{S}(1)|/(6T)$, and the central limit theorem for the numerator gives $\mathbb{E}|\hat{\sigma}(0) - \frac{1}{2}| \ge c\sqrt{(1-q)/(qT)}$ for a universal $c > 0$ and all $T \ge T_0(q)$. With constant iterates the noise cancels in the normalization; the noise the average keeps is in the relative weighting of iterates that play different actions.

Algorithm 1 states the rule as implemented, for one player's traversal of one iteration.

B EXPERIMENTAL SETUP AND METRICS

B.1 GAMES

The four games of Table 2 are the OpenSpiel 1.6.11 games `kuhn_poker` (12 information sets, payoffs in $[-2, 2]$), `leduc_poker` (936, payoffs in $[-13, 13]$), `liars_dice(numdice=1,dice_sides=4)` (1,024, payoffs ± 1, bidding rule reset-face) and `goofspiel(num_cards=4,imp_info=True,points_order=random)` wrapped as a turn-based game (3,608, win-loss payoffs ± 1). The two other instances are `liars_dice(numdice=1,dice_sides=6)` (24,576 information sets) and `goofspiel(num_cards=5,imp_info=True,points_order=descending)` (2,124, no chance node). Goofspiel(6) of the published comparison is `goofspiel(num_cards=6,imp_info=True,points_order=descending)` (34,482). The largest instance is `goofspiel(num_cards=5,imp_info=True,points_order=random)` (236,450), wrapped like the other Goofspiel instances in `turn_based_simultaneous_game`. Liar's Dice(6) has the information-set count of Bluff(1,1) in Lanctot (2013), and it uses the reset-face bidding rule of OpenSpiel. All games are two-player and zero-sum.

B.2 ITERATIONS, CHECKPOINTS AND SEEDS

One outcome-sampling iteration is one sampled history per player. The traversing player explores with the mixture $\varepsilon\,\mathrm{unif} + (1-\varepsilon)\sigma$. Under one-sided exploration the opponent is sampled from its current strategy; under two-sided exploration both players use the same ε (Table 4). At an intermediate evaluation we add the tail of the parent-read rule to a copy of the running sums, so every checkpoint reports S_I^t at that iteration. Every cell of a gap or exploitability table has thirty seeds (0–29), except in Table 13, which uses the first twelve. One generator per run, seeded with the seed, drives both players' traversals and chance. The per-information-set diagnostics of Figure 3 use the $\varepsilon = 0.6$ runs of all thirty seeds, which log every visit. Standard deviations are sample standard deviations over seeds.

B.3 RUNS AND EVALUATION

The main grid runs 10^5 iterations per configuration with exploitability computed exactly every 10^3 iterations. A valid floor for the sampling probability under two-sided exploration is $(\varepsilon/\max_I |A(I)|)^L$ times the smallest chance probability of a terminal history, with L the largest number of decisions of both players on a terminal history. The regret bound of Lanctot et al. (2009) improves as δ grows toward uniform sampling. For the per-history rule, a history h is sampled with probability $q_t(h)$ and then adds $\pi_i^{\sigma_t}(I)\,\sigma_t(I,\cdot)/q_t(h)$, so each of the $|I|$ histories contributes $\pi_i^{\sigma_t}(I)\,\sigma_t(I,\cdot)$ in expectation. The horizon experiment runs 10^6 iterations on Leduc poker with evaluation every 10^4. On each of the three larger games, twelve of the thirty seeds run to 10^7 with evaluation every 10^5. The per-set rate experiment runs 10^6 iterations under two-sided exploration over thirty seeds and records, at seven log-spaced horizons, every information set visited by then.

B.4 ADDING ON THE OPPONENT'S TRAVERSAL

Every opponent-traversal column comes from a pass that maintains the running sums on both traversals for the optimistic and stochastic rules. For seeds 0–4 of Tables 2 and 11 and each seed of the 10^6 rows of Table 14 a pass that maintains only the own-traversal sums replays the same trajectory. The running sums added on the player's own traversal agree between the two passes to machine precision. On the opponent's traversal a rule adds at a set of player i using player i's own reach and the strategy then in force. $q_t(h)$ is then the sampling probability of that traversal under chance, the traversing player's mixture and player i on-policy. $q_t(I)$ is its sum over I. The last-visit iteration of the optimistic rule on the opponent's traversal counts only visits on that traversal. For the player who moved first the added strategy is the one after the update of that iteration. These are alternating updates, for which Burch et al. (2019) repair the CFR+ error bound. At that player's sets the running sums added on the opponent's traversal index the strategy one update later than (1) does.

B.5 External sampling

The external-sampling runs use 10^5 iterations with evaluation every 10^4. One external-sampling iteration is one traversal per traversing player. Chance and the opponent are sampled from their current strategies and all of the traversing player's actions are expanded, with no exploration parameter and no importance weight. OpenSpiel's external-sampling rule, of simple averaging type, adds the current strategy at weight one to an opponent's information set. For the rules attached to the traversing player's own information sets under external sampling, $q_t(h)$ is the reach of h under chance and the opponent. $q_t(I)$ is its sum over I.

B.6 Setting of Lanctot (2013)

We run the setting of Lanctot (2013) with $\varepsilon \in \{0.5, 0.6, 0.8\}$. Lanctot (2013) finds $0.5 \leq \varepsilon \leq 0.8$ a good interval for alternating updates and calls $\varepsilon = 0.6$ a good choice overall. We run 10^7 iterations on Liar's Dice(6) and 5×10^6 on Goofspiel(6). Evaluation is at the iterations $c \cdot 10^p$ for $c \in \{1, 2, 5\}$ and $p \geq 3$ up to the horizon. We record the number of prefix histories enumerated at every evaluation. About 10^8 prefix histories are enumerated per run, the top of the axis in Lanctot (2013). The final update of Lanctot (2013) for the optimistic rule is one more increment at every information set when the average is evaluated. We apply it to the optimistic rule on the opponent's traversal. On either traversal the result is the optimistic rule plus tail.

B.7 Metrics

Exploitability is half the sum over both players of the gain of an exact best response against the other player's normalized average. The OpenSpiel evaluator computes it on the tabular policy of each rule at the same iteration of the same run. The relative gap of a rule in per cent is its exploitability divided by that of the exact average of the same run, minus one. The ℓ_1 distance of Figure 3 is $\|\hat{\sigma}(I) - \bar{\sigma}^T(I)\|_1$ averaged with weights $\pi_i^{\bar{\sigma}}(I)$ over the information sets of one depth $d(I)$ (the depth class). We average the class means over the seeds. The weight error is $|\rho_k - 1|$ averaged over the visits of an information set at which the optimistic and the parent-read increment are both positive. The class value is the mean of these per-set values with weights $\pi_i^{\bar{\sigma}}(I)$. The depth classes hold between 3 and 3360 information sets, except the depth-four class of Liar's Dice with one to four. The window length of Appendix C is the mean window $t_k - t_{k-1}$ of each depth-zero information set of Liar's Dice under one-sided exploration, averaged over the thirty seeds. The per-seed value ranges from 40 to 336, and the longest single window of a depth-zero set has a median of 24,063 iterations. The unbiasedness check compares the sampled regrets on a frozen profile with the exact counterfactual regrets and reports the worst ratio of error to standard error over all information sets with at least the minimum number of samples.

B.8 Controls and compute

On a full traversal of Kuhn poker over three hundred iterations all rules and the full-traversal solver of OpenSpiel agree to 10^{-15}. A brute-force accumulation of (1) from the stored per-iteration strategies agrees with the running sums of the parent-read rule to 2×10^{-10} at every information set and to 2×10^{-8} for the linear-weight variant $\alpha_\tau = \tau$ of Proposition 1. This holds over three thousand iterations on sampled runs of Kuhn under one-sided exploration at $\varepsilon = 0.6$ and two-sided exploration at $\varepsilon = 0.3$, and of Leduc under one-sided exploration at $\varepsilon = 0.6$. The check covers never-visited sets and sets last visited long before the final evaluation. On Kuhn and Leduc under both exploration settings, strategies are frozen after 300 warm-up iterations. The mean sampled regret increment over 50,000 traversals per traversing player agrees with the exact counterfactual regret within 3.5 standard errors at every information set sampled at least 50 times. That is every set on Kuhn under two-sided exploration, all but one per player under one-sided exploration, and 254 to 292 of the 468 sets per player on Leduc. All runs are single-threaded Python on CPU. A 10^5-iteration run of the main grid takes a median of 23 seconds on Kuhn poker and of four to nineteen minutes on the other three games. The 10^7-iteration Liar's Dice(6) runs take about an hour each and the 5×10^6-iteration Goofspiel(6) runs about three hours, of which the exact best-response evaluations take under half. The 10^7-iteration runs of the three larger games take a median of 28 to 63 minutes each.

Table 4: The two traversals and the two exploration settings of every outcome-sampling result below. An iteration contains one sampled traversal per player, and the regrets of player i are updated on player i's own traversal.

A rule adds on	One-sided exploration: only the traversing player mixes (OpenSpiel)	Two-sided exploration: both players mix
player i's own traversal (assumed unless stated otherwise)	I can stay unvisited while player i reaches it, so the stochastic rules miss those iterations (Section 4)	every history is sampled with probability at least δ, and the running sum of the per-set rule is unbiased (Section 4)
the opponent's traversal, as Gibson et al. (2012b) do	under either setting every set that player i reaches can be visited, but the strategy at I changes between the iterations at which a rule adds, so no rule is exact	

Under external sampling OpenSpiel averages at the opponent's nodes with weight one, in `external_sampling_mccfr.py` at lines 159 to 165 of version 1.6.11.

What OpenSpiel runs. In version 1.6.11 its `outcome_sampling_mccfr.py` adds `my_reach * policy / sample_reach` at line 141, inside the branch taken when the acting player is the one being updated. That is the per-history rule adding on the player's own traversal.

Control against OpenSpiel's own solver. We also run OpenSpiel's outcome-sampling solver unchanged, at its own exploration parameter of 0.6 and with one episode per player per iteration as in our implementation. After 10^5 iterations the exploitability of its average over thirty seeds is 0.0070 ± 0.0025 on Kuhn poker, 0.505 ± 0.044 on Leduc poker, 0.114 ± 0.016 on Liar's Dice and 0.147 ± 0.015 on Goofspiel. The per-history rule of our implementation gives 0.0078 ± 0.0029, 0.517 ± 0.055, 0.115 ± 0.015 and 0.146 ± 0.014 over thirty seeds. None of the four differences is significant (Welch $p \geq 0.26$).

