The AAAI-17 Workshop on
Computer Poker and Imperfect Information Games
WS-17-06

Targeted CFR

Eric Jackson

Abstract

In recent years, Counterfactual Regret Minimization (CFR) has emerged as the standard technique for computing near-equilibrium solutions to large games of imperfect information. This paper describes a new sampling variant of Counterfactual Regret Minimization, called Targeted CFR. We compare with other sampling variants including Outcome Sampling and External Sampling, and present experimental results on poker. We find that Targeted CFR outperforms other sampling variants on certain types of large games.

Counterfactual Regret Minimization

Counterfactual regret minimization (CFR) (Zinkevich et al. 2007) is a technique for solving large games of imperfect information that has become the dominant approach in the field of computer poker. It takes regret minimization techniques that have traditionally been applied to normal-form games and adapts them to work efficiently on extensive-form games like poker.

Below we briefly recap the key elements of CFR. For a fuller exposition, the reader is invited to consult the prior literature such as (Zinkevich et al. 2007). For background on extensive-form games, the reader can consult a text such as (Osborne and Rubenstein 1994). Familiarity with terms such as information set, history and abstraction is assumed below.

Through a form of self-play, CFR computes (behavioral) strategies that provably converge to a Nash equilibrium for two-player zero-sum extensive-form games with perfect recall. CFR maintains a quantity known as "regret" for every action. Regret is a measure how much the player could have gained over previous iterations by playing that action as opposed to playing the actual sequence of actions that he did. At each iteration we compute a current strategy for each player from the regrets. We also maintain the cumulative profile for each player which sums the probabilities assigned to each action over the entire sequence of iterations. The cumulative profile allows us to compute the average strategy; it is this average strategy which can be shown to converge to equilibrium.

To compute regrets, a quantity known as "counterfactual value" is computed for actions at an information set. Given a current strategy profile σ for the two players, the counterfactual value $v_i(\sigma, I)$ of an information set for player i is the expected utility at that information set, assuming that player i plays to that information set:

$$v_i(\sigma, I) = \sum_{z \in Z_I} (u_i(z)\pi^{\sigma}_{-i}(z[I])\pi^{\sigma}(z[I], z)) \quad (1)$$

Here Z_I is the set of terminal histories passing through I, and $z[I]$ is the prefix of z contained in I. $\pi^{\sigma}(h)$ is the probability of history h occurring if all players play according to σ. $\pi^{\sigma}_{-i}(h)$ only incorporates the probabilities of players other than i (including chance). $\pi^{\sigma}(h_1, h_2)$ is the product of the probabilities along the path between h_1 and h_2. $u_i(z)$ is the utility accruing to player i at terminal history z.

The counterfactual regret at iteration t of an action a is how much player i could have gained by playing a as opposed to the current strategy that he actually employed.

$$r^t_i(I, a) = v_i(\sigma^t_{I \to a}, I) - v_i(\sigma^t, I) \quad (2)$$

Here $\sigma^t_{I \to a}$ is the strategy profile identical to σ^t except that at information set I action a is selected with probability 1.

We maintain the sum of the regrets across all iterations:

$$R^T_i(I, a) = \sum_{t=1}^{T} r^t_i(I, a) \quad (3)$$

The current strategy for iteration $T + 1$ is calculated from the regrets at iteration T using an approach known as "regret matching" (Hart and Mas-Colell 2000) in which probabilities are proportional to positive regret:

$$\sigma^{T+1}(I, a) = \frac{R^{T,+}_i(I, a)}{\sum_{b \in A(I)} R^{T,+}_i(I, b)} \quad (4)$$

x^+ means $\max\{x, 0\}$.

The original version of CFR is known as "Vanilla" CFR. In Vanilla CFR, on each iteration an exhaustive traversal of the game tree is performed, with exact counterfactual values being computed at every information set. Likewise regrets and the cumulative profile are updated at each information set. We alternate between iterations on which we are updating player 1's regrets and iterations on which we are updating player 2's regrets. On each iteration we say that one player is the "target player" and the other player is the opponent.

