Techniques for Speeding Up CFR
(some of these apply to other algorithms also)

Tuomas Sandholm

Angel Jordan Professor of Computer Science, Carnegie Mellon University

Co-Director, CMU AI

Founder and Director, Electronic Marketplaces Laboratory

Focus of this lecture

- Theoretically sound techniques that are
- game-independent
- …although game-specific techniques also exist
  - Johanson et al. *IJCAI*-11 faster traversal; applies at least to poker
  - Farina & Sandholm 2021 sequence form LP sparsification; applies at least to poker

TECHNIQUE 1: “ALTERNATION”

Alternation

- In CFR, instead of updating in iteration *t* both agents 1 and 2 based on the opponent's strategy in iteration *t-1*,
  - update agent 1 based on agent 2's strategy at *t-1*, and
  - then update agent 2 based on agent 1's strategy at *t*
- Motivation: updates each agent based on the newest strategy of the opponent
- Converges faster in practice
- Still provable $O(\sqrt{T})$ cumulative regret [Burch et al. *JAIR*-19]

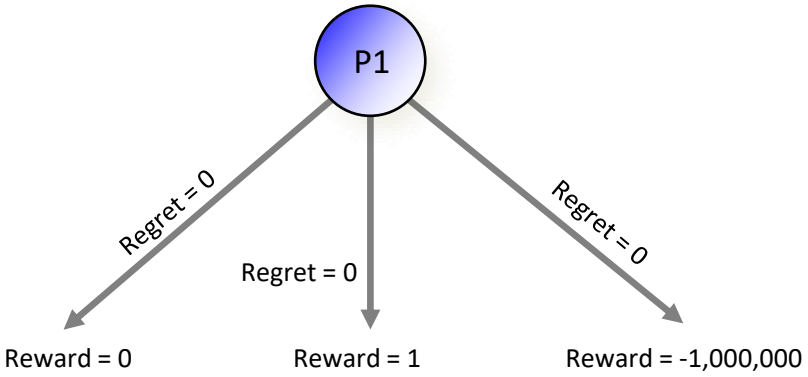
TECHNIQUE FAMILY 2: “RE-WEIGHTING”

[BROWN & SANDHOLM *AAAI*-19 DISTINGUISHED PAPER HONORABLE MENTION]

THIS WAS THE FASTEST ALGORITHM FOR 0-SUM GAMES (NORMAL- AND EXTENSIVE-FORM) AT THE TIME
…LATER IN THE COURSE WE WILL TEACH YOU SOME FURTHER IMPROVEMENTS