Table 5 measures the cost of each rule on one core of an Intel Xeon Gold 6426Y over twelve repeats: the median wall time of the training loop over 10^5 iterations at $\varepsilon = 0.6$ under one-sided exploration, the time of the final pass that produces every average strategy of the rule, and the peak resident memory of the process. The two stochastic rules on the opponent's traversal are timed on their own, each in a separate training loop. Each repeat runs every rule in the same order. The ratios below are medians of the per-repeat ratios. Their quartiles lie within 1.9 percentage points of their median on Leduc poker and Goofspiel and within 3 on Liar's Dice. The loop of the parent-read rule takes 4% longer than the optimistic one on Leduc poker and Goofspiel and 1.6% less on Liar's Dice. It is within 3% of the per-history rule and 11 to 17% below the per-set rule, both adding on the opponent's traversal. The twelve repeats of a row span 1.8 to 5.1% of its median on Leduc poker and Goofspiel and 7.4 to 24.7% on Liar's Dice, which was timed when the host carried more load.

The parent-read rule stores one extra scalar (the cumulative own reach at its last write) and one parent pointer per information set, and reads the parent. At a visit this is one extra read and one subtraction, against the regret matching over $|A(I)|$ actions that the traversal already performs at that node, so the extra work is of lower order in $|A(I)|$ than the work the visit costs anyway. The percentages above are measured in our single-threaded Python loop. The traversal records the scalar at the first visit, inside the timed loop, and writes the parent's entry a few nodes earlier; the read then touches a running sum the same iteration has just used. A compiled implementation can instead pass $W_I(t)$ down the recursion as the lazy rule passes its running weight. The lazy rule stores one pending weight per action and its loop takes 1.1 to 1.2 times the time of the parent-read rule.

Table 6 repeats the measurement on Liar's Dice(6) and Goofspiel(6), where $9{,}015$ and $19{,}859$ information sets are visited. The loop of the parent-read rule takes 3 to 5% more time than the optimistic rule at 1 to 2 MB more memory. It is within 2% of the per-history rule adding on the opponent's traversal and takes 20 to 63% less time than the per-set one. Exact $q_t(I)$ costs the per-set rule 1.7 to 3.0 times the parent-read memory. The twelve repeats of a row span 0.9 to 9.1% of its median.

C Additional experiments

Appendix C.1 gives the cells of Table 2, with the per-set rule added, in exploitability rather than in per cent (Table 8). Appendix C.2 gives the exploitability of the exact average under both exploration settings (Table 9): the cost of two-sided exploration stated in Section 5.2. Appendix C.3 gives the relative gaps of every rule over the full exploration grid (Tables 11 and 12). Appendix C.4 gives the horizon experiments (Tables 13 and 14), Appendix C.5 the tail added at its exact weight and the

Table 5: Cost of each averaging rule in its own training loop (the last row times the four rules on the player's own traversal together): median wall time of 10^5 training iterations, time of the final pass and peak resident memory; lower is better; twelve repeats, one core, seed 0, $\varepsilon = 0.6$, one-sided exploration.

Rule	Leduc			Liar's Dice			Goofspiel		
	train (s)	final (s)	RSS (MB)	train (s)	final (s)	RSS (MB)	train (s)	final (s)	RSS (MB)
optimistic	12.0	0.00	51	10.2	0.00	62	15.9	0.01	55
parent-read	12.5	0.00	51	9.9	0.00	64	16.6	0.01	55
per-history	12.1	0.00	51	10.1	0.00	63	16.5	0.00	54
per-set	13.5	0.00	52	11.1	0.00	64	18.4	0.00	59
per-history, opponent's traversal	12.3	0.00	51	10.4	0.00	64	16.6	0.00	54
per-set, opponent's traversal	14.1	0.00	52	12.0	0.00	64	19.5	0.00	59
lazy rule of Lanctot (2013)	13.8	0.02	51	11.5	0.02	63	18.6	0.08	56
all four own-traversal rules	15.5	0.01	53	12.9	0.01	65	21.1	0.03	61

Table 6: Cost of the rules on the larger instances of Appendix C.10, same protocol as Table 5: median wall time of 10^5 training iterations, time of the final pass and peak resident memory (lower is better) over twelve repeats.

Rule	Liar's Dice(6)			Goofspiel(6)		
	train (s)	final (s)	RSS (MB)	train (s)	final (s)	RSS (MB)
optimistic	8.8	0.04	63	20.3	0.10	84
parent-read	9.2	0.03	64	21.2	0.06	86
per-history, opponent's traversal	9.2	0.01	62	20.9	0.03	81
per-set, opponent's traversal	11.6	0.01	111	56.9	0.03	250

weight error by depth (Figure 3), Appendix C.6 the weighted averages (Table 15), and Appendix C.8 the other instances (Table 17). Appendix C.9 gives external sampling (Table 18) and Appendix C.10 the setting of Lanctot (2013) (Table 19). Appendix C.7 gives the largest instance (Table 16) and Appendix C.11 the rate of the per-set rule at a set. Appendix C.12 gives what every rule returns when it adds on the opponent's traversal, and Appendix C.14 the lazy rule of Lanctot (2013) as written. Appendix C.17 gives the interval of the mean, the range over seeds and the count of worse seeds for cells of Tables 2, 11, 14, 18 and 19 that the text describes by a seed count. On Kuhn poker the per-history rule's cell of Table 2 is noisy: the interval of its mean is $[+1.7, +21.9]$, yet it is the worse rule on only seventeen of thirty seeds.

C.1 The cells of the main table in exploitability

The relative gap of Table 2 is a mean of per-seed ratios, so it does not reproduce a ratio of the means and it hides how large the exploitabilities themselves are. Table 8 gives the same cells of the same runs, with the per-set rule added, in exploitability. On Kuhn poker every rule is within one standard deviation of the exact average, and on Goofspiel the per-history rule on the player's own traversal ends about half again as exploitable as the exact average.

C.2 At $\varepsilon \geq 0.6$ two-sided exploration costs the exact average a factor of 1.4 to 3.3 on the three larger games

Table 9 keeps the protocol of Table 2 and reports the exploitability of the exact average rather than a gap. The two-sided values are 1.4 to 1.8 times the one-sided ones at $\varepsilon = 0.6$ and 2.2 to 3.3 times at $\varepsilon = 1.0$ on Leduc, Liar's Dice and Goofspiel.

C.3 The gaps hold across the exploration grid and shrink on the three larger games when both players explore

Table 10 repeats the cells of Table 2, with the per-set rule added, with both players exploring, at the same exploration parameter, seed count and iteration count; on the four games of Table 2 the lazy rule is run under one-sided exploration only and has no column there. The full exploration grid behind Table 2 is in Tables 11 and 12: the gaps hold across the grid under one-sided exploration and shrink to a few percent on the three larger games under two-sided exploration. The exploration parameter runs over the whole grid and everything else is the protocol of Table 2. In Table 11 the own-traversal rules are positive in every cell of Leduc, Liar's Dice and Goofspiel under one-sided exploration, except the stochastic per-history rule on Leduc at $\varepsilon = 0.1$ (-0.3 ± 4.7). In Table 12 the cells with $\varepsilon \geq 0.5$ do not exceed 2.9% on the three larger games. The rows marked opp. in Table 11

Table 7: Outcome of each robustness variant of Table 2 and the part that reports it.

Variant	Outcome	Reported in
Two-sided exploration	the gaps close on the three larger games, and at $\varepsilon \geq 0.6$ the exact average there becomes 1.4 to 3.3 times as exploitable	Appendices C.2 and C.3, Tables 9 and 12
Other instances	the gaps persist, and on Goofspiel(5) and Goofspiel(6) the weight error is the larger of the two	Appendix C.8, Table 17; Goofspiel(6) in the text
External sampling	OpenSpiel's rule, which adds on the opponent's traversal, is below the exact average where the own-traversal rules are above it	Appendix C.9, Table 18
Instances of Lanctot (2013)	the exact average matches the best rule reported there	Appendix C.10, Table 19
Lazy rule	it comes close to the exact average on the three larger games, at 1.1 to 1.2 times the parent-read training time	Appendix C.14, Table 5
CFR+ and discounted CFR	the largest gap falls by roughly three quarters and each larger game keeps its worst rule	Appendix C.6, Table 15
The largest instance	on the player's own traversal the optimistic and the two stochastic rules stay above the exact average on every seed	Appendix C.7, Table 16
A longer horizon	on all three larger games the gaps widen up to 10^7 iterations	Appendix C.4, Tables 13 and 14

Table 8: The cells of Table 2, with the per-set rule added, in exploitability rather than in per cent (lower is better); one-sided exploration, thirty seeds, $\varepsilon = 0.6$, 10^5 iterations, $\pm$ the sample standard deviation. The standard rules are the optimistic rule and the per-history one on the player's own traversal.

Rule	Kuhn	Leduc	Liar's Dice	Goofspiel
own traversal				
optimistic	$0.0071_{\pm 0.0021}$	$0.528_{\pm 0.047}$	$0.114_{\pm 0.016}$	$0.141_{\pm 0.012}$
optimistic + tail	$0.0070_{\pm 0.0021}$	$0.487_{\pm 0.042}$	$0.107_{\pm 0.014}$	$0.117_{\pm 0.010}$
per-set	$0.0071_{\pm 0.0021}$	$0.514_{\pm 0.055}$	$0.109_{\pm 0.015}$	$0.141_{\pm 0.013}$
per-history	$0.0078_{\pm 0.0029}$	$0.517_{\pm 0.055}$	$0.115_{\pm 0.015}$	$0.146_{\pm 0.014}$
both				
lazy (Lanctot)	$0.0077_{\pm 0.0031}$	$0.464_{\pm 0.035}$	$0.100_{\pm 0.012}$	$0.100_{\pm 0.007}$
opponent's traversal				
optimistic + tail	$0.0070_{\pm 0.0021}$	$0.480_{\pm 0.033}$	$0.101_{\pm 0.012}$	$0.103_{\pm 0.008}$
per-set	$0.0070_{\pm 0.0019}$	$0.467_{\pm 0.035}$	$0.097_{\pm 0.011}$	$0.095_{\pm 0.008}$
per-history	$0.0069_{\pm 0.0020}$	$0.467_{\pm 0.035}$	$0.098_{\pm 0.011}$	$0.096_{\pm 0.008}$
exact average	$0.0070_{\pm 0.0021}$	$0.467_{\pm 0.033}$	$0.098_{\pm 0.011}$	$0.098_{\pm 0.008}$

come from the pass of Appendix B that maintains both traversals. There the stochastic rules and the optimistic rule add on the opponent's traversal. In these rows the stochastic rules stay within 2.8% above or below the exact average at every ε on the three larger games. The optimistic rule plus tail is 0.8 to 6.4% worse there.

C.4 Under one-sided exploration the gap widens out to 10^7 iterations on all three larger games

Leduc poker runs to 10^6 iterations over all thirty seeds, and under one-sided exploration the relative gap does not close there (Table 14). The columns of the rules adding on the opponent's traversal come from the pass of Appendix B that maintains both traversals and are filled under one-sided exploration only; a dash marks their two-sided cells, and the 10^5 ones at $\varepsilon = 0.6$ are in Table 10. In the one-sided rows the stochastic rules stay within 0.2% of the exact average at 10^6, while the optimistic rule plus tail is 2.9 to 4.4% worse.