Sampling Variants

Numerous sampling variants of CFR have been introduced over the last several years including External Sampling and Outcome Sampling (Lanctot et al. 2008) and Average Strategy Sampling (Gibson et al. 2012a). Sampling variants of CFR traverse only a small fraction of the game tree on each iteration, and replace an exact computation of the counterfactual value $v(\sigma, I)$ with an unbiased estimate $\hat{v}(\sigma, I)$. We often find faster convergence to near-equilibrium solutions, especially for large games. One reason for the success of these sampling variants may be that they spend more time updating the important parts of the strategy, specifically those areas that get played to more frequently.

In the framework of (Gibson et al. 2012b), sampling variants of CFR must satisfy a couple of requirements. The first is that the estimate of counterfactual value they produce, $\hat{v}(\sigma, I)$, be an unbiased estimate. The second, which we will term the "reachability requirement", requires that all portions of the game tree that the opponent plays to be sampled with some non-zero probability [1]

Sampling variants of CFR have been employed with good results in the Computer Poker Competition. The top three competitors from the 2014 competition all employed some variation on External Sampling. On the other hand, Heads-Up Limit Texas Hold'em was recently solved (Tammelin et al. 2015) with CFR+, a variant of Vanilla CFR that employs no sampling. It seems likely that games that can be solved very exactly (to a very low exploitability, and using no card abstraction) will be best tackled with CFR+. But games that are too large to solve with CFR+ will still best be attacked using some form of sampling.

Outcome Sampling

In Outcome Sampling, a single action is sampled at every information set. At an information set where chance acts, we sample according to the fixed chance distribution. At an information set where the opponent acts, we sample according to the opponent's current strategy. At an information set where the target player acts, we sample *approximately* according to the target player's current strategy. The complication is that we require some "exploration" to ensure that the reachability requirement is satisfied. For example, with some small probability ϵ we may sample an action according to the uniform distribution.

A single iteration of Outcome Sampling involves following a single trajectory from the root of the game tree to a terminal history z. This is depicted in figure 1.

External Sampling

External Sampling is like Outcome Sampling at information sets where chance acts or where the opponent acts; i.e., we sample a single action according to the current strategy profile. External Sampling differs at information sets where the target player acts. At these information sets we evaluate all actions. Since we evaluate all of the target player actions we meet the reachability requirement.

External Sampling is depicted in figure 2.

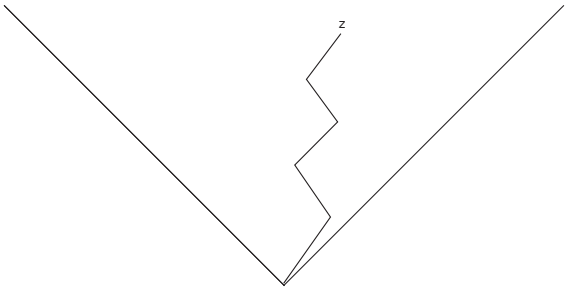

Figure 1: Outcome Sampling

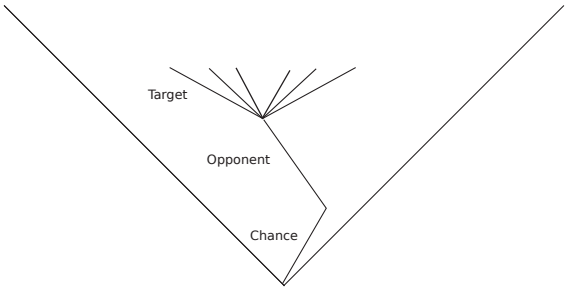

Figure 2: External Sampling

Average Strategy Sampling

Average Strategy Sampling is identical to External Sampling for information sets where chance or the opponent acts. Where the target player acts, we sample more or less according to the average strategy up to this point in time. As with Outcome Sampling, the catch is that we must ensure some exploration to ensure that the reachability condition is satisfied. In Average Strategy Sampling, we achieve this by always evaluating a target player action with at least probability ϵ. See(Gibson et al. 2012a) for more details.