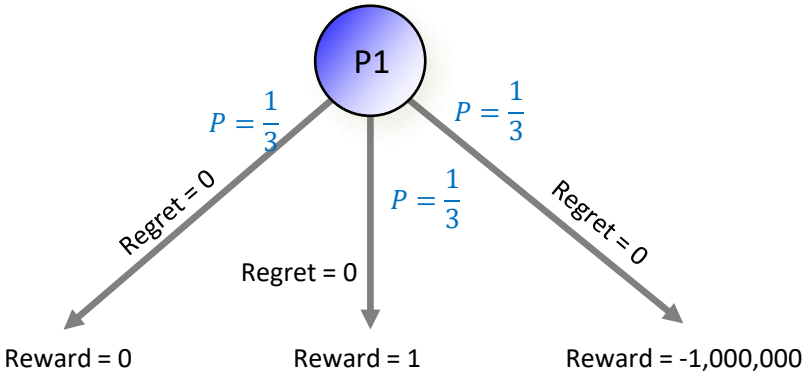
Motivation: limitations of CFR+

i.e., the previously fastest algorithm in practice



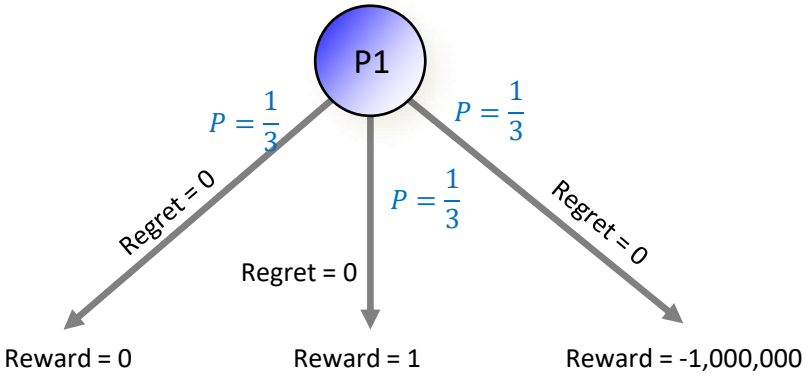
Motivation: limitations of CFR+

- On first iteration, pick all actions with equal probability



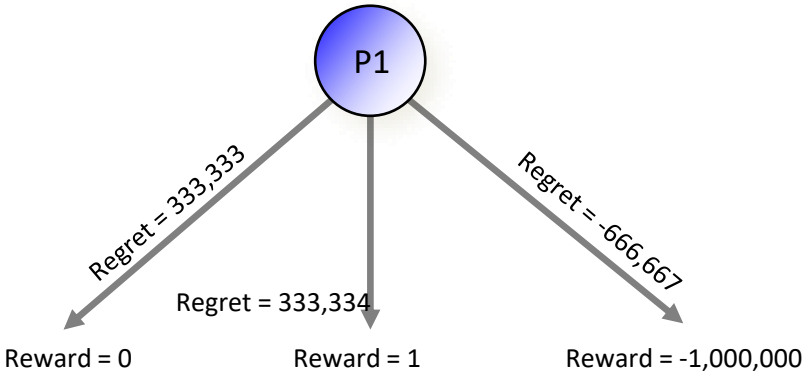
Motivation: limitations of CFR+

- On first iteration, pick all actions with equal probability
- Expected reward is -333,333



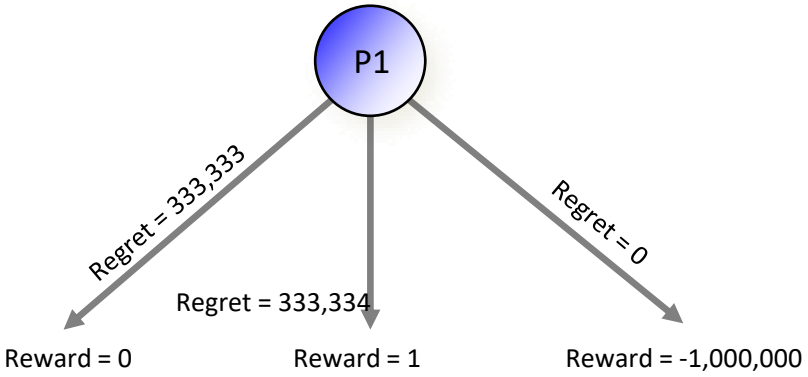
Motivation: limitations of CFR+

- On first iteration, pick all actions with equal probability
- Expected reward is -333,333
- Update regret as Action EV – Achieved EV



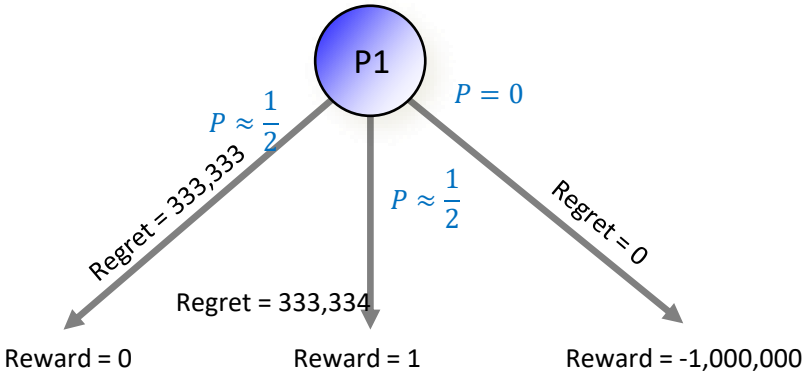
Motivation: limitations of CFR+

- On first iteration, pick all actions with equal probability
- Expected reward is -333,333
- Update regret as Action EV – Achieved EV
- CFR+ floors regret at zero



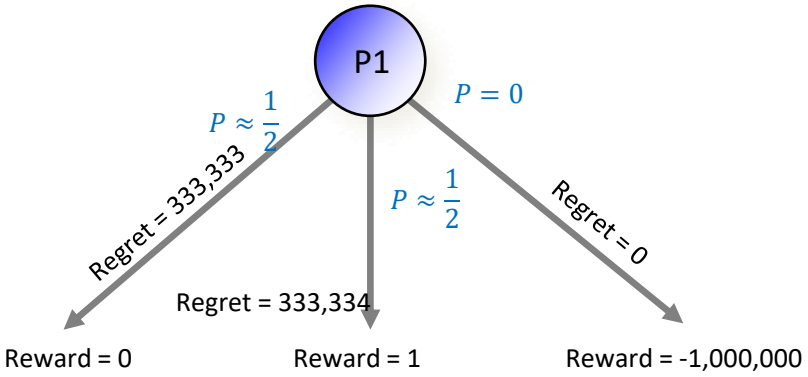
Motivation: limitations of CFR+

- On second iteration, pick actions **proportional to their regret**



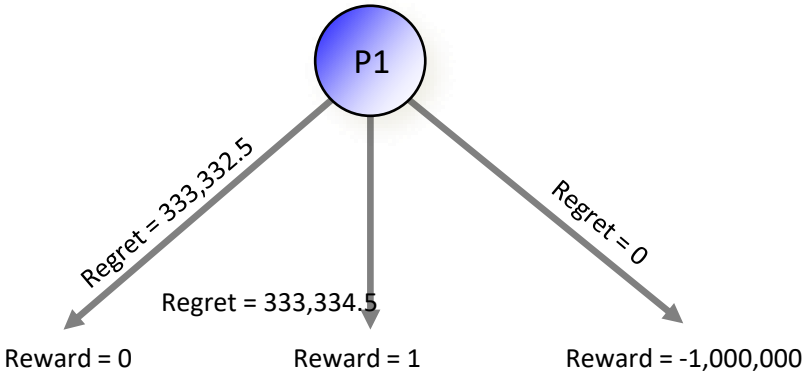
Motivation: limitations of CFR+

- On second iteration, pick actions **proportional to their regret**
- Expected reward ≈ 0.5