Twelve of the thirty seeds of each of the three larger games run on to 10^7 iterations with the same driver, evaluated every 10^5 iterations (Table 13). At 10^5 iterations all 36 runs give exploitabilities equal to those of the same seeds in Table 2, so each longer run is the shorter one continued. On Leduc poker a second set of runs of the same twelve seeds, evaluated every 5×10^5 iterations, matches these runs to the last bit at all 2,160 shared values. For both standard rules the mean gap grows over each decade on all three games, and it is larger at 10^7 than at 10^6 on every seed except one seed of the optimistic rule on Leduc poker. The per-history rule on the opponent's traversal stays within 2.3% of the exact average in the mean at every checkpoint on Leduc poker and Liar's Dice, and on Goofspiel it stays below the exact average by at most 5.1%. We fit the exponent of T to the exploitability over 10^6 to 10^7, seed by seed. The checkpoints are evenly spaced in T, so the fit weights the upper half of the decade most. A slope fitted over one decade is local and not an

Table 9: Exploitability of the exact average after 10^5 iterations (lower is better). Mean and sample s.d. over thirty seeds for every game, exploration parameter and exploration setting.

Game	Exploration	$\varepsilon = 0.1$	$\varepsilon = 0.3$	$\varepsilon = 0.5$	$\varepsilon = 0.6$
Kuhn	one-sided	$0.010_{\pm 0.004}$	$0.008_{\pm 0.004}$	$0.007_{\pm 0.003}$	$0.007_{\pm 0.002}$
	two-sided	$0.011_{\pm 0.006}$	$0.009_{\pm 0.005}$	$0.009_{\pm 0.003}$	$0.008_{\pm 0.003}$
Leduc	one-sided	$0.720_{\pm 0.063}$	$0.566_{\pm 0.045}$	$0.473_{\pm 0.032}$	$0.467_{\pm 0.033}$
	two-sided	$0.774_{\pm 0.081}$	$0.650_{\pm 0.052}$	$0.647_{\pm 0.046}$	$0.683_{\pm 0.058}$
Liar's Dice	one-sided	$0.176_{\pm 0.032}$	$0.120_{\pm 0.018}$	$0.106_{\pm 0.018}$	$0.098_{\pm 0.011}$
	two-sided	$0.190_{\pm 0.028}$	$0.146_{\pm 0.022}$	$0.156_{\pm 0.017}$	$0.175_{\pm 0.023}$
Goofspiel	one-sided	$0.269_{\pm 0.034}$	$0.149_{\pm 0.016}$	$0.102_{\pm 0.008}$	$0.098_{\pm 0.008}$
	two-sided	$0.281_{\pm 0.031}$	$0.179_{\pm 0.019}$	$0.160_{\pm 0.013}$	$0.168_{\pm 0.009}$

Game	Exploration	$\varepsilon = 0.7$	$\varepsilon = 0.9$	$\varepsilon = 1.0$
Kuhn	one-sided	$0.008_{\pm 0.003}$	$0.006_{\pm 0.002}$	$0.006_{\pm 0.003}$
	two-sided	$0.009_{\pm 0.003}$	$0.010_{\pm 0.004}$	$0.009_{\pm 0.003}$
Leduc	one-sided	$0.454_{\pm 0.030}$	$0.470_{\pm 0.031}$	$0.488_{\pm 0.027}$
	two-sided	$0.739_{\pm 0.059}$	$0.955_{\pm 0.084}$	$1.114_{\pm 0.133}$
Liar's Dice	one-sided	$0.104_{\pm 0.013}$	$0.121_{\pm 0.011}$	$0.137_{\pm 0.013}$
	two-sided	$0.218_{\pm 0.034}$	$0.309_{\pm 0.046}$	$0.378_{\pm 0.040}$
Goofspiel	one-sided	$0.091_{\pm 0.005}$	$0.097_{\pm 0.005}$	$0.109_{\pm 0.004}$
	two-sided	$0.188_{\pm 0.013}$	$0.270_{\pm 0.017}$	$0.358_{\pm 0.024}$

Table 10: Relative gap to the exact average of the same run, in per cent; positive is more exploitable. Two-sided exploration, thirty seeds, $\varepsilon = 0.6$, 10^5 iterations, $\pm$ the sample standard deviation. $*$: the standard rules. †: the per-set rule, an idealized baseline. Table 2 is the same measurement with one-sided exploration.

Game	own traversal				opponent's traversal			exact average (absolute expl.)
	optimistic*	optimistic + tail	per-set†	per-history*	optimistic + tail	per-set†	per-history	
Kuhn	$+0.1_{\pm 0.3}$	$+0.1_{\pm 0.3}$	$+3.4_{\pm 10.8}$	$+3.9_{\pm 11.2}$	$+0.1_{\pm 0.5}$	$-0.3_{\pm 8.6}$	$-0.2_{\pm 9.6}$	$0.008_{\pm 0.003}$
Leduc	$+0.5_{\pm 0.4}$	$+0.3_{\pm 0.3}$	$+0.1_{\pm 0.7}$	$+0.0_{\pm 0.9}$	$-0.1_{\pm 0.4}$	$-0.0_{\pm 0.6}$	$+0.0_{\pm 0.7}$	$0.683_{\pm 0.058}$
Liar's Dice	$+1.9_{\pm 1.4}$	$+1.5_{\pm 1.3}$	$+1.1_{\pm 1.4}$	$+1.4_{\pm 1.8}$	$+0.7_{\pm 1.0}$	$-0.0_{\pm 1.5}$	$+0.1_{\pm 1.8}$	$0.175_{\pm 0.023}$
Goofspiel	$+1.5_{\pm 0.9}$	$+1.3_{\pm 0.9}$	$+2.1_{\pm 2.0}$	$+2.5_{\pm 2.3}$	$-0.1_{\pm 1.0}$	$-2.0_{\pm 1.9}$	$-1.2_{\pm 2.1}$	$0.168_{\pm 0.009}$

asymptotic rate. Over this decade the exponents of the standard rules lie between -0.11 and -0.42 and those of the exact average between -0.56 and -0.77, and the per-history rule on the opponent's traversal matches the exact average on each game.

C.5 On the three larger games the exact tail weight, read from the parent, more than halves the gap of the optimistic rule

On depth-zero information sets the optimistic rule plus tail coincides with the parent-read rule. The exact cumulative tail weight $W_I(T) - W_I(t_K)$ is available from the running sums of the parent-read rule. Added with this weight instead of the reach in force at the end times the tail length, the tail leaves the optimistic rule 2.4%, 7.2% and 13.7% above the exact average on Leduc, Liar's Dice and Goofspiel at $\varepsilon = 0.6$. The standard deviations are 1.5, 4.3 and 7.2 points. The optimistic rule plus tail of Table 2 ends 4.3, 9.9 and 19.6% above. On Liar's Dice the depth-zero sets have $\rho_k \equiv 1$ and the optimistic rule plus tail is exact there. The optimistic rule has ℓ_1 distance 0.07, because the on-policy opponent leaves those sets unvisited for windows of mean length 106 iterations.

Figure 3 stratifies both the weight error and the ℓ_1 distance by depth. Two-sided exploration visits every set with probability at least δ in each iteration, and there the reach-weighted ℓ_1 distance of the optimistic rule on the three larger games is 6.2 to 8.1 times smaller. The weight error increases through the available depths on Kuhn poker and Goofspiel and through depth three on Leduc poker and Liar's Dice. Under two-sided exploration the weight error of the optimistic rule is not always smaller, as the depth-two sets of Goofspiel show. There frequent visits keep the strategies of neighbouring windows close, so misweighting them moves the average less.

C.6 Under CFR+ and discounted CFR the gaps shrink and each larger game keeps its worst rule

CFR+ gives iteration t the weight t and floors the regrets after each iteration at the information set; discounted CFR, at the parameters of Brown & Sandholm (2019a), gives it t^2 and multiplies

Table 11: Relative exploitability gap (%) to the exact average under one-sided exploration after 10^5 iterations (lower is better); mean and sample s.d. over thirty seeds. In rows marked opp. the rules add on the opponent's traversal.

Game	Rule	$\varepsilon = 0.1$	$\varepsilon = 0.3$	$\varepsilon = 0.5$	$\varepsilon = 0.6$
Kuhn	optimistic	$+0.7_{\pm 1.9}$	$-0.1_{\pm 10.2}$	$+0.3_{\pm 4.8}$	$+0.9_{\pm 4.2}$
	optimistic + tail	$+0.4_{\pm 0.9}$	$+0.1_{\pm 1.0}$	$+0.0_{\pm 0.4}$	$+0.1_{\pm 0.3}$
	per-set	$+13.2_{\pm 73.4}$	$+13.8_{\pm 33.4}$	$+8.0_{\pm 31.3}$	$+1.8_{\pm 14.5}$
	per-history	$+13.4_{\pm 68.6}$	$+32.5_{\pm 54.2}$	$+20.0_{\pm 39.2}$	$+11.8_{\pm 27.0}$
	optimistic + tail, opp.	$-0.6_{\pm 5.1}$	$-0.6_{\pm 1.8}$	$-0.2_{\pm 1.1}$	$-0.0_{\pm 0.9}$
	per-set, opp.	$+0.4_{\pm 13.7}$	$+2.6_{\pm 9.1}$	$-1.4_{\pm 12.4}$	$-0.0_{\pm 8.6}$
	per-history, opp.	$-1.7_{\pm 12.9}$	$+1.3_{\pm 14.1}$	$-4.3_{\pm 12.7}$	$-0.2_{\pm 9.1}$
Leduc	optimistic	$+5.3_{\pm 3.2}$	$+8.7_{\pm 5.6}$	$+11.4_{\pm 9.8}$	$+13.2_{\pm 7.2}$
	optimistic + tail	$+2.0_{\pm 1.7}$	$+3.1_{\pm 2.4}$	$+3.5_{\pm 2.4}$	$+4.3_{\pm 3.0}$
	per-set	$+0.5_{\pm 4.2}$	$+3.1_{\pm 4.5}$	$+8.0_{\pm 9.2}$	$+10.1_{\pm 8.9}$
	per-history	$-0.3_{\pm 4.7}$	$+3.6_{\pm 5.0}$	$+8.1_{\pm 9.4}$	$+10.8_{\pm 8.8}$
	optimistic + tail, opp.	$+1.5_{\pm 1.2}$	$+2.4_{\pm 2.2}$	$+2.4_{\pm 2.3}$	$+2.8_{\pm 1.7}$
	per-set, opp.	$-0.1_{\pm 0.9}$	$-0.0_{\pm 0.8}$	$-0.1_{\pm 0.8}$	$+0.1_{\pm 0.8}$
	per-history, opp.	$+0.5_{\pm 1.9}$	$+0.1_{\pm 1.2}$	$-0.1_{\pm 1.0}$	$-0.0_{\pm 0.9}$
Liar's Dice	optimistic	$+7.7_{\pm 7.0}$	$+12.9_{\pm 13.6}$	$+16.2_{\pm 8.9}$	$+16.4_{\pm 10.5}$
	optimistic + tail	$+3.4_{\pm 3.1}$	$+8.3_{\pm 6.6}$	$+10.8_{\pm 4.6}$	$+9.9_{\pm 5.8}$
	per-set	$+1.3_{\pm 7.8}$	$+5.9_{\pm 12.9}$	$+8.6_{\pm 8.9}$	$+11.5_{\pm 10.1}$
	per-history	$+0.8_{\pm 10.3}$	$+9.6_{\pm 13.8}$	$+11.2_{\pm 10.3}$	$+18.0_{\pm 14.0}$
	optimistic + tail, opp.	$+1.9_{\pm 1.8}$	$+3.0_{\pm 2.6}$	$+3.7_{\pm 4.1}$	$+3.6_{\pm 3.0}$
	per-set, opp.	$-0.0_{\pm 2.2}$	$-0.1_{\pm 2.1}$	$-0.4_{\pm 1.8}$	$-0.1_{\pm 2.6}$
	per-history, opp.	$+1.4_{\pm 4.5}$	$-0.3_{\pm 3.8}$	$-0.2_{\pm 2.2}$	$+0.1_{\pm 3.0}$
Goofspiel	optimistic	$+10.3_{\pm 8.5}$	$+30.9_{\pm 14.7}$	$+45.5_{\pm 13.4}$	$+45.3_{\pm 14.0}$
	optimistic + tail	$+3.3_{\pm 4.0}$	$+12.3_{\pm 7.4}$	$+20.4_{\pm 8.3}$	$+19.6_{\pm 8.7}$
	per-set	$+3.0_{\pm 11.2}$	$+26.4_{\pm 14.0}$	$+42.0_{\pm 11.8}$	$+44.9_{\pm 12.6}$
	per-history	$+3.9_{\pm 11.9}$	$+29.3_{\pm 14.2}$	$+47.0_{\pm 14.3}$	$+50.1_{\pm 13.8}$
	optimistic + tail, opp.	$+1.1_{\pm 2.2}$	$+3.4_{\pm 2.1}$	$+4.8_{\pm 3.7}$	$+5.6_{\pm 3.6}$
	per-set, opp.	$-0.9_{\pm 2.2}$	$-1.4_{\pm 2.1}$	$-1.9_{\pm 1.9}$	$-2.7_{\pm 2.7}$
	per-history, opp.	$+0.6_{\pm 4.1}$	$-0.2_{\pm 3.1}$	$-1.3_{\pm 2.0}$	$-1.9_{\pm 3.0}$

