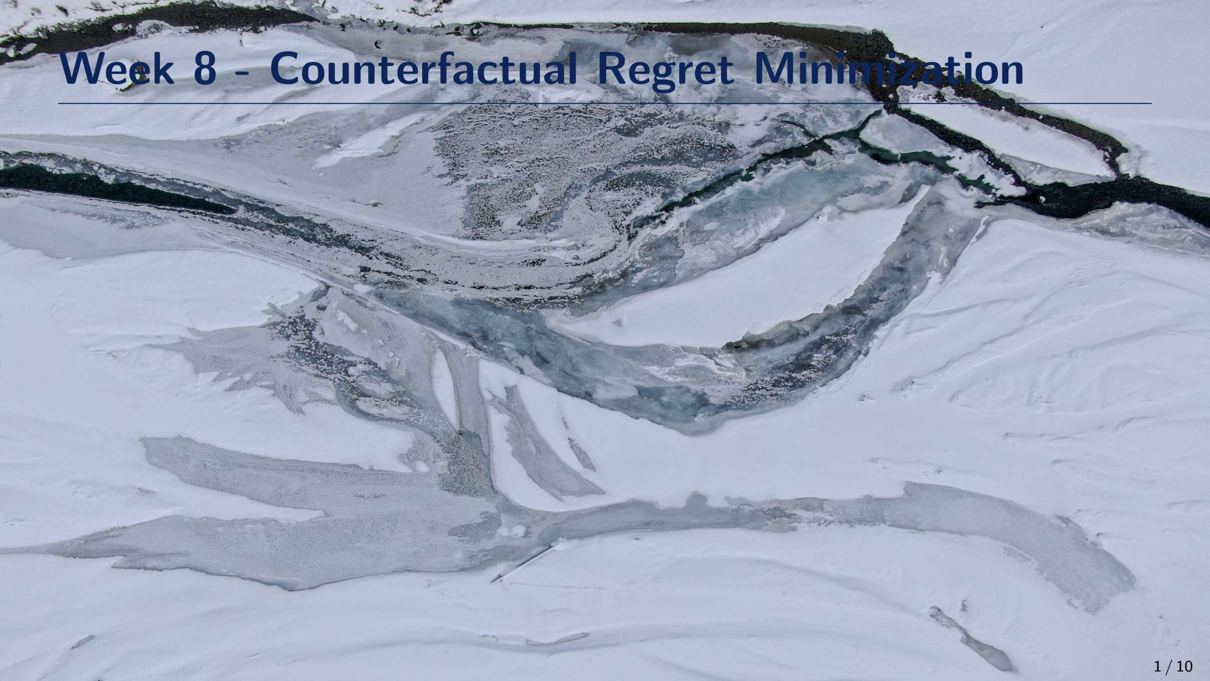
Week 8 - Counterfactual Regret Minimization



Regret Minimization In Extensive Form Games

- We have already seen application of regret minimization to matrix games (see Week 5)
- We will now see how to minimize regret in sequential games using counterfactual regret minimization (CFR) [1]
- The core of the idea is to decompose the full regret into individual per-infoset regrets that can be then minimized independently

[1] Zinkevich M, Johanson M, Bowling M, Piccione C. Regret minimization in games with incomplete information. Advances in neural information processing systems. 2007;20.

State and State-Action Values

The state value $v_i^{\pi}(h)$ is the expected future reward under policy π to player i given the history h. The state-action value $q_i^{\pi}(h, a_i)$ is defined analogously, except that the player i first takes the action a_i. Recall that, counterfactual reach $P^{\pi}_{-i}(h)$ of a given history h for player i is the reach probability of that history when player i attempts to reach h.

$$P^{\pi}_{-i}(h) := P_{\mathcal{T}}(h) \prod_{j \in \mathcal{N} \setminus \{i\}} P_j^{\pi}(h)$$

Counterfactual State-Action Values

Now we can define counterfactual-weighted state and state-action values. The counterfactual-weighted state-action value for player i of action $a \in \mathcal{A}_i$ at state $s \in \mathcal{S}_i$ is (1).

$$q_{i,c}^{\pi}(s, a) := \sum\nolimits_{h \in \mathcal{H}(s)} P_{-i}^{\pi}(h) q_i^{\pi}(h, a) \tag{1}$$