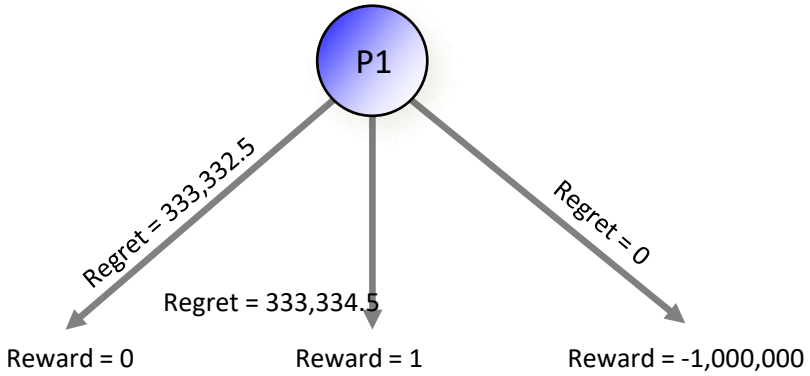


Motivation: limitations of CFR+

- On second iteration, pick actions **proportional to their regret**
- Expected reward ≈ 0.5
- Update regret



Motivation: limitations of CFR+



- Problem:
  Takes CFR+ **471,407** iterations to learn to pick the middle action with 100% probability!

- Solution:
  Discount early "bad" iterations' **regrets and average strategy** by weighing iteration t by t
  - We coin this **Linear CFR**
  - Takes only **970** iterations to learn to pick the middle action
  - Worst-case convergence bound increases by only a factor $\frac{2}{\sqrt{3}}$ …

Weighted Averaging Schemes for CFR+

- Works for any sequence of nondecreasing weights:

Theorem 1. *Suppose T iterations of RM+ are played in a two-player zero-sum game. Then the weighted average strategy profile, where iteration t is weighed proportional to $w_t > 0$ and $w_i \leq w_j$ for all $i < j$, is a $\frac{w_T}{\sum_{t=1}^{T} w_t} \Delta |\mathcal{I}| \sqrt{|A|} \sqrt{T}$-Nash equilibrium.*

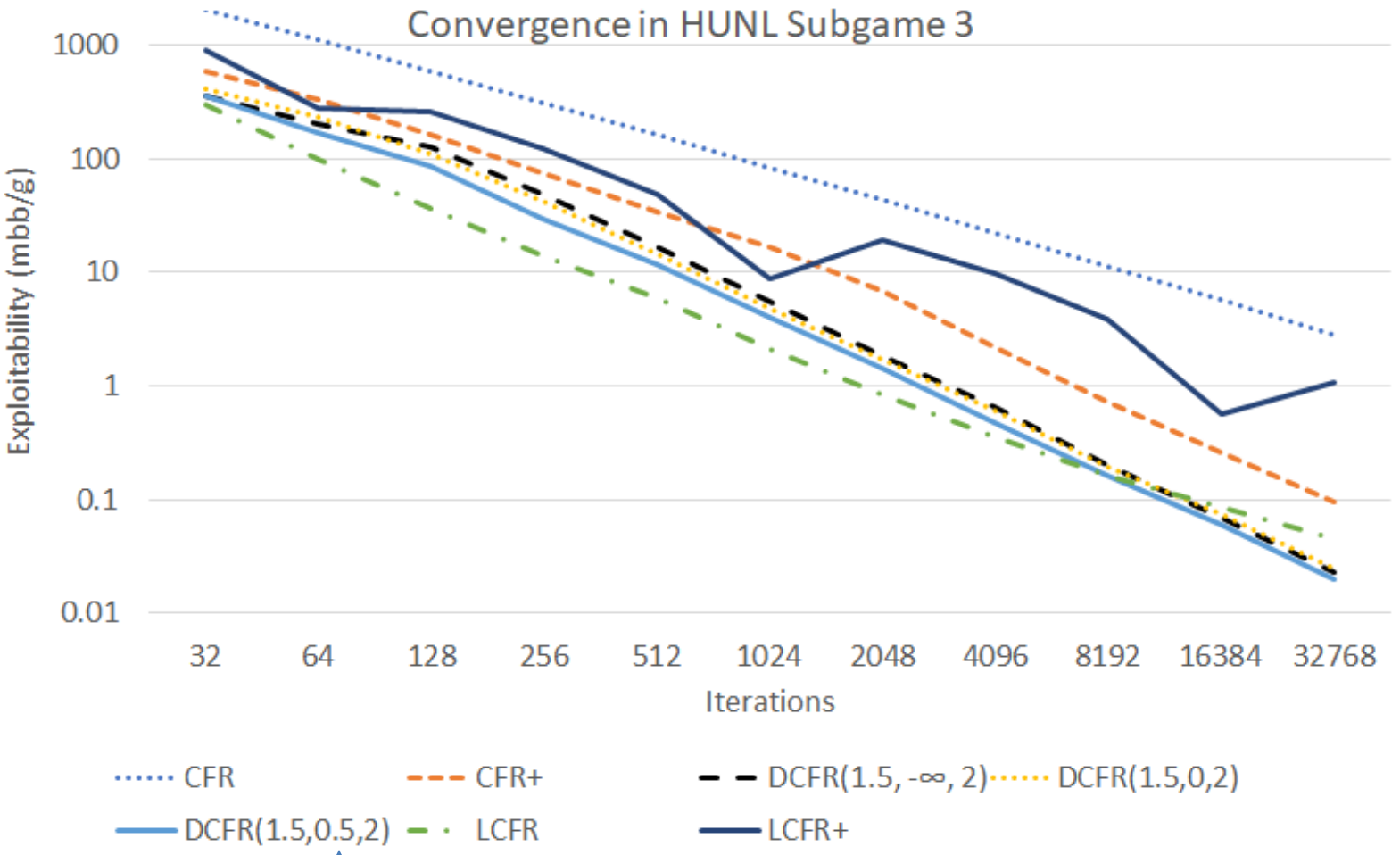
Bound is never lower than with uniform weights.