Probes

In (Gibson et al. 2012b), a new wrinkle is introduced. Imagine we are at an information set and we wish to perform an update. In Outcome Sampling, we would sample one action according to the current strategy, but no others. Gibson et al. observe that this is effectively estimating a zero counterfactual value for all the unsampled actions at that information set. They propose instead computing a rough estimate of the counterfactual value of those unsampled actions with a "probe". A probe is a walk down a single trajectory from the current information set to a terminal history according to the current strategy profile. It produces an unbiased estimate of the counterfactual value, as desired.

A probe is depicted in figure 3. Here we have reached

[1] The "reachability requirement" is not explicitly stated in (Gibson et al. 2012b), but arises from the fact that the estimated counterfactual value is scaled up by $1/q_i(I)$ where $q_i(I)$ is player i's contribution to the probability of sampling information set I. If $q_i(I)$ were zero, this estimated counterfactual value would be undefined.

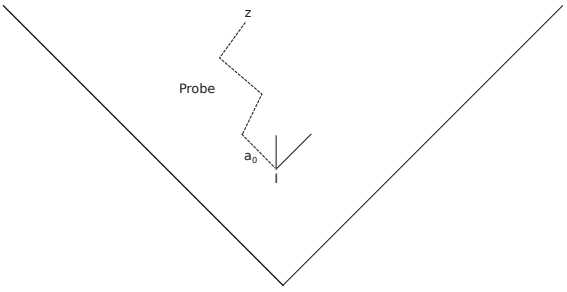

Figure 3: A probe

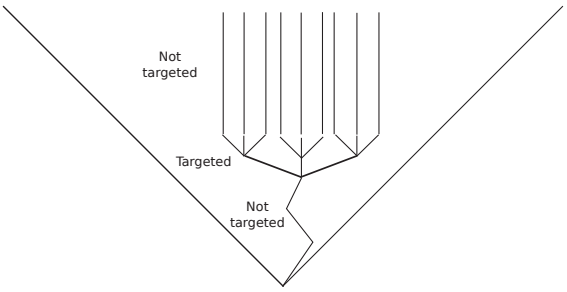

Figure 4: Targeted CFR

information set I and we choose to estimate the value of action a_0 with a probe. Instead of fully evaluating it, we follow a single on-policy trajectory to the terminal history z which is represented by the dashed line.

One interesting thing about probes is that we calculate counterfactual values in a portion of the game tree in which we do not update regrets (nor the cumulative profile). (In figure 3 we do not update regrets along the path marked by the dashed line.) In this respect incorporating probes takes us outside the family of sampling algorithms defined in (Lanctot et al. 2008).

Targeted CFR

The new approach described in this paper, Targeted CFR, takes some inspiration from all of the above approaches. It can be viewed as falling somwhere between Outcome Sampling and External Sampling on the spectrum of sampling algorithms, in the sense that a single iteration of Targeted CFR will typically visit more of the game tree than Outcome Sampling, but less than External Sampling.

In Targeted CFR, we divide the game tree into a number of partitions. On a single iteration, one or more partitions will be "targeted". When within a targeted partition, sampling and regret updates are performed as in External Sampling. When not within an targeted partition, we essentially perform a probe. That is, we follow a single trajectory sampling a single action at every information set according to each player's current strategy. We do not perform any updates (of regrets or the cumulative profile) in the non-targeted regions, but we do return an unbiased estimate of the counterfactual value.

Targeted CFR is depicted in figure 4.