Game	Rule	$\varepsilon = 0.7$	$\varepsilon = 0.9$	$\varepsilon = 1.0$
Kuhn	optimistic	$-1.4_{\pm 6.6}$	$+0.7_{\pm 7.1}$	$+4.7_{\pm 22.8}$
	optimistic + tail	$+0.1_{\pm 0.3}$	$+0.2_{\pm 0.4}$	$-0.0_{\pm 0.4}$
	per-set	$+3.4_{\pm 20.0}$	$+10.1_{\pm 52.7}$	$+21.3_{\pm 40.6}$
	per-history	$+8.4_{\pm 26.4}$	$+24.2_{\pm 59.0}$	$+30.0_{\pm 47.7}$
	optimistic + tail, opp.	$+0.1_{\pm 1.0}$	$+0.6_{\pm 1.8}$	$-0.4_{\pm 1.0}$
	per-set, opp.	$+0.3_{\pm 11.6}$	$+1.2_{\pm 13.5}$	$+6.6_{\pm 17.3}$
	per-history, opp.	$+0.3_{\pm 11.2}$	$+1.3_{\pm 13.7}$	$+6.6_{\pm 17.3}$
Leduc	optimistic	$+13.0_{\pm 7.6}$	$+12.2_{\pm 6.9}$	$+14.3_{\pm 7.0}$
	optimistic + tail	$+3.3_{\pm 2.5}$	$+3.6_{\pm 2.9}$	$+2.8_{\pm 1.9}$
	per-set	$+9.6_{\pm 6.9}$	$+10.3_{\pm 8.3}$	$+12.4_{\pm 8.1}$
	per-history	$+10.6_{\pm 7.6}$	$+12.0_{\pm 8.4}$	$+13.6_{\pm 8.5}$
	optimistic + tail, opp.	$+2.2_{\pm 2.3}$	$+1.9_{\pm 1.7}$	$+1.5_{\pm 2.0}$
	per-set, opp.	$-0.3_{\pm 0.8}$	$-0.2_{\pm 0.8}$	$+0.1_{\pm 0.9}$
	per-history, opp.	$-0.2_{\pm 1.0}$	$-0.3_{\pm 0.8}$	$+0.1_{\pm 0.9}$
Liar's Dice	optimistic	$+19.8_{\pm 9.8}$	$+14.9_{\pm 7.0}$	$+10.9_{\pm 8.0}$
	optimistic + tail	$+12.4_{\pm 6.9}$	$+9.2_{\pm 5.7}$	$+6.8_{\pm 4.4}$
	per-set	$+13.8_{\pm 8.9}$	$+10.7_{\pm 7.9}$	$+7.8_{\pm 7.0}$
	per-history	$+18.3_{\pm 10.6}$	$+13.0_{\pm 8.5}$	$+12.3_{\pm 9.2}$
	optimistic + tail, opp.	$+3.8_{\pm 3.5}$	$+2.3_{\pm 2.2}$	$+0.8_{\pm 3.0}$
	per-set, opp.	$+0.2_{\pm 2.1}$	$+0.4_{\pm 1.9}$	$-0.6_{\pm 1.9}$
	per-history, opp.	$+0.4_{\pm 2.4}$	$+0.5_{\pm 1.8}$	$-0.6_{\pm 1.9}$
Goofspiel	optimistic	$+50.1_{\pm 17.0}$	$+51.0_{\pm 14.2}$	$+39.6_{\pm 8.5}$
	optimistic + tail	$+20.5_{\pm 9.3}$	$+19.8_{\pm 9.1}$	$+15.6_{\pm 7.0}$
	per-set	$+48.4_{\pm 14.5}$	$+48.9_{\pm 11.0}$	$+43.4_{\pm 10.3}$
	per-history	$+56.9_{\pm 14.0}$	$+54.6_{\pm 11.5}$	$+48.9_{\pm 10.2}$
	optimistic + tail, opp.	$+5.6_{\pm 3.4}$	$+6.4_{\pm 2.9}$	$+5.6_{\pm 3.2}$
	per-set, opp.	$-2.7_{\pm 2.5}$	$-2.6_{\pm 2.0}$	$-2.6_{\pm 2.0}$
	per-history, opp.	$-1.9_{\pm 2.3}$	$-2.2_{\pm 2.0}$	$-2.6_{\pm 2.0}$

positive and negative cumulative regrets by $t^{1.5}/(t^{1.5}+1)$ and $1/2$. Remark 1 gives the averages they target. Table 15 repeats the cells of Table 2, with the per-set rule added, at $\varepsilon = 0.6$ under one-sided exploration with thirty seeds each. On the player's own traversal the largest gap falls from 50.1% under equal weights to 13.5% under CFR+ and 10.6% under discounted CFR, both on Goofspiel; on the other three games no gap exceeds 1.4 and 2.0%. On the opponent's traversal every mean lies within 0.6 and 0.3% of the exact average. The lazy rule is within 3.1% under both weightings. Under equal weights it is $+9.9\%$ on Kuhn poker, because it drops the weight of paths that end before the player's next decision. On a full traversal our CFR+ reproduces the average policy of OpenSpiel's own CFR+ solver, to machine precision on Kuhn poker. On Leduc poker the two average policies part, because truncating negative regrets at zero amplifies the last bits

Table 12: Relative exploitability gap (%) to the exact average under two-sided exploration after 10^5 iterations (lower is better); mean and sample s.d. over thirty seeds.

Game	Rule	$\varepsilon = 0.1$	$\varepsilon = 0.3$	$\varepsilon = 0.5$	$\varepsilon = 0.6$
Kuhn	optimistic	$+0.1_{\pm 0.3}$	$+0.2_{\pm 0.6}$	$+0.1_{\pm 0.3}$	$+0.1_{\pm 0.3}$
	optimistic + tail	$+0.0_{\pm 0.3}$	$+0.1_{\pm 0.5}$	$+0.1_{\pm 0.2}$	$+0.1_{\pm 0.3}$
	per-set	$-2.3_{\pm 11.0}$	$+3.3_{\pm 10.0}$	$+0.3_{\pm 6.0}$	$+3.4_{\pm 10.8}$
	per-history	$-1.6_{\pm 10.7}$	$+4.1_{\pm 12.5}$	$+0.5_{\pm 6.8}$	$+3.9_{\pm 11.2}$
Leduc	optimistic	$+1.4_{\pm 1.1}$	$+1.0_{\pm 0.6}$	$+0.6_{\pm 0.5}$	$+0.5_{\pm 0.4}$
	optimistic + tail	$+1.1_{\pm 1.0}$	$+0.9_{\pm 0.6}$	$+0.4_{\pm 0.5}$	$+0.3_{\pm 0.3}$
	per-set	$+0.4_{\pm 0.9}$	$+0.2_{\pm 0.7}$	$+0.1_{\pm 0.8}$	$+0.1_{\pm 0.7}$
	per-history	$+1.0_{\pm 1.9}$	$+0.5_{\pm 1.3}$	$+0.3_{\pm 1.1}$	$+0.0_{\pm 0.9}$
Liar's Dice	optimistic	$+3.2_{\pm 2.0}$	$+3.1_{\pm 1.5}$	$+2.9_{\pm 1.5}$	$+1.9_{\pm 1.4}$
	optimistic + tail	$+2.7_{\pm 1.6}$	$+2.9_{\pm 1.4}$	$+2.5_{\pm 1.4}$	$+1.5_{\pm 1.3}$
	per-set	$+0.9_{\pm 2.1}$	$+0.4_{\pm 1.7}$	$+0.9_{\pm 1.9}$	$+1.1_{\pm 1.4}$
	per-history	$+3.5_{\pm 4.6}$	$+0.7_{\pm 2.5}$	$+1.2_{\pm 2.3}$	$+1.4_{\pm 1.8}$
Goofspiel	optimistic	$+1.1_{\pm 1.5}$	$+2.4_{\pm 1.7}$	$+1.6_{\pm 1.5}$	$+1.5_{\pm 0.9}$
	optimistic + tail	$+1.0_{\pm 1.3}$	$+2.2_{\pm 1.6}$	$+1.3_{\pm 1.4}$	$+1.3_{\pm 0.9}$
	per-set	$+0.0_{\pm 1.7}$	$+1.1_{\pm 2.0}$	$+2.0_{\pm 2.3}$	$+2.1_{\pm 2.0}$
	per-history	$+3.6_{\pm 2.9}$	$+2.9_{\pm 2.6}$	$+2.5_{\pm 2.6}$	$+2.5_{\pm 2.3}$