Discounted CFR

- **Linear CFR**: Weigh iteration t by t (in regrets and in averaging)
- **CFR+**: Floor regrets at zero (and weigh iteration t by t in averaging)
- Can we combine both into **Linear CFR+**?
  - Theory: Yes!
  - Practice: **No!** Does very poorly

- **But** less-aggressive combinations do well:
  Discounted CFR (DCFR)
  - On each iteration, multiply positive regrets by $\frac{t^\alpha}{t^\alpha+1}$
  - On each iteration, multiply negative regrets by $\frac{t^\beta}{t^\beta+1}$
  - Weight contributions toward average strategy by $(\frac{t}{t+1})^\gamma$
  - Worst-case convergence bound only a small constant worse than CFR
  - For $\alpha = 1.5, \beta = 0, \gamma = 2$, consistently outperforms CFR+ in practice

Experiment on heads-up no-limit Texas hold'em endgames used by *Libratus*



$\beta > 0$ facilitates regret-based pruning (to be discussed next in this lecture) because regrets can get negative, so we also show results for β =1/2.

Empirically we observed that CFR+ converges faster when assigning iteration t a weight of t^2 rather than a weight of t when calculating the average strategy. We therefore use this weight for CFR+ and its variants throughout this paper when calculating the average strategy.

Further improvements in this paper

- In Discounted CFR, can discount contributions to the *average strategy* in different ways
  - CFR+ does this in a specific way: linear
  - We show that many other ways are theoretically valid: **any nondecreasing weight sequence where final weight/sum of weights -> 0** works
    - E.g., t^x for any $x > 0$
    - Many choices in the valid space are empirically faster
- Monte Carlo Linear CFR
  - CFR+ and Discounted CFR do poorly with sampling, but Linear CFR does very well
- …

Conclusions on this paper

- Superior performance is achieved by discounting early CFR iterations
- **Discounted CFR (DCFR)** matches or exceeds CFR+ in all domains tested and was the **state-of-the-art equilibrium-finding algorithm in large imperfect-information games**
- **Linear CFR (LCFR)** does even better in games with extremely suboptimal action, but is inferior to DCFR otherwise
  - Also works well with sampling, unlike CFR+ and DCFR

TECHNIQUE FAMILY 3:
“DYNAMIC PRUNING TECHNIQUES”

Why not permanent pruning like αβ-pruning in perfect-information games?

Early idea: “Partial Pruning”

[Lanctot et al. *ICML*-09]

- If on some path, the opponent’s *reach* (i.e., probability of playing there) is 0,
  then that path can be pruned because it will not affect the regrets

- This is a no-brainer to use, but one can prune more …

Next idea: "Interval Regret-Based Pruning (Interval RBP)"

[Brown & Sandholm *NeurIP*S-15]