The choice of how to partition the game tree and how often to target each partition is a design choice that the implementer must make. In Texas Hold'em poker, there are four betting rounds, which provides us with a natural way to partition the game tree. We may elect to target either a single betting round or a combination of betting rounds.

In order to satisfy the reachability requirement, Targeted CFR requires occasional iterations on which *all* partitions are targeted. We refer to these as "full" iterations. Full iterations are identical to External Sampling iterations. Since External Sampling satisfies the reachability constraint, Targeted CFR, which performs these iterations with non-zero probability also satisfies the reachability constraint.

We use a parameter τ to designate the distribution that determines how often each combination of partitions — betting rounds, in the case of poker — is targeted.

Pseudocode

Algorithm 1 Targeted CFR

Require: Parameters τ
Require: Initialize regrets: $\forall a, I : r(I, a) \leftarrow 0$
Require: Initialize cumulative profile: $\forall a, I : s(I, a) \leftarrow 0$

function ITERATION(player i)
  Sample targeted partitions $R \sim \tau$
  Targeted($\emptyset, i$)
function TARGETED(history h, player i)
  $r \leftarrow$ index of partition
  $I \leftarrow$ information set containing h
  $\sigma(I, \cdot) \leftarrow$ RegretMatching($r(I, \cdot)$)
  if $h \in Z$ **then**
    return $u_i(h)$
  else if $h \in P(c)$ **then**
    Sample action $a \sim \sigma_c(h, \cdot)$
    return Targeted(ha, i)
  else if $h \notin P(i)$ **then**
    if R contains all partitions **then**
      for $a \in A(I)$ **do**
        $s(I, a) \leftarrow s(I, a) + \sigma(I, a)$
    Sample action $a \sim \sigma(I, \cdot)$
    return Targeted(ha, i)
  else
    if $r \in R$ **then**
      for $a \in A(I)$ **do**
        $\hat{v}(a) \leftarrow$ Targeted(ha, i)
      $\hat{c} \leftarrow \sum_{a \in A(I)} \sigma(I, a)\hat{v}(a)$
      for $a \in A(I)$ **do**
        $r(I, a) \leftarrow r(I, a) + \hat{v}(a) - \hat{c}$
      return $\hat{c}$
    else
      Sample action $a \sim \sigma(I, \cdot)$
      return Targeted(ha, i)

Pseudocode for Targeted CFR is shown in Algorithm 1.

A reader familiar with Monte Carlo CFR (MCCFR) methods will recall that in MCCFR quantities often need to be

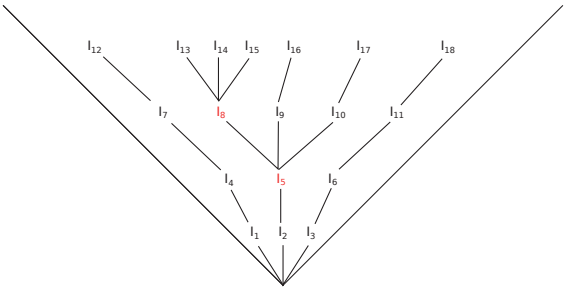

Figure 5: Targeting Important Information Sets

scaled by $1/q$ where q is the probability of the given information set having been sampled. This eliminates biases that would result from certain information sets being sampled more often than others. But there is no such rescaling in the pseudocode above.

With respect to the cumulative profile, we make the expedient choice of updating it only during the "full" iterations. (Gibson et al. 2012b) shows how the cumulative profile can be updated during the opponent phase of External Sampling. It turns out that no rescaling is needed when you do this.

For updating regrets, it is theoretically required to scale counterfactual values by $1/q$ as is done in other MCCFR variants. However, perhaps surprisingly, we have found empirically that such weighting actually hurts the rate of convergence. We speculate that these rescaling factors (which will be quite large when q is small) introduce a lot of variance into the regrets as compared to the unscaled values.