Game	Rule	$\varepsilon = 0.7$	$\varepsilon = 0.9$	$\varepsilon = 1.0$
Kuhn	optimistic	$+0.2_{\pm 0.3}$	$+0.1_{\pm 0.2}$	$+0.0_{\pm 0.2}$
	optimistic + tail	$+0.1_{\pm 0.3}$	$+0.1_{\pm 0.2}$	$+0.1_{\pm 0.2}$
	per-set	$+2.3_{\pm 12.6}$	$+1.5_{\pm 10.9}$	$+2.4_{\pm 8.0}$
	per-history	$+2.1_{\pm 13.1}$	$+1.4_{\pm 10.8}$	$+2.4_{\pm 8.0}$
Leduc	optimistic	$+0.5_{\pm 0.4}$	$+0.4_{\pm 0.3}$	$+0.5_{\pm 0.4}$
	optimistic + tail	$+0.3_{\pm 0.3}$	$+0.2_{\pm 0.2}$	$+0.2_{\pm 0.4}$
	per-set	$+0.3_{\pm 0.7}$	$+0.1_{\pm 0.7}$	$+0.4_{\pm 0.7}$
	per-history	$+0.3_{\pm 0.7}$	$+0.1_{\pm 0.8}$	$+0.4_{\pm 0.7}$
Liar's Dice	optimistic	$+1.5_{\pm 1.1}$	$+0.7_{\pm 0.9}$	$+0.6_{\pm 1.0}$
	optimistic + tail	$+1.3_{\pm 0.9}$	$+0.5_{\pm 0.8}$	$+0.2_{\pm 0.6}$
	per-set	$+0.8_{\pm 1.5}$	$+0.6_{\pm 1.3}$	$+0.2_{\pm 1.2}$
	per-history	$+0.7_{\pm 1.5}$	$+0.7_{\pm 1.4}$	$+0.2_{\pm 1.2}$
Goofspiel	optimistic	$+1.3_{\pm 1.3}$	$+1.1_{\pm 0.8}$	$+0.8_{\pm 0.5}$
	optimistic + tail	$+1.0_{\pm 1.1}$	$+0.7_{\pm 0.7}$	$+0.2_{\pm 0.5}$
	per-set	$+1.8_{\pm 2.2}$	$+1.5_{\pm 1.7}$	$+0.8_{\pm 1.1}$
	per-history	$+1.9_{\pm 2.1}$	$+1.5_{\pm 1.7}$	$+0.8_{\pm 1.1}$

Table 13: Twelve seeds carried to 10^7 iterations, $\varepsilon = 0.6$, one-sided exploration. Exact average: exploitability; other rules: relative gap in per cent (lower is better). Exponent: fitted over 10^6 to 10^7, with 95% interval. Work: median work, relative to the exact average's, to reach its exploitability at 10^6 (Appendix C.15); $>$: lower bound; –: not computed. $*$: standard rule; †: per-set rule.

Game	Rule	relative gap (%) at			exponent	work
		10^5	10^6	10^7	10^6 to 10^7	to reach $T = 10^6$
Leduc	exact average	0.4723	0.1248	0.0325	-0.56 $[-0.59, -0.52]$	1
	optimistic*	$+12.1$	$+29.9$	$+117.1$	-0.42 $[-0.52, -0.31]$	1.44
	optimistic + tail	$+4.9$	$+7.9$	$+31.4$	-0.50 $[-0.54, -0.45]$	–
	per-set†	$+9.6$	$+36.4$	$+173.4$	-0.29 $[-0.37, -0.21]$	–
	per-history*	$+10.6$	$+41.5$	$+199.5$	-0.27 $[-0.35, -0.20]$	1.96
	per-history, opponent's traversal	-0.1	$+0.1$	-1.0	-0.56 $[-0.60, -0.52]$	0.99
Liar's Dice	exact average	0.0966	0.0169	0.0028	-0.77 $[-0.84, -0.69]$	1
	optimistic*	$+13.9$	$+95.0$	$+444.7$	-0.31 $[-0.48, -0.14]$	3.25
	optimistic + tail	$+9.7$	$+35.9$	$+120.8$	-0.60 $[-0.77, -0.43]$	–
	per-set†	$+10.2$	$+112.5$	$+460.8$	-0.34 $[-0.49, -0.19]$	–
	per-history*	$+19.3$	$+152.1$	$+557.2$	-0.34 $[-0.48, -0.19]$	> 9.7
	per-history, opponent's traversal	-0.1	-0.5	$+0.3$	-0.76 $[-0.84, -0.68]$	0.98
Goofspiel	exact average	0.0994	0.0242	0.0068	-0.56 $[-0.59, -0.53]$	1
	optimistic*	$+44.5$	$+180.6$	$+686.1$	-0.11 $[-0.18, -0.03]$	> 9.6
	optimistic + tail	$+20.2$	$+60.9$	$+181.1$	-0.31 $[-0.41, -0.21]$	–
	per-set†	$+41.7$	$+137.6$	$+490.8$	-0.16 $[-0.22, -0.10]$	–
	per-history*	$+46.4$	$+151.8$	$+508.0$	-0.17 $[-0.23, -0.11]$	> 9.9
	per-history, opponent's traversal	-2.4	-2.6	-3.2	-0.56 $[-0.59, -0.53]$	0.95

of a different summation order: the largest difference between them grows from 10^{-15} after five iterations to 10^{-2} after two hundred. Their exploitabilities are 0.004970 and 0.004963, a difference of 0.14%.

Table 14: Leduc poker: relative exploitability gap (%) to the exact average after 10^5 and 10^6 iterations (lower is better); mean and sample s.d. over thirty seeds. Last column: exploitability of the exact average. $*$ marks the standard rules and † the per-set rule, an idealized baseline.

Iterations	ε	own traversal				opponent's traversal			exact average (absolute expl.)
		optimistic*	optimistic + tail	per-set†	per-history*	optimistic + tail	per-set†	per-history	
One-sided exploration									
10^5	0.6	$+13.2_{\pm 7.2}$	$+4.3_{\pm 3.0}$	$+10.1_{\pm 8.9}$	$+10.8_{\pm 8.8}$	$+2.8_{\pm 1.7}$	$+0.1_{\pm 0.8}$	$-0.0_{\pm 0.9}$	$0.467_{\pm 0.033}$
	1.0	$+14.3_{\pm 7.0}$	$+2.8_{\pm 1.9}$	$+12.4_{\pm 8.1}$	$+13.6_{\pm 8.5}$	$+1.5_{\pm 2.0}$	$+0.1_{\pm 0.9}$	$+0.1_{\pm 0.9}$	$0.488_{\pm 0.027}$
10^6	0.6	$+34.9_{\pm 25.0}$	$+7.8_{\pm 6.5}$	$+46.7_{\pm 26.7}$	$+53.3_{\pm 29.5}$	$+4.4_{\pm 4.4}$	$+0.2_{\pm 1.1}$	$+0.1_{\pm 1.2}$	$0.121_{\pm 0.009}$
	1.0	$+39.7_{\pm 25.4}$	$+8.1_{\pm 6.7}$	$+50.0_{\pm 32.2}$	$+58.1_{\pm 33.8}$	$+2.9_{\pm 2.5}$	$+0.1_{\pm 1.3}$	$+0.1_{\pm 1.3}$	$0.124_{\pm 0.006}$
Two-sided exploration									
10^5	0.6	$+0.5_{\pm 0.4}$	$+0.3_{\pm 0.3}$	$+0.1_{\pm 0.7}$	$+0.0_{\pm 0.9}$	–	–	–	$0.683_{\pm 0.058}$
	1.0	$+0.5_{\pm 0.4}$	$+0.2_{\pm 0.4}$	$+0.4_{\pm 0.7}$	$+0.4_{\pm 0.7}$	–	–	–	$1.114_{\pm 0.133}$
10^6	0.6	$+0.2_{\pm 0.2}$	$+0.2_{\pm 0.2}$	$+0.1_{\pm 0.8}$	$+0.1_{\pm 1.0}$	–	–	–	$0.217_{\pm 0.026}$
	1.0	$+0.1_{\pm 0.1}$	$+0.1_{\pm 0.1}$	$+0.2_{\pm 0.5}$	$+0.2_{\pm 0.5}$	–	–	–	$0.408_{\pm 0.035}$

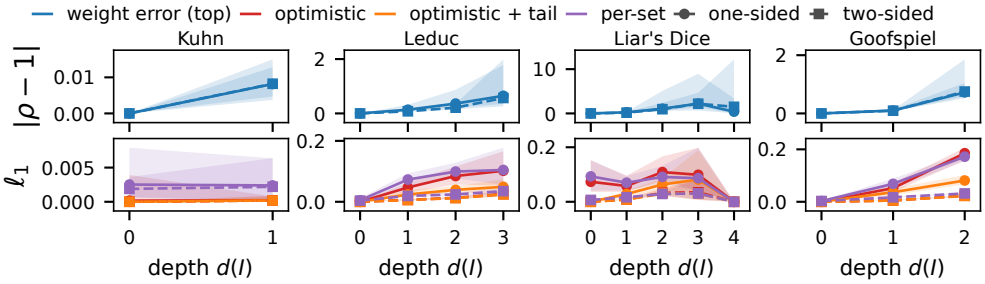


Figure 3: The weight error of the optimistic rule is zero at depth zero and grows with depth up to depth three. Top: weight error $|\rho_k - 1|$ over visits where both rules add positive weight. Bottom: ℓ_1 distance to the exact average. Lines: reach-weighted means per depth over thirty seeds; shaded: the range over the seeds; the scales differ.

C.7 The own-traversal rules stay the worse ones on a game seven times larger

Goofspiel(5) with a random point order has 236,450 information sets and 3.3×10^6 histories, against 34,482 and 9.7×10^5 for Goofspiel(6) of Appendix C.10. One exact best response there takes about three and a half minutes. Each run is evaluated once, at the end, and the horizon is 10^6 rather than 10^5. In three separate runs of 10^5 iterations, all rules lie within 3.1% of the exact average. At 10^6 the rules on the player's own traversal lie above those on the opponent's, as in Table 2. On the player's own traversal the optimistic rule is 10.9% more exploitable than the exact average, the per-set rule 11.5% and the per-history rule that OpenSpiel runs 14.9%, each of them on all thirty seeds. The lazy rule is 1.5% worse. On the opponent's traversal the two stochastic rules are 0.8 and 0.2% below the exact average and the optimistic rule plus tail is 1.6% above it. Two-sided exploration closes the gap here as it does on the three larger games of Table 2: no rule is then more than 1.2% from the exact average on either traversal. The exact average itself is 1.8 times as exploitable at the same iteration count.