- While partial pruning allows one to prune paths that an *opponent* reaches with zero probability, this allows one to **also prune paths that the *agent* reaches with zero probability**
- Such pruning is necessarily temporary. Consider an action $a \in A(I)$ such that $\sigma^t(I, a) = 0$, and assume that it is known action a will not be played with positive probability until some far-future iteration t' (in RM, this would be the case *if* $R^t(I, a) << 0$)
  - To determine t', the condition on when regret might turn positive can be projected conservatively or checked dynamically
- Since action a is played with zero probability on iteration t, the strategy played and reward received following action a (that is, in $D(I, a)$) will not contribute to the regret for any information set preceding action a on iteration t. In fact, what happens in $D(I, a)$ has no bearing on the rest of the game tree until iteration t'. So one can **procrastinate** in deciding what happened beyond action a on iterations $t, ..., t' - 1$
- Upon reaching iteration t', rather than individually making up the $t' - t$ iterations over $D(I, a)$, one can do a single iteration, playing against the average of the opponents' strategies in those iterations that were missed, and declare that we played that strategy on all the missed iterations
- Moreover, since player i never plays action a with positive probability between iterations t and t', every other player can apply partial pruning on that part of the game tree for those iterations, i.e., skip it completely
- This, in turn, means that player i has free rein to play whatever she wants in $D(I, a)$ without affecting the regrets of the other players. In light of that, and of the fact that player i gets to decide what is played in $D(I, a)$ after knowing what the other players have played, player i might as well play a strategy that ensures zero regret for all information sets $I' \in D(I, a)$ in the iterations $t \ldots t'$. A CBR to the average of the opponent strategies on iterations $t \ldots t'$ would qualify as such a zero-regret strategy
  - **Definition.** A ***counterfactual best response (CBR)*** is a strategy similar to a best response, except that it maximizes counterfactual value even at information sets that it does not reach due to its earlier actions

Next idea: "Total Regret-Based Pruning (Total RBP)"

[Brown & Sandholm *ICML*-17]

- When pruning ends and regret must be updated in the pruned branch:
  - *Interval RBP* calculates a CBR to the average opponent strategy over the skipped iterations *t...t'*, and updates regret in the pruned branch as if that CBR strategy were played in those iterations
  - *Total RBP* calculates a CBR in the pruned branch against the opponent's average strategy over **all iterations *1...t'*** played so far, and sets regret as if that CBR strategy were played in all those iterations
    - Instead of CBR, an approximate CBR can be used. The paper shows how approximate it can be to still get the theoretical convergence guarantee
      - In practice CFR converges much faster than the theoretical bound, so the potential function (used in the convergence proof) is typically far lower than the theoretical bound. Thus, while choosing a near-CBR rather than an exact CBR may allow for slightly longer pruning according to the theory, it may actually result in worse performance. Clever heuristics for deciding on a near-CBR may lead to even better performance in practice

- Again, to determine *t'*, the condition on when regret might turn positive can be projected conservatively or checked dynamically (formula is in the paper)

Total RBP Has a Space Advantage

- Storing regrets:
  - **Pruned subtrees need not be stored because their regrets are computed from scratch at iteration *t'***
    - **Theorem.** For any information set *I* and action $a \in A(I)$ that is not part of a best response to a Nash equilibrium, there is an iteration $T_{I,a}$ such that for all $T \geq T_{I,a}$, action *a* in information set *I* (and its descendants) could be pruned forever
      - The algorithm won't know whether T has met the condition, so every once in a while a (near) best response computation in *D(I, a)* needs to be done, but that doesn't require storing regrets
- Storing average strategies: Fortunately, if action *a* in information set *I* is pruned for long enough, the stored cumulative strategy in *D(I, a)* can be discarded at the cost of a small increase in the distance of the final average strategy from a Nash equilibrium (by saying we never play *a* in *I*)

- **Corollary.** In a zero-sum game (with some threshold on the average strategy $C/\text{sqrt}(T)$ for $C > 0$), after a finite number of iterations, CFR with Total RBP requires only