Extensions

So far we have discussed targeting entire betting rounds on each iteration. A somewhat different type of approach is to identify specific information sets that are important, and to fully explore all actions at those information sets. This is depicted in figure 5. The figure depicts only information sets at which the target player acts. The ones in red are those deemed important. (As usual, where chance or the opponent act, we always sample a single action.)

There are many ways one could define which information sets are most important. We have typically defined them as information sets in which the gap in regrets between the top two actions is less than some threshold. Important information sets are thus identified as those where there is a close decision. This threshold is likely game-dependent, but in our experiments on poker a value of around one thousand big blinds has worked well.

Generally we find it advantageous to combine this form of targeting with the targeting of betting rounds described earlier.

Discussion

Targeted CFR is in a certain sense between Outcome Sampling and External Sampling on the spectrum of sampling algorithms. Outcome Sampling employs the fastest but least accurate iterations in that each iteration follows only a single trajectory from root to leaf. External Sampling traverses much more of the game tree on each iteration because it evaluates all actions for the target player. Targeted CFR is between these two because it pursues all actions for the target player only in some portions of the tree; otherwise it is following a single trajectory as in Outcome Sampling.

As observed in (Gibson et al. 2012b), Outcome Sampling leads to a high variance in the estimated counterfactual value computed for an action. On iterations where an action is not evaluated, the counterfactual value is (implicitly) considered to be zero. This may explain why Outcome Sampling has not been preferred for most forms of poker. Adding probes to Outcome Sampling gives us a superior (although still very rough) estimate of every action's counterfactual value. Likewise, in Targeted CFR, at information sets where we are performing regret updates (i.e., in the targeted portion of the game tree) we always have an estimated counterfactual value for every action.

In all our experiments with poker, it has proven advantageous to target the later betting rounds more often than the earlier betting rounds. We conjecture that the reason is that game trees for poker exhibit a high degree of fan-out and there are far more histories on the later betting rounds than the earlier betting rounds. All forms of sampling prior to Targeted CFR visit histories in the earlier betting rounds far more frequently than histories in the later betting rounds. Targeted CFR addresses this imbalance by targeting the later betting rounds more often.

Results

We performed a variety of experiments, mostly comparing Targeted CFR to External Sampling, but also in some cases Average Strategy Sampling. Two different measures of the quality of our strategies were employed: exploitability and head-to-head performance. Exploitability measures how much a given strategy would lose to a best response, and can be thought of as a measure of distance from equilibrium. A true equilibrium strategy would have zero exploitability.

Head-to-head performance was measured by comparison to a reference strategy. For the reference strategy we typically use the best strategy we have available for the given game. We use head-to-head performance for a couple of reasons. First, for some games it is not feasible to compute exploitability. Second, depending on his or her goals, an implementer may care more about head-to-head performance than exploitability.

Measuring head-to-head performance in this way has the limitation that we are only comparing to a single reference system. It is possible that a strategy may have excellent head-to-head performance against one opponent, but not against another.

We discuss below results on games of different sizes. For poker we think of the size of the game along two dimensions; the number of different betting sequences allowed, and the number of "buckets" in the card abstraction. For a game with no card abstraction, the number of buckets is simply the number of permutations of the cards from the viewpoint of one player (modulo isomorphism). In Texas Hold'em, on the final betting round, this would be the num-