C.8 The gaps persist on two other instances, and on Goofspiel(5) and Goofspiel(6) the weight error is the larger part

Liar's Dice(6) and Goofspiel(5) with a fixed point order replace the four games of Table 2. They have 24,576 and 2,124 information sets, and the second has a longer horizon of decisions than the Goofspiel of Table 2. There the gaps of Table 2 persist under one-sided exploration and close under two-sided exploration. Table 17 reports them. In the rows under one-sided exploration at $\varepsilon = 0.6$ the optimistic rule is 38.5% worse than the exact average and the stochastic rules 44.9–53.5% on Goofspiel(5); on Liar's Dice(6) the gaps are 9.1% and 5.2–8.1%. The rows under two-sided exploration lie within 1% of the exact average. Adding the tail at its exact weight to the running sum of the optimistic rule (Appendix C.5) brings its gap at $\varepsilon = 0.6$ from 9.1 to 2.8% on Liar's Dice(6), from 38.5 to 29.7% on Goofspiel(5) and from 9.7 to 6.7% on the Goofspiel(6) of

Table 15: Weighted averages, relative gap (%) to the exact average of the same run under the same weighting, thirty seeds, $\varepsilon = 0.6$, one-sided exploration, 10^5 iterations. Compare Table 2, which is the same measurement with equal weights. $*$ marks the standard rules and † the per-set rule, an idealized baseline.

Game	own traversal				both	opponent's traversal		
	optimistic*	optimistic + tail	per-set†	per-history*	lazy (Lanctot)	optimistic + tail	per-set†	per-history
CFR+								
Kuhn	$+0.0_{\pm 0.0}$	$+0.0_{\pm 0.0}$	$+0.5_{\pm 1.8}$	$+0.2_{\pm 3.4}$	$-2.8_{\pm 2.3}$	$-0.0_{\pm 0.1}$	$-0.2_{\pm 1.8}$	$-0.0_{\pm 1.9}$
Leduc	$+0.4_{\pm 0.3}$	$+0.2_{\pm 0.2}$	$+0.2_{\pm 0.4}$	$+0.1_{\pm 0.8}$	$-0.3_{\pm 0.4}$	$-0.1_{\pm 0.3}$	$+0.2_{\pm 0.3}$	$+0.2_{\pm 0.4}$
Liar's Dice	$+1.3_{\pm 0.8}$	$+0.5_{\pm 0.5}$	$+0.5_{\pm 1.5}$	$+1.4_{\pm 2.2}$	$-0.1_{\pm 2.4}$	$-0.2_{\pm 0.6}$	$+0.0_{\pm 0.8}$	$+0.2_{\pm 1.0}$
Goofspiel	$+6.0_{\pm 5.0}$	$+0.6_{\pm 0.7}$	$+7.1_{\pm 5.4}$	$+13.5_{\pm 5.9}$	$+0.2_{\pm 2.1}$	$+0.3_{\pm 0.5}$	$-0.5_{\pm 1.3}$	$-0.1_{\pm 1.4}$
discounted CFR								
Kuhn	$-0.0_{\pm 0.0}$	$+0.0_{\pm 0.0}$	$-0.5_{\pm 2.9}$	$-0.0_{\pm 4.0}$	$-3.1_{\pm 3.9}$	$-0.0_{\pm 0.1}$	$-0.3_{\pm 2.2}$	$-0.2_{\pm 2.3}$
Leduc	$+0.4_{\pm 0.2}$	$+0.1_{\pm 0.1}$	$+0.0_{\pm 0.6}$	$+0.1_{\pm 1.0}$	$-0.4_{\pm 0.6}$	$-0.1_{\pm 0.3}$	$+0.1_{\pm 0.4}$	$+0.1_{\pm 0.5}$
Liar's Dice	$+1.3_{\pm 0.8}$	$+0.3_{\pm 0.5}$	$+1.0_{\pm 1.2}$	$+1.9_{\pm 2.2}$	$+0.3_{\pm 1.5}$	$-0.2_{\pm 0.7}$	$+0.1_{\pm 1.0}$	$+0.1_{\pm 1.1}$
Goofspiel	$+4.3_{\pm 4.2}$	$+0.5_{\pm 0.8}$	$+4.5_{\pm 4.5}$	$+10.6_{\pm 6.3}$	$-0.1_{\pm 1.9}$	$+0.2_{\pm 0.6}$	$-0.2_{\pm 1.5}$	$+0.0_{\pm 1.6}$

Table 16: The largest instance, Goofspiel(5) with a random point order: relative gap (%) to the exact average of the same run, thirty seeds per row, $\varepsilon = 0.6$, 10^6 iterations. Same measurement as Table 2. $*$ marks the standard rules and † the per-set rule, an idealized baseline.

	own traversal				both	opponent's traversal			
Exploration	optimistic*	optimistic + tail	per-set†	per-history*	lazy (Lanctot)	optimistic + tail	per-set†	per-history	exact average (absolute expl.)
one-sided	$+10.9_{\pm 1.3}$	$+6.9_{\pm 1.1}$	$+11.5_{\pm 1.7}$	$+14.9_{\pm 1.8}$	$+1.5_{\pm 0.9}$	$+1.6_{\pm 0.7}$	$-0.8_{\pm 0.5}$	$-0.2_{\pm 0.5}$	$0.252_{\pm 0.004}$
two-sided	$+0.6_{\pm 0.4}$	$+0.3_{\pm 0.4}$	$+0.9_{\pm 0.3}$	$+1.2_{\pm 0.4}$	$+0.1_{\pm 0.5}$	$+0.0_{\pm 0.3}$	$+0.0_{\pm 0.4}$	$+0.4_{\pm 0.4}$	$0.443_{\pm 0.007}$

Appendix C.10. That instance is not in Table 17. The optimistic rule plus tail, which adds the tail at the reach in force at the end, ends at 3.8, 35.6 and 7.1%. Most of the gap of the optimistic rule is then weight error on the two Goofspiel instances and the tail on Liar's Dice(6).

C.9 Under external sampling the rules on the player's own traversal lose by more on the three larger games

Under external sampling the rule that adds on the opponent's traversal wins at the same iteration count, and on the three larger games the rules that add on the player's own traversal lose by more than under outcome sampling. Table 18 reports it. OpenSpiel's external-sampling rule adds the current strategy at weight one to the opponent's visited information sets, while the rules on the player's own traversal in Table 2 add to that player's own information sets. OpenSpiel's external-sampling rule is as good as the exact average or slightly better: -0.7 to -7.9% on Leduc, Liar's Dice and Goofspiel, and better on all thirty seeds of Goofspiel and on 23 of the 30 seeds of Leduc. Under external sampling an opponent's information set is visited in expectation in proportion to the own reach of the player who acts there. This suggests OpenSpiel's external-sampling rule needs neither a window weight nor an inverse probability. On Goofspiel the sign of its margin matches the per-history opponent-traversal column of Table 2, which is also below the own-traversal exact average there. On the three larger games the optimistic and the stochastic rules adding to the traversing player's own information sets are 2.5 to 3.9 times as exploitable as the exact average, and the optimistic rule plus tail 1.2 to 1.8 times. In these columns the tail dominates (optimistic rule 151% against optimistic rule plus tail 25% on Leduc), because the opponent is still sampled on-policy. On the three larger games the exact average is 5 to 9 times less exploitable at 10^5 iterations than under outcome sampling.

C.10 At equal iterations the exact average matches the best rule of Lanctot (2013)

At equal iterations the exact average matches the best of the rules that Lanctot (2013) compares. Table 19 reports it, and it is the row of Table 7 for those instances. The games, the sampling form and the schedule are those of Appendix B.6. Table 19 sets the opponent-traversal and the own-traversal columns beside the exact column. There the exact average is within 1% of the stochastic rules adding on the opponent's traversal and 6 to 17% better than the optimistic rule plus tail on that traversal. The same stochastic rules adding on the player's own traversal are 73 to 159% worse.

Table 17: Other instances after 10^5 iterations: relative exploitability gap (%) to the exact average (lower is better); mean and sample s.d. over thirty seeds. Last column: exploitability of the exact average. $*$ marks the standard rules and † the per-set rule, an idealized baseline.

Game	ε	optimistic*	optimistic + tail	per-set†	per-history*	exact average (absolute expl.)
One-sided exploration						
Liar's Dice (6)	0.6	$+9.1_{\pm 4.0}$	$+3.8_{\pm 1.5}$	$+5.2_{\pm 3.1}$	$+8.1_{\pm 4.4}$	$0.358_{\pm 0.024}$
	1.0	$+7.5_{\pm 3.5}$	$+2.0_{\pm 1.3}$	$+5.2_{\pm 3.6}$	$+6.9_{\pm 4.2}$	$0.448_{\pm 0.031}$
Goofspiel (5), fixed order	0.6	$+38.5_{\pm 20.7}$	$+35.6_{\pm 17.4}$	$+44.9_{\pm 18.8}$	$+53.5_{\pm 20.4}$	$0.089_{\pm 0.011}$
	1.0	$+25.0_{\pm 16.4}$	$+20.1_{\pm 16.3}$	$+36.3_{\pm 25.2}$	$+49.7_{\pm 28.9}$	$0.116_{\pm 0.014}$
Two-sided exploration						
Liar's Dice (6)	0.6	$+0.7_{\pm 0.4}$	$+0.6_{\pm 0.4}$	$-0.1_{\pm 0.6}$	$+0.1_{\pm 0.8}$	$0.498_{\pm 0.020}$
	1.0	$-0.1_{\pm 0.3}$	$-0.1_{\pm 0.2}$	$-0.4_{\pm 0.5}$	$-0.4_{\pm 0.5}$	$0.702_{\pm 0.035}$
Goofspiel (5), fixed order	0.6	$+0.5_{\pm 1.2}$	$+0.5_{\pm 1.2}$	$+0.3_{\pm 2.8}$	$+0.3_{\pm 2.6}$	$0.190_{\pm 0.031}$
	1.0	$+0.2_{\pm 0.6}$	$+0.1_{\pm 0.5}$	$+0.4_{\pm 1.4}$	$+0.4_{\pm 1.4}$	$0.613_{\pm 0.097}$

Table 18: External sampling (ES) after 10^5 iterations: relative exploitability gap (%) to the exact average of the same run (lower is better); mean and sample s.d. over thirty seeds. $*$ marks the standard rules of outcome sampling, here adding at the traversing player's own sets, and † the per-set rule, an idealized baseline; OpenSpiel's external-sampling rule is the first column.

Game	OpenSpiel ES rule	optimistic*	optimistic + tail	per-set†	per-history*	exact average (absolute expl.)
Kuhn	$-3.8_{\pm 17.3}$	$+4.2_{\pm 11.9}$	$+0.0_{\pm 0.1}$	$+41.6_{\pm 91.9}$	$+50.9_{\pm 86.9}$	$0.003_{\pm 0.001}$
Leduc	$-1.8_{\pm 2.3}$	$+150.8_{\pm 59.5}$	$+24.6_{\pm 16.7}$	$+159.4_{\pm 55.0}$	$+186.9_{\pm 59.9}$	$0.074_{\pm 0.005}$
Liar's Dice	$-0.7_{\pm 2.6}$	$+267.8_{\pm 245.7}$	$+40.1_{\pm 54.4}$	$+224.2_{\pm 222.9}$	$+283.1_{\pm 239.4}$	$0.011_{\pm 0.002}$
Goofspiel	$-7.9_{\pm 3.8}$	$+278.6_{\pm 61.0}$	$+78.6_{\pm 37.3}$	$+269.7_{\pm 71.4}$	$+283.8_{\pm 72.2}$	$0.019_{\pm 0.001}$

The stochastic rules adding on the opponent's traversal are within 0.9% above or below the exact average in the mean at every checkpoint. The parent-read rule adding on the opponent's traversal is the variant of Appendix C.12. It ends 1.0% below the own-traversal exact average on Liar's Dice(6) and 0.5% below it on Goofspiel(6) after 10^5 iterations (thirty seeds, standard deviations 0.4 and 0.6 points). These two differences measure the effect of that choice of traversal. Lanctot (2013) reports the same order of its stochastic and optimistic rules on Bluff and Goofspiel(6), the stochastic one fastest by its count of nodes touched. On Goofspiel(6) it also finds the lazy rule slower than the optimistic one.