  $O(|\#\text{infosets not pruned by above theorem}||A|)$ space

  - Can be useful, e.g., if the abstraction is grown dynamically

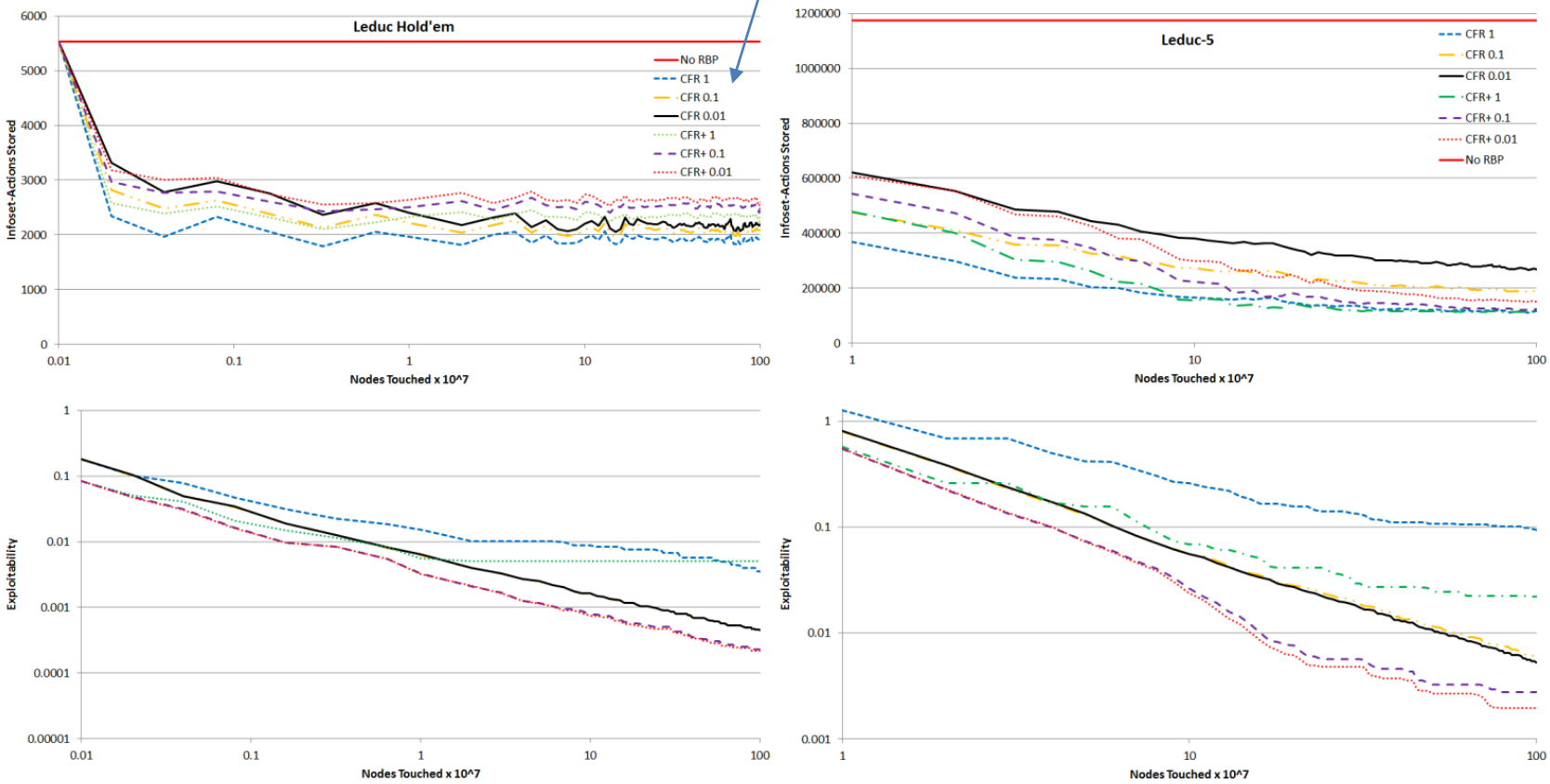
Total RBP Is Faster than CFR

- Same number of iterations
- Iterations are faster
  - Intuitively, as both players converge to a Nash equilibrium, actions that are not a CBR will eventually do worse than actions that are, so those suboptimal actions will accumulate increasing amounts of negative regret. This allows those action to be pruned for increasingly long periods
  - **Theorem.** In a zero-sum game, if both players choose strategies according to CFR with Total RBP, conducting T iterations traverses only $O(|S|T + |H| \ln(T))$ nodes

Game paths that are part of some CBR to some equilibrium

All game paths (i.e., infosets)

The number is the *C* in the criterion of setting an action to 0 probability when its average probability is less than C/sqrt(*T*). Same idea can be used also for other regret-based algorithms and even for other algorithms such as EGT [Brown, Kroer & Sandholm *AAAI*-17]

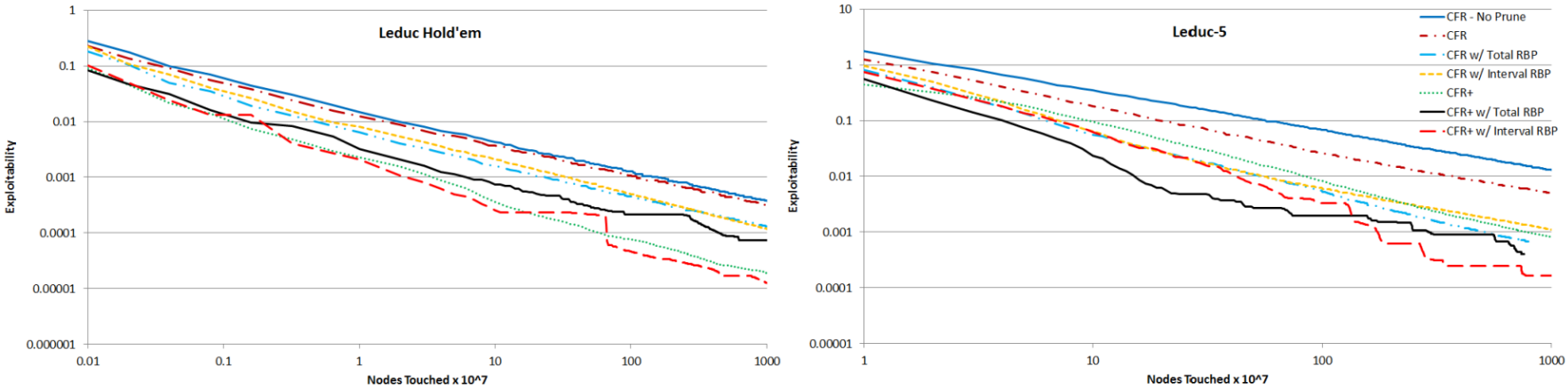


(a) Leduc Hold'em

(b) Leduc-5 Hold'em

Figure 1: Convergence and space required for CFR and CFR+ with Total RBP.