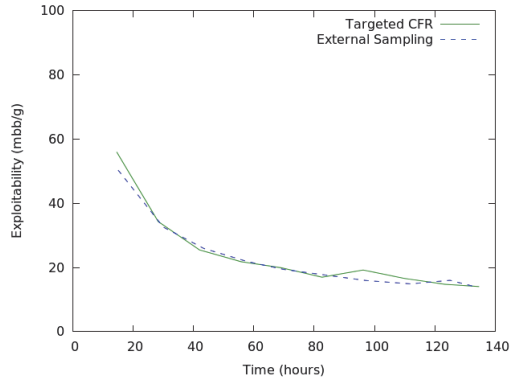

Figure 6: Game 1: Exploitability Over Time

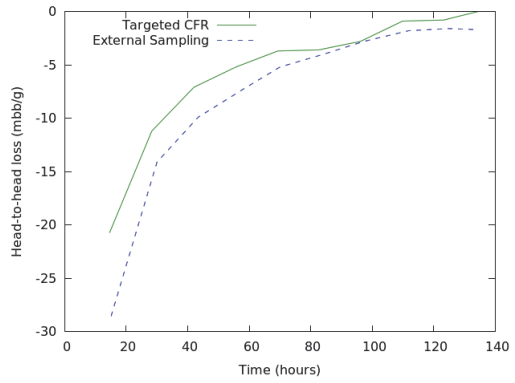

Figure 7: Game 1: Head-to-head loss over time

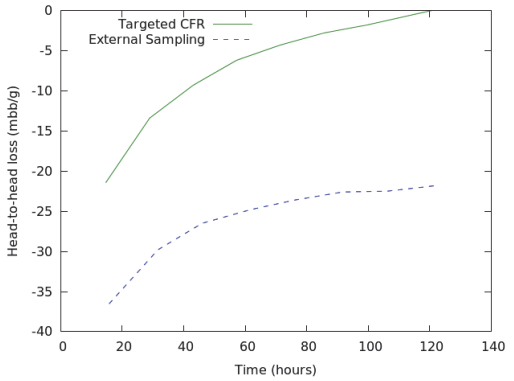

Figure 8: Game 2: Head-to-head loss over time

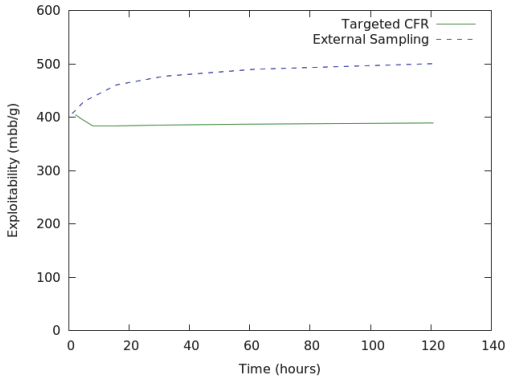

Figure 9: Game 2: Real-game exploitability over time

ber of ways of dealing five community cards and two private cards to the player.

For the first game ("game 1") we used a 24 card deck with no card abstraction. There are approximately 6.25 million buckets on the final betting round. The game permits a single pot-size bet on every betting round, which leads to a total of 322 betting sequences. For the parameter τ which controls how often we target each betting round we used this distribution:

Betting Rounds	Probability
$0, 1, 2, 3$	0.2
$2, 3$	0.8

Note that we are always targeting multiple betting rounds, and that we prioritize the later betting rounds over the earlier betting rounds.

Figure 6 shows exploitability over time. As you can see, Targeted CFR and External CFR differ hardly at all on this measure. Figure 7 shows the head-to-head results. Targeted CFR offers a small improvement which diminishes over time.

We turn now to games that do not employ a perfect card abstraction. For game 2, we use a full 52-card deck, but cluster the hands on each of the last three betting rounds into approximately one million buckets. The card abstraction exhibits imperfect recall (Waugh et al. 2009). We use the same betting structure as game 1, with only 322 betting sequences possible.

For the parameter τ which controls how often we target each betting round we used this distribution:

Betting Rounds	Probability
$0, 1, 2, 3$	0.025
$2, 3$	0.75
$1, 3$	0.175
$0, 3$	0.05

The head-to-head results are shown in figure 8. Surprisingly we see much larger head-to-head outperformance with this game than with game 1. We also computed real-game exploitability, with results shown in figure 9.