Counterfactual State Values

Having these counterfactually-weighted action-values, we define the corresponding state-values as (2).

$$v^{\pi}_{i,c}(s) := \sum\nolimits_{a \in \mathcal{A}_i(s)} \pi_i(s,a) q^{\pi}_{i,c}(s,a) \tag{2}$$

Counterfactual Regret

Given a strategy sequence $\pi^0, \ldots, \pi^{t-1}$, we can use counterfactual state and state-action values to define the counterfactual regrets as follows:

$$R_i^t(s,a) = \sum\nolimits_{k=0}^{t-1} \left(q_{i,c}^{\pi^k}(s,a) - v_{i,c}^{\pi^k}(s) \right) \tag{3}$$

$$R_i^t(s) = \max_{a \in \mathcal{A}(s)} R_i^t(s,a) \tag{4}$$

Setting $R_i^t(s)^+ = \max(R_i^t(s), 0)$, the key result is that one can bound the full regret by the sum of the positive counterfactual regrets (Theorem 6)

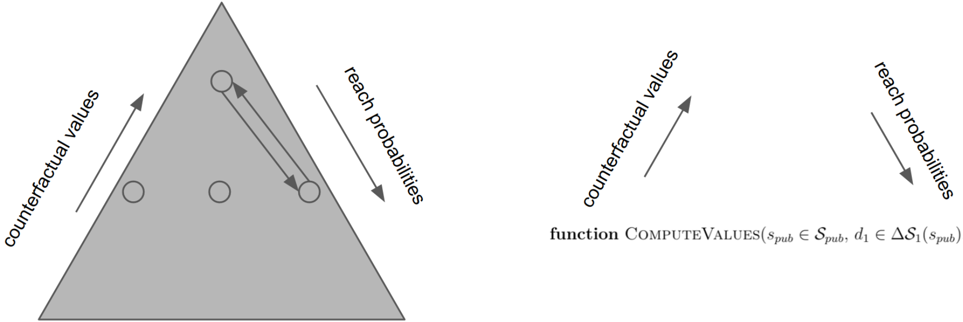
Theorem

$$R_i^t \leq \sum_{s \in \mathcal{S}_i} R_i^t(s)^+$$

Convergence

Combining the last result and Folk Theorem guarantees that if we use Hannan consistent regret minimizer in each infostate, the average strategy converges to optimal. During the averaging, one must not forget to properly weight the behavioral strategy by the current reach probability (to recall the details of that, see Figure **??**)

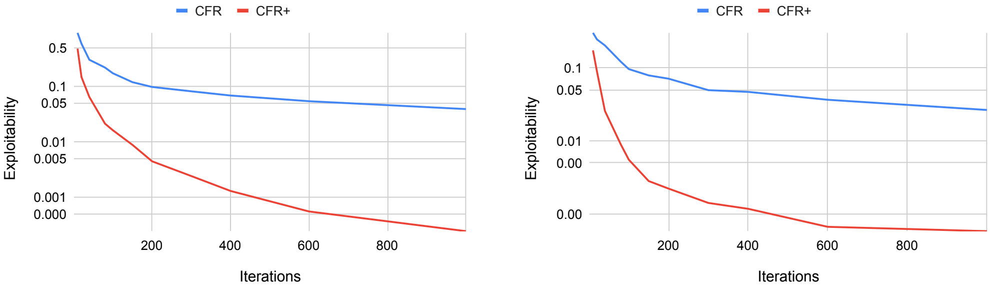
Implementation



function ComputeValues($s_{pub} \in \mathcal{S}_{pub}$, $d_1 \in \Delta\mathcal{S}_1(s_{pub})$, $d_2 \in \Delta\mathcal{S}_2(s_{pub})$)

Figure: Public tree implementation of CFR traverses the public tree and send the reach probabilities down the tree and counterfactual values up the tree.

Empirical Performance



(a) Leduc poker.

(b) Small graph chase game.

Figure: Convergence of CFR and CFR+ tabular solvers.

Week 8 Homework

This week, your task is to implement:

- the Counterfactual Regret Minimization algorithm

Once implemented, plot the exploitability of the sequence of average strategy profiles produced by the algorithm in Kuhn Poker and compare it to the Fictitious Play algorithm from Week 6.