Detail about CFR+: Since Interval RBP can only prune negative-regret actions, Interval RBP modifies the definition of CFR+ so that regret can be negative, but immediately jumps up to zero as soon as regret increases. Total RBP does not require this modification. Both forms of RBP modify the behavior of CFR+ because without pruning, CFR+ would put positive probability on an action as soon as its regret increases, while RBP waits until pruning is over. CFR+'s linear weighting of the average strategy is only guaranteed to converge to a Nash equilibrium if pruning does not occur. While pruning does well empirically with CFR+, the convergence is noisy. This noise can be reduced by using the lowest-exploitability average strategy profile found so far. We did this in the experiment



(a) Leduc Hold'em

(b) Leduc-5 Hold'em

Figure 2: Convergence for CFR and CFR+ with only partial pruning, with Interval RBP, and with Total RBP. "CFR - No Prune" is CFR without any pruning.

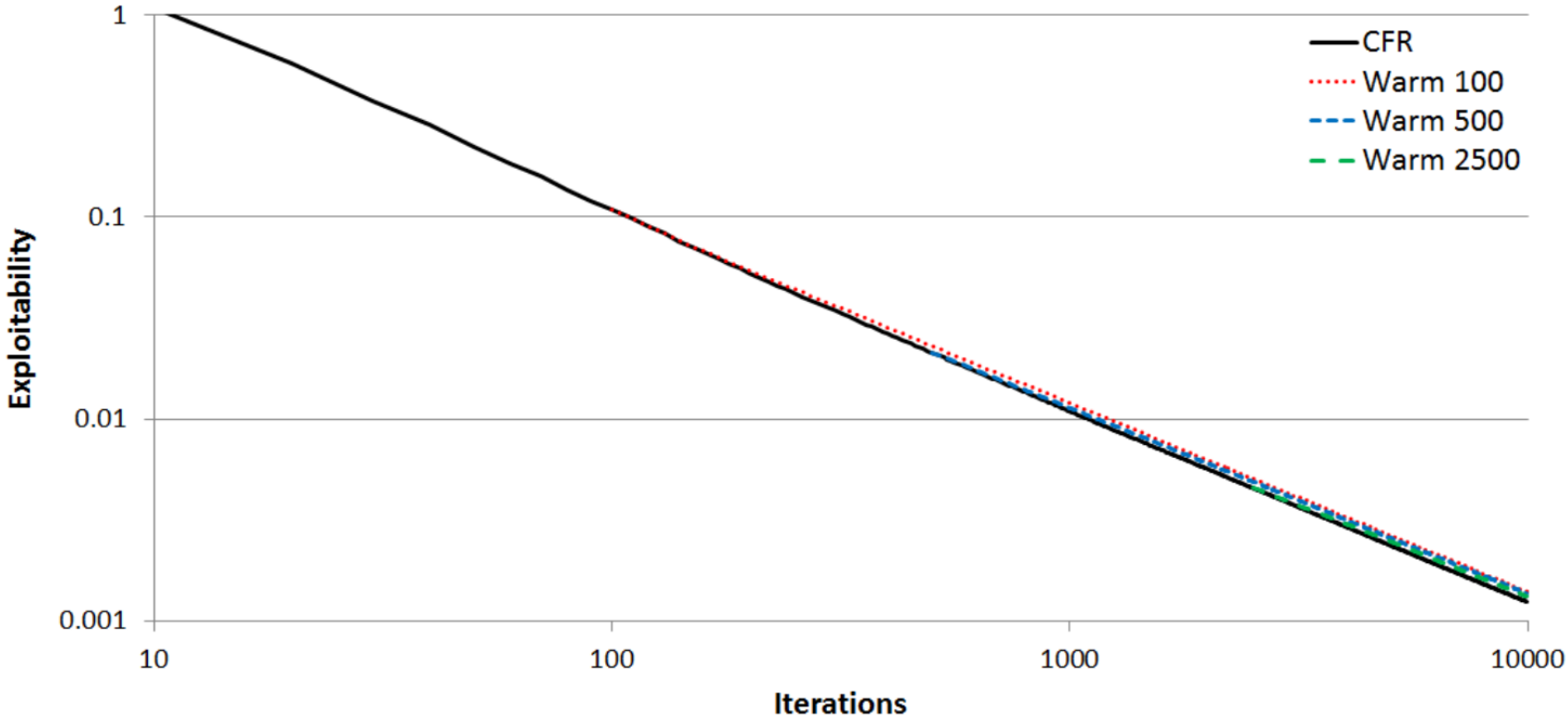
TECHNIQUE FAMILY 4:
“WARM STARTING”