On Goofspiel(6) the traversal moves a rule's gap by more than the tail does, while on Liar's Dice(6) the tail is the larger of the two for the optimistic rule. On the player's own traversal the optimistic rule is 85 to 153% worse than the exact average and the optimistic rule plus tail 25 to 87%. On both games the gap of the optimistic rule is 8 to 14% at 10^5 iterations and grows to its final value on every seed.

C.11 The rate of the per-set rule holds where its condition does

Corollary 1 is a statement about one information set whose own reach stays above a level on a positive fraction of the iterations, so we evaluate it in bands of mean own reach. On Leduc poker under two-sided exploration at $\varepsilon = 0.6$, over thirty seeds and 10^6 iterations, we record at seven log-spaced horizons the ℓ_1 distance between the per-set rule's average and the exact average at every set, and S_I^t summed over actions, whose ratio to t is the set's mean own reach. Averaging the distance over the sets of a band at each horizon and fitting the exponent of t gives -0.55 ± 0.04 where the mean own reach is above 0.5 and -0.52 ± 0.04 between 0.2 and 0.5, against -0.14 ± 0.04 below 0.05; moving the first cut from 0.3 to 0.7 moves that exponent between -0.56 and -0.52. Under one-sided exploration, where Theorem 2(b) says the rule is biased, the same fit over all sets together also gives -0.14. Under two-sided exploration the sets of low mean own reach are why a fit over all sets reads -0.42 rather than the rate.

C.12 What every rule returns on the opponent's traversal

The optimistic and stochastic rules also add on the opponent's traversal, as the pseudocode of Lanctot (2013) and Gibson et al. (2012b) do. Once the parent-read rule has added to I in iteration t, a second increment in that iteration is $W_I(t) - W_I(t) = 0$ times whatever strategy is in force. The parent-read rule adding only on the opponent's traversal instead applies each window's weight to the strategy at

Table 19: Our runs in the setting of Lanctot (2013): relative exploitability gap (%) to the exact average of the same run at the final iteration (10^7 on Liar's Dice(6); 5×10^6 on Goofspiel(6)); lower is better. Mean and sample s.d. over thirty seeds; rules adding on the own and on the opponent's traversal; last column: exploitability of the exact average.

Game	ε	own traversal (as in OpenSpiel)				exact average (absolute expl.)
		optimistic	optimistic + tail	per-set	per-history	
Liar's Dice (6)	0.5	$+140.5_{\pm 71.4}$	$+25.3_{\pm 8.2}$	$+96.5_{\pm 52.4}$	$+140.5_{\pm 70.9}$	$0.052_{\pm 0.005}$
	0.6	$+148.5_{\pm 55.1}$	$+27.7_{\pm 9.1}$	$+105.9_{\pm 43.8}$	$+144.6_{\pm 47.3}$	$0.050_{\pm 0.004}$
	0.8	$+152.6_{\pm 45.8}$	$+28.9_{\pm 7.4}$	$+116.9_{\pm 40.9}$	$+159.0_{\pm 51.3}$	$0.055_{\pm 0.004}$
Goofspiel(6)	0.5	$+85.1_{\pm 29.9}$	$+78.4_{\pm 31.4}$	$+73.2_{\pm 33.1}$	$+86.9_{\pm 34.1}$	$0.044_{\pm 0.004}$
	0.6	$+91.1_{\pm 35.7}$	$+83.8_{\pm 32.5}$	$+87.5_{\pm 50.2}$	$+105.0_{\pm 48.3}$	$0.040_{\pm 0.005}$
	0.8	$+91.6_{\pm 27.9}$	$+86.1_{\pm 25.1}$	$+85.6_{\pm 27.4}$	$+103.3_{\pm 29.4}$	$0.041_{\pm 0.003}$

Game	ε	on the opponent's traversal (as in Lanctot, 2013)			exact average (absolute expl.)
		optimistic + tail	per-set	per-history	
Liar's Dice (6)	0.5	$+8.9_{\pm 2.9}$	$+0.1_{\pm 0.8}$	$+0.6_{\pm 0.8}$	$0.052_{\pm 0.005}$
	0.6	$+9.2_{\pm 3.0}$	$-0.2_{\pm 0.8}$	$+0.3_{\pm 1.1}$	$0.050_{\pm 0.004}$
	0.8	$+7.8_{\pm 2.8}$	$+0.2_{\pm 0.9}$	$+0.2_{\pm 1.1}$	$0.055_{\pm 0.004}$
Goofspiel(6)	0.5	$+6.6_{\pm 6.7}$	$-0.4_{\pm 1.1}$	$-0.4_{\pm 1.8}$	$0.044_{\pm 0.004}$
	0.6	$+9.4_{\pm 8.2}$	$-0.2_{\pm 1.0}$	$-0.2_{\pm 1.4}$	$0.040_{\pm 0.005}$
	0.8	$+16.1_{\pm 10.4}$	$-0.2_{\pm 0.9}$	$-0.0_{\pm 1.0}$	$0.041_{\pm 0.003}$

the window's end, because player i's strategy changes at its own visits in between. On the trajectories of Table 2 the parent-read rule on the opponent's traversal ends -0.0% on Kuhn, -0.7% on Leduc, -1.3% on Liar's Dice and -1.4% on Goofspiel relative to the own-traversal exact average (standard deviations 0.2 to 0.9 points over seeds). Its normalized strategies differ from that average by 0.02 to 0.07 per entry in mean absolute difference on the larger games. The optimistic rule without its tail, which Table 2 shows only on the player's own traversal, ends $+3.1\%$ on Leduc poker, $+3.9\%$ on Liar's Dice and $+5.8\%$ on Goofspiel there.

Because every rule reads the same sampled histories within a run, the two effects the body compares can be differenced seed by seed. On a seed, write the rule effect for the gap between the optimistic rule and the per-history rule on the player's own traversal, and the traversal effect for the gap between the per-history rule on the two traversals, both as magnitudes. Their means over thirty seeds are 4.6 against 11.7 on Leduc poker, 8.0 against 19.3 on Liar's Dice and 12.6 against 52.0 on Goofspiel. The paired difference is $+7.1$, $+11.2$ and $+39.4$ points with 95% intervals $[3.7, 10.4]$, $[6.6, 15.8]$ and $[34.7, 44.1]$. None of the three contains zero. Taken as differences of the means in Table 2, the two standard rules are 2.4, 1.6 and 4.8 points apart, and the largest of the three is the bound Section 5.2 quotes. All three are smaller than the mean rule effects above (4.6, 8.0 and 12.6), because which of the two rules is the more exploitable changes from seed to seed.

The rule of Gibson et al. (2012b) adds $\sigma_t(I, \cdot)$ divided by the probability of the sampled history contributed by the traversing player's actions, with no own-reach factor. The per-history rule adds $\pi_i^{\sigma_t}(I)\,\sigma_t(I, \cdot)/q_t(h)$. Under one-sided exploration the non-traversing player is sampled from σ_t. Then $q_t(h)$ is that same probability times the chance probability of h times $\pi_i^{\sigma_t}(h)$, which perfect recall makes $\pi_i^{\sigma_t}(I)$. The two increments therefore differ by a factor of the chance probability of h alone. Wherever chance gives every history of a set the same probability, the two running sums differ by a factor that the normalization removes. The rules then return the same strategy. Enumerating the tree, we find that this holds at every one of the 12, 936, $1{,}024$ and $3{,}608$ information sets of the four games. Over thirty seeds on the same trajectories the two rules agree in exploitability to 1.1×10^{-14}. The two part where a set holds histories of different chance probability, as when a player observes a signal whose outcomes are not equally likely. Under two-sided exploration $q_t(h)$ carries the explored reach of player i instead, and the factor $\pi_i^{\sigma_t}(I)$ does not cancel. The rules then part: at $\varepsilon = 0.6$ after 2000 iterations they differ at 353 of the 859 visited sets of Leduc poker, by up to 0.73 per entry.

Adding on the opponent's traversal also makes the per-set running sum unbiased under one-sided exploration, for the strategies added on that traversal. The opponent explores, so every history with positive own reach has positive sampling probability. Every nonzero target contribution can then be sampled. The per-history rule normalizes by $1/q_t(h)$ and the per-set rule by $1/q_t(I)$. At $\varepsilon = 0.6$ their strategies have exploitabilities within 1% of each other.

In the setting of Lanctot (2013) (Table 19) the optimistic rule plus tail on the opponent's traversal is within 2% of the exact average up to 10^5 iterations, 2 to 6% worse at 5×10^5 and 6 to 17% worse at the end. At the end it is worse than the exact average on all thirty seeds in two of six cells and on 26 to 29 of the thirty in the other four. Adding the tail changes the result by at most 0.3 point on the mean and by at most 2.1% on any seed, because visits are frequent and the tail is short on the opponent's traversal. The remaining gap of the optimistic rule plus tail is the weight error of Section 4 with one added effect. The strategy at I changes between two increments, since the regret update happens on the player's own traversal.

C.13 How much of the player's own reach the sampler cannot reach

Under one-sided exploration the opponent is sampled from its own strategy. An information set I of player i can then be visited on i's traversal at iteration t only if $\pi_{-i}^{\sigma_t}(I) = \sum_{h \in I} \pi_{-i}^{\sigma_t}(h)$ is positive. When it is zero and the own reach is not, the own traversal cannot visit I: the stochastic rules omit that iteration, the optimistic rule counts it at the next visit, if any, with the reach seen there, and the parent-read rule adds its exact weight at the next visit or at the end. The omission by the stochastic rules is the effect of Theorem 2(b), and it is separate from the weight error of Theorem 1(b), which is what a set with no earlier decision of the player escapes. We size it by enumerating the whole tree every 200 iterations of a run of 2×10^4 iterations at $\varepsilon = 0.6$, over three seeds, and reporting the share of own reach that the sampler cannot reach,

$$\Big(\sum_{I,t} \pi_i^{\sigma_t}(I)\, \mathbb{1}\big[\pi_{-i}^{\sigma_t}(I) = 0\big] \Big) \Big/ \Big(\sum_{I,t} \pi_i^{\sigma_t}(I) \Big). \tag{7}$$

It reads 9.3% on Kuhn poker, 21.7% on Leduc poker, 38.6% on Liar's Dice and 49.6% on Goofspiel, the three seeds of a game spanning at most 6.8 points. The order is the order of the optimistic rule's gaps in Table 2, with Kuhn poker smallest in both. Under two-sided exploration the opponent gives every action probability at least $\varepsilon / \max_{I'} |A(I')|$. Then (7) is zero whatever the strategies are, and the measurement returns exactly zero on Kuhn poker.