Note that for both systems real-game exploitability bottoms out very quickly and actually gets slightly worse over time. This is typical of systems with imperfect recall card abstractions; they tend not to optimize real-game exploitability as well as one might hope. Having said that, it does appear that Targeted CFR produces a substantially lower real-game exploitability.

Our final game, game 3, employs a much larger betting system than either of the previous two games. We allow a maximum of two bets per betting round and four bet sizes, which leads to around a million betting sequences. We used a smaller imperfect recall card abstraction with 169, 10,000,

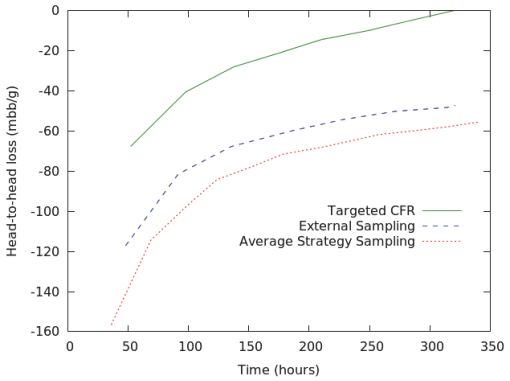

Figure 10: Game 3: Head-to-head loss over time

10,000 and 1,980 buckets on the four betting rounds respectively.

We compare Targeted CFR to both External Sampling and Average Strategy Sampling, with the head-to-head results shown in figure 10. As you can see Targeted CFR outperforms by a substantial margin. [2]

The results shown above are a mixed bag, but Targeted CFR outperforms on the two games that employ imperfect recall card abstractions. Games that can be solved with no card abstraction are probably best attacked through other methods (e.g., CFR+). Targeted CFR is best suited for larger games that require card abstraction and cannot feasibly be solved with methods like CFR+.

Targeted CFR is being employed in the construction of Slumbot 2017, our entry into the 2017 Computer Poker Competition.

Conclusions

We have presented Targeted CFR, a new variant of counterfactual regret minimization, along with results comparing it to External Sampling and Average Strategy Sampling. Targeted CFR generally outperforms on large games with imperfect recall card abstractions.

The present paper explores only a couple of methods of targeting; there are no doubt other variations that would prove fruitful.

References

Gibson, R.; Burch, N.; Lanctot, M.; and Szafron, D. 2012a. Efficient monte carlo counterfactual regret minimization in games with many player actions. In *Advances in Neural Information Processing Systems 25 (NIPS)*.

Gibson, R.; Lanctot, M.; Burch, N.; Szafron, D.; and Bowling, M. 2012b. Generalized sampling and variance in counterfactual regret minimization. In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*.

Hart, S., and Mas-Colell, A. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*.

Lanctot, M.; Waugh, K.; Zinkevich, M.; and Bowling, M. 2008. Monte carlo sampling for regret minimization in extensive games. In *Advances in Neural Information Processing Systems 22 (NIPS)*.

Osborne, M. J., and Rubenstein, A. 1994. *A Course in Game Theory*. The MIT Press.

Tammelin, O.; Burch, N.; Johanson, M.; and Bowling, M. 2015. Solving heads-up limit texas hold'em. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), 2015*.

Waugh, K.; Zinkevich, M.; Johanson, M.; Kan, M.; Schnizlein, D.; and Bowling, M. 2009. A practical use of imperfect recall. In *Proceedings of the Eighth Symposium on Abstraction, Reformulation and Approximation (SARA)*.

Zinkevich, M.; Bowling, M.; Johanson, M.; and Piccione, C. 2007. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS)*.

[2] It is unclear what would happen if we ran this experiment indefinitely long. If this were a game with a perfect recall abstraction we would be guaranteed that the External Sampling and Average Strategy Sampling systems would converge to equilibria within the abstract game, and so their head-to-head losses should disappear. But since this is an imperfect recall abstraction, convergence to equilibrium is not guaranteed and it is at least possible that this head-to-head gap could persist indefinitely.