Warm starting CFR from any strategies

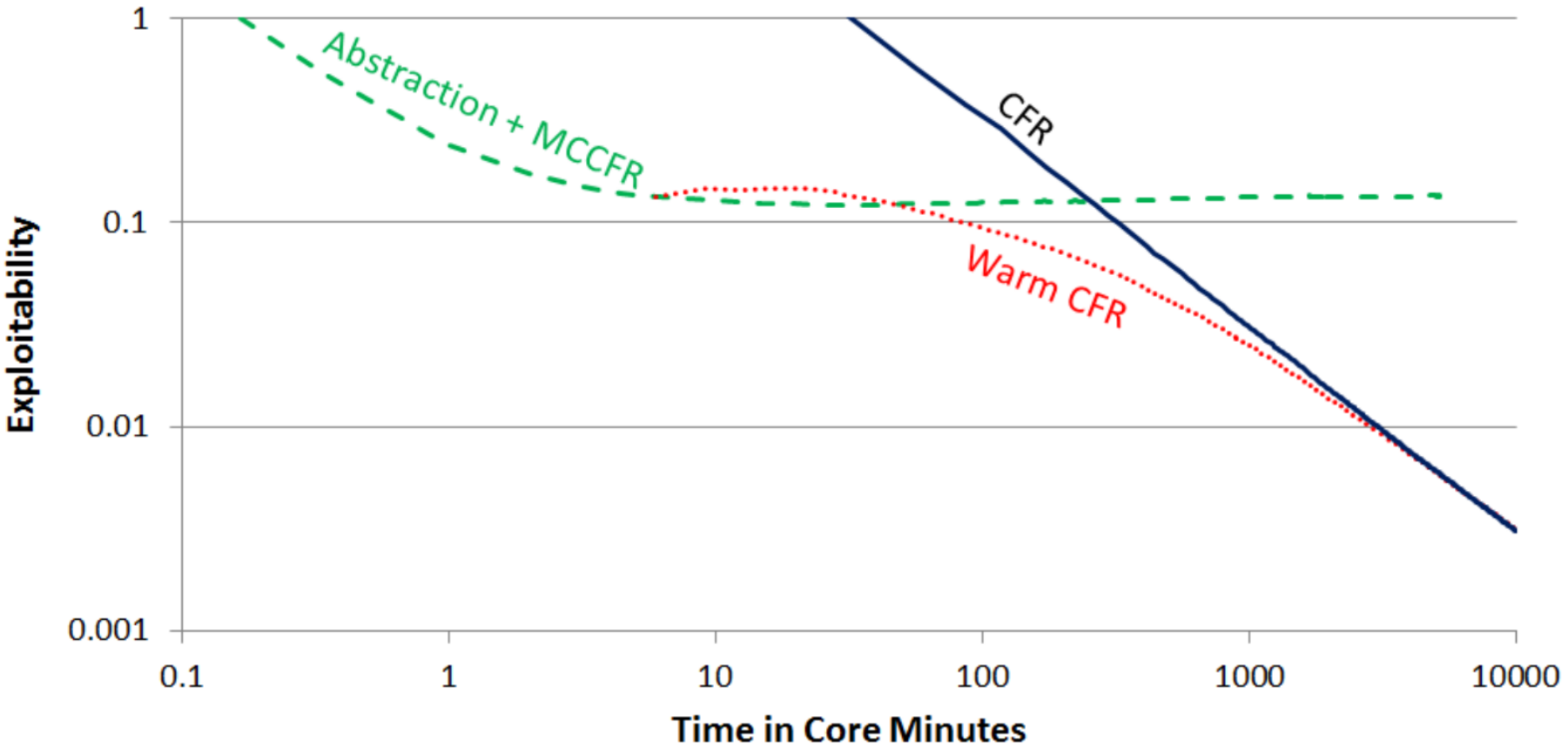
[Brown & Sandholm *AAAI*-16]

- Just starting CFR from given strategies naively doesn't help (and often hurts) compared to starting CFR from uniform strategies
  - It was believed CFR cannot be warm started, but we showed it can
- The input strategies can come from any source: hand crafted, a different algorithm, CFR run on a coarser abstraction, etc.
- Enables simultaneous abstraction and equilibrium finding
  - See also earlier paper on that [Brown & Sandholm *IJCAI*-15]
- Idea: Pretend that the input strategies came from CFR by not just setting the average strategies but also **setting the number of iterations so far *T* and the regrets appropriately**
  - The paper shows the constraints that need to be satisfied to make the CFR proof go through, and suggests practically well-performing choices within those constraints

Experiment on Flop Texas hold'em



Another experiment on Flop Texas hold'em



TECHNIQUE FAMILY 5: "OPTIMISM/PREDICTIVITY"

THIS WILL BE COVERED IN THE NEXT CLASS…