C.14 The lazy rule of Lanctot (2013) as written

The final push that Lanctot (2013) mentions without specifying moves the figures of Section 5.1 by at most a tenth of a point. Lanctot (2013) compares the rules at equal numbers of nodes touched and lists the final push of its lazy rule as extra computation time, whereas we compare at equal iterations with a parent-read rule. The lazy rule pushes weights to the children, and the parent-read rule reads the preceding action entry.

We also run the lazy rule as written on the trajectories of Table 2. At every visited decision node of player i it pushes the pending weight of each action to the sampled child, and a final push over the whole tree closes the run. Its relative gap to the exact average after 10^5 iterations is -0.6% on Leduc, 2.7% on Liar's Dice, 2.2% on Goofspiel and 9.9% on Kuhn before the final push, which moves each by under a tenth of a point. The standard deviations over seeds are 3, 8, 5 and 25 points. On Kuhn the weight pushed along the sampled path is lost whenever that path ends before another information set of player i. Pushing the weight to every own successor of (I, a) instead would visit the subtree below (I, a) at each push. For instance, let action a at I be followed by an opponent decision with two actions, one ending the game and one leading to the own set J. The pending weight $w(I, a)$ reaches J only on the traversals that sample the second action and is discarded on those that sample the first. The parent-read rule instead gives J the full weight $s_I(a)$ at its next visit. Algorithm 2 transcribes the rule as run. Its lines 2 to 5 are lines 14 to 16 and 21 to 24 of Algorithm 5 of Lanctot (2013), and its line 6 is the final push that Lanctot (2013) mentions without specifying.

The lazy rule's normalized strategies differ from the exact average by 0.0018 on Kuhn and 0.022 to 0.031 on the larger games in mean absolute difference per entry, against 0.041 to 0.076 for the optimistic rule there. Table 5 measures the lazy rule at 1.1 to 1.2 times the training time of the parent-read rule.

C.15 What a rule costs in compute

The tables report how far a rule lands from the exact average at a fixed iteration count. A solver's operator asks instead how much work a rule needs to reach a given exploitability. Every rule reads

Algorithm 2 The lazy rule of Lanctot (2013) as run here, at a decision node $h \in I$ of player p with sampled action a_0, on either player's traversal. r_p is the running weight of player p passed from above and carried unchanged through the other player's nodes; π_p is player p's own reach of h.

1: Once, initialize $w(I, a) \leftarrow 0$ for every (I, a); at the root of every traversal, set $r_p \leftarrow 0$ for both players.
2: $r'_p \leftarrow r_p\, \sigma(I, a_0) + w(I, a_0)$.
3: Recurse into ha_0 with running weight r'_p.
4: $s_I \leftarrow s_I + (r_p + \pi_p)\, \sigma(I, \cdot)$.
5: $w(I, a) \leftarrow w(I, a) + (r_p + \pi_p)\, \sigma(I, a)$ for $a \neq a_0$, and $w(I, a_0) \leftarrow 0$.
6: At the end of the run, one traversal of the whole tree pushes every pending $w(I, a)$ down in the same way and adds to each set the running weight that reaches it.

the same sampled histories within a run, so the answer is a comparison inside the run. For a target iteration count T we take the exploitability the exact average has at T, find the first checkpoint at which a rule is at or below it, and count the work done there as the nodes touched times that rule's training-time overhead on the same game from Table 5, relative to the same quantity for the exact average at T. We report the median over all runs and count a run that never reaches the target as above every run that does. When the median falls among the runs that miss the target, counting each of them at the work of its whole run gives a lower bound for the median. The resolution is the checkpoint spacing, 10^4 iterations for $T \leq 3 \times 10^5$ and 10^5 iterations for $T \geq 10^6$. On Leduc poker over thirty seeds with $T = 10^5$, the optimistic rule needs 1.26 times the work of the exact average, the per-history rule 1.33 and the per-history rule on the opponent's traversal 0.99. With $T = 3 \times 10^5$ they need 1.43, 1.45 and 1.00. For the twelve seeds of each larger game carried to 10^7, the last column of Table 13 gives the medians for $T = 10^6$. On Goofspiel one run of each standard rule reaches the target within 10^7 iterations, and on Liar's Dice six runs of the per-history rule do. At $T = 2 \times 10^6$ on Leduc poker three of the twelve runs of the per-history rule reach the target, while all twelve of the same rule on the opponent's traversal do, at 0.94 times the work.

C.16 The lazy rule is accurate because of where it adds

Algorithm 2 updates player p at p's own decision nodes whichever player is being updated, and an iteration traverses once per player. The lazy rule accumulates for a player on both traversals. To separate that from the weight it targets, we run it again confined to the traversal of the player being updated. The weight it targets is the same in both runs. The position alone separates them. Across the 120 cells of the four games and thirty seeds, every other rule is unchanged to the last bit at all $12{,}000$ checkpoints. The two readings are paired within a seed. The relative gap of the lazy rule then moves from 9.9 to 19.7% on Kuhn poker, from -0.6 to 12.3% on Leduc poker, from 2.7 to 18.7% on Liar's Dice and from 2.2 to 54.1% on Goofspiel. The paired differences are 9.8, 13.0, 16.0 and 51.9 points, with 95% intervals $[3.3, 16.2]$, $[10.2, 15.7]$, $[13.0, 19.0]$ and $[47.0, 56.9]$. The confined rule is the worse one on 20, 29, 30 and 30 of the thirty seeds. Confined, it lands inside or above the range that the two standard rules occupy on every one of the four games.

C.17 Intervals, ranges and seed counts

Table 20 gives the interval of the mean, the range over seeds and the count of worse seeds for each of these cells: the optimistic rule on the three larger games, the stochastic rules on the player's own traversal on all four games and the stochastic rules adding on the opponent's traversal on Goofspiel at $\varepsilon = 0.6$ (Table 2, with the per-set cells in Table 11), the Leduc horizon cells at 10^5 and 10^6 iterations (Table 14), OpenSpiel's external-sampling rule under external sampling (Table 18) and the optimistic rule plus tail on the opponent's traversal in the setting of Lanctot (2013) (Table 19).

Table 20: Relative exploitability gap (%) to the exact average of the same run, for the cells the text describes seed by seed; negative means better. Columns: seeds, mean, a 95% interval of the mean, the range over seeds, and the number of seeds on which the rule is worse.

Cell	seeds	mean	95% interval	range	worse
Leduc, optimistic, own traversal	30	+13.2	[+10.5, +15.9]	+0.3 to +31.1	30
Liar's Dice, optimistic, own traversal	30	+16.4	[+12.5, +20.3]	−3.1 to +44.7	29
Goofspiel, optimistic, own traversal	30	+45.3	[+40.1, +50.6]	+15.1 to +71.3	30
Kuhn, per-set, own traversal	30	+1.8	[−3.6, +7.2]	−25.7 to +42.5	19
Kuhn, per-history, own traversal	30	+11.8	[+1.7, +21.9]	−37.1 to +78.9	17
Leduc, per-set, own traversal	30	+10.1	[+6.8, +13.4]	−9.1 to +32.6	27
Leduc, per-history, own traversal	30	+10.8	[+7.5, +14.1]	−7.6 to +34.2	27
Liar's Dice, per-set, own traversal	30	+11.5	[+7.7, +15.3]	−4.3 to +33.2	27
Liar's Dice, per-history, own traversal	30	+18.0	[+12.8, +23.2]	−8.9 to +61.7	27
Goofspiel, per-set, own traversal	30	+44.9	[+40.2, +49.6]	+19.4 to +72.7	30
Goofspiel, per-history, own traversal	30	+50.1	[+45.0, +55.3]	+23.0 to +77.3	30
Goofspiel, per-set, opp. traversal	30	−2.7	[−3.7, −1.7]	−10.6 to +3.0	3
Goofspiel, per-history, opp. traversal	30	−1.9	[−3.0, −0.8]	−8.9 to +4.5	6
Leduc, $\varepsilon = 0.6$, optimistic, 10^6	30	+34.9	[+25.6, +44.2]	+3.0 to +91.5	30
Leduc, $\varepsilon = 0.6$, per-set, 10^6	30	+46.7	[+36.7, +56.6]	+6.5 to +110.0	30
Leduc, $\varepsilon = 0.6$, per-history, 10^6	30	+53.3	[+42.3, +64.3]	+7.8 to +119.0	30
Leduc, $\varepsilon = 1.0$, optimistic, 10^5	30	+14.3	[+11.6, +16.9]	+3.5 to +31.7	30
Leduc, $\varepsilon = 1.0$, optimistic, 10^6	30	+39.7	[+30.3, +49.2]	+5.5 to +94.4	30
Leduc, $\varepsilon = 1.0$, per-set, 10^5	30	+12.4	[+9.4, +15.4]	−0.4 to +32.1	29
Leduc, $\varepsilon = 1.0$, per-set, 10^6	30	+50.0	[+38.0, +62.0]	+11.7 to +132.2	30
Leduc, $\varepsilon = 1.0$, per-history, 10^5	30	+13.6	[+10.4, +16.8]	−0.2 to +31.8	29
Leduc, $\varepsilon = 1.0$, per-history, 10^6	30	+58.1	[+45.5, +70.7]	+15.7 to +142.0	30
Leduc, OpenSpiel ES rule, external sampling	30	−1.8	[−2.7, −0.9]	−6.9 to +1.6	7
Liar's Dice, OpenSpiel ES rule, external sampling	30	−0.7	[−1.6, +0.3]	−4.9 to +4.8	12
Goofspiel, OpenSpiel ES rule, external sampling	30	−7.9	[−9.3, −6.5]	−15.2 to −1.4	0
Liar's Dice(6), $\varepsilon = 0.5$, optimistic + tail, opp. traversal	30	+8.9	[+7.8, +10.0]	+4.2 to +16.7	30
Liar's Dice(6), $\varepsilon = 0.6$, optimistic + tail, opp. traversal	30	+9.2	[+8.1, +10.3]	+2.6 to +16.2	30
Liar's Dice(6), $\varepsilon = 0.8$, optimistic + tail, opp. traversal	30	+7.8	[+6.8, +8.9]	−0.1 to +13.4	29
Goofspiel(6), $\varepsilon = 0.5$, optimistic + tail, opp. traversal	30	+6.6	[+4.1, +9.1]	−8.9 to +23.1	27
Goofspiel(6), $\varepsilon = 0.6$, optimistic + tail, opp. traversal	30	+9.4	[+6.3, +12.5]	−6.3 to +36.2	26
Goofspiel(6), $\varepsilon = 0.8$, optimistic + tail, opp. traversal	30	+16.1	[+12.2, +19.9]	−0.1 to +46.4	29