Dresden University of Technology

Computational Logic ∴ Group

Hannes Strass

Faculty of Computer Science, Institute of Artificial Intelligence, Computational Logic Group

Counterfactual Regret Minimisation

Lecture 9, 15th Jun 2026 // Algorithmic Game Theory, SS 2026

Previously . . .

- A **belief system** assigns probabilities to histories in information sets.
- An **assessment** is a pair (behaviour strategy profile, belief system).
- A **sequentially rational** assessment plays best responses “everywhere”.
- An assessment is **weakly consistent** whenever the belief system’s probabilities match what is expected from everyone playing according to the behaviour strategy profile.
- An assessment is **consistent** iff it is the limit of a (justifying) sequence of assessments with all-positive-probability strategies.
- An assessment is a **(weak) sequential equilibrium** iff it is both sequentially rational and (weakly) consistent.
- A sequential equilibrium is **perfect** iff every player plays a best response against every opponent profile in the equilibrium’s justifying sequence.

Motivation

Main Question

- How to algorithmically solve imperfect-information games …
- …or at least devise good strategies or play them well in practice?

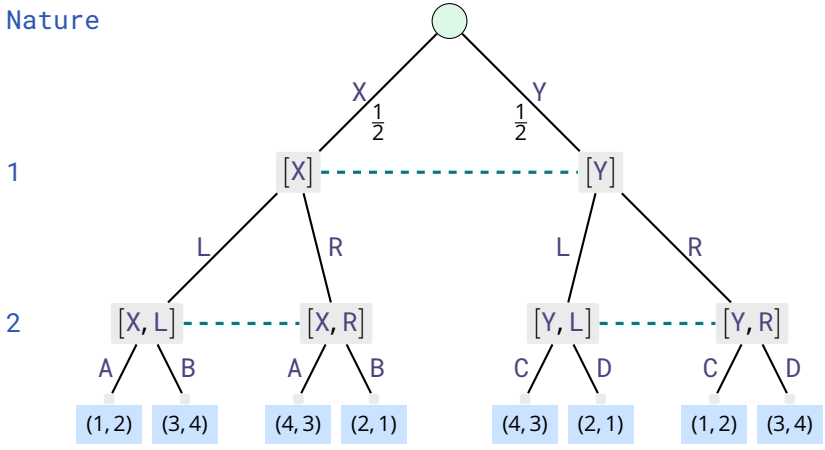
Transformation to Normal Form?

- Incurs an exponential blowup: For every player $i \in P$, there are up to $|M_i|^{|p^{-1}(i)|}$ many pure strategies in the normal-form game.
- Nash equilibria of the normal form do not always respect sequentiality.

Algorithms for sequential (perfect-information) games?

- Player i's best move in $\mathcal{I}_j \in \mathcal{I}$ depends on the player's beliefs $\beta_i : \mathcal{I}_j \to [0, 1]$.
- Consistent beliefs about $\mathcal{I}_j$ in turn depend (in general) on probabilities of moves in other information sets (even on other paths of play).

Motivation: Example



The best move for 2 in {[X, L], [X, R]} depends on what 2 does in {[Y, L], [Y, R]}:
If 2 prefers C, then 1 will prefer L and thus 2 should prefer B. (Same for D and A.)

Motivation: Regret Matching

Before minimising regret in imperfect-information extensive-form games, we start with the simpler case of normal-form games …

Recall

Let $(P, \mathbf{S}, \mathbf{u})$ be a noncooperative game in normal form, $i \in P$, and $s_j \in S_i$.
The **regret** of i playing s_j w.r.t. opponent profile $\boldsymbol{\pi}_{-i}$ is

$$r_{\boldsymbol{\pi}_{-i}, s_j} := \left(\max_{\pi_k \in \Pi_i} U_i(\boldsymbol{\pi}_{-i}, \pi_k) \right) - U_i(\boldsymbol{\pi}_{-i}, s_j)$$

Difference between what player i could have had optimally vs. what they got.

Regret is zero iff a best response is played.

$\rightsquigarrow$ Minimise regret over time in order to approach playing best responses.

Overview

Regret Matching

Counterfactual Regret Minimisation

Regret Matching

Learning to Play

Learning in Games: General Setting

- A (normal-form) game is played repeatedly for time points $t = 1, 2, \ldots$.
- After the game at time point t has ended, player (say) i has access to all strategy profiles $\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^t$ played previously, and their payoffs to i.
- Using this information, the player can devise a (mixed) strategy π_i^{t+1} to play at time point $t + 1$.

How can we evaluate whether a learner (player) is "doing well"?

Hindsight Rationality

After playing the game for $t \to \infty$ time points, the player "cannot think of" a function $\Phi\colon \Pi_i \to \Pi_i$ that would strictly increase their payoff in hindsight.

Can learning (dynamic, local) lead to equilibria (static, global)?

Regret Matching: Definition

In what follows, we assume a fixed normal-form game $G = (P, \mathbf{S}, \mathbf{u})$ to be played at time points $t = 1, 2, \ldots, T$ and take the perspective of $i \in P$.

At each time step $t \leq T$, i's one-time regret of not having played $s_k \in S_i$ is:

$$r_i^t(s_k) := u_i(s_k, \mathbf{s}_{-i}^t) - u_i(\mathbf{s}^t)$$

At time point T, the accumulated regret of a strategy $s_k \in S_i$ is thus:

$$R_i^T(s_k) := \sum_{1 \leq t \leq T} r_i^t(s_k)$$

The probabilities at $T+1$ are then set to be proportional to positive regret:

$$\pi_i^{T+1}(s_j) := \begin{cases} \dfrac{\left[R_i^T(s_j)\right]^+}{R_i^{T,+}} & \text{if } R_i^{T,+} > 0, \quad \text{where} \quad R_i^{T,+} := \displaystyle\sum_{s_k \in S_i} \left[R_i^T(s_k)\right]^+, \quad \text{for } s_j \in S_i \\ \dfrac{1}{|S_i|} & \text{otherwise.} \end{cases}$$

$([x]^+ := \max\{x, 0\} \quad \text{for all } x \in \mathbb{R}.)$

Regret Matching: Example

(Cat, Dee)	Cinema	Dancing
Cinema	(10, 7)	(2, 2)
Dancing	(0, 0)	(7, 10)

We denote $\overline{\texttt{Cinema}} = \texttt{Dancing}$ and $\overline{\texttt{Dancing}} = \texttt{Cinema}$.

T	$\mathbf{s}^T = \left(s^T_{\texttt{Cat}}, s^T_{\texttt{Dee}}\right)$	$r^T_{\texttt{Cat}}\left(\overline{s^T_{\texttt{Cat}}}\right)$	$R^T_{\texttt{Cat}}(\texttt{Cinema})$	$R^T_{\texttt{Cat}}(\texttt{Dancing})$	$\pi^{T+1}_{\texttt{Cat}}$
1	(Cinema, Dancing)	5	0	5	$\{\texttt{Cinema} \mapsto 0, \texttt{Dancing} \mapsto 1\}$
2	(Dancing, Cinema)	10	10	5	$\left\{\texttt{Cinema} \mapsto \frac{2}{3}, \texttt{Dancing} \mapsto \frac{1}{3}\right\}$
3	(Cinema, Dancing)	5	10	10	$\left\{\texttt{Cinema} \mapsto \frac{1}{2}, \texttt{Dancing} \mapsto \frac{1}{2}\right\}$
4	(Cinema, Cinema)	-10	10	0	$\{\texttt{Cinema} \mapsto 1, \texttt{Dancing} \mapsto 0\}$
5	(Cinema, Dancing)	5	10	5	$\left\{\texttt{Cinema} \mapsto \frac{2}{3}, \texttt{Dancing} \mapsto \frac{1}{3}\right\}$
6	(Cinema, Cinema)	-10	10	-5	$\{\texttt{Cinema} \mapsto 1, \texttt{Dancing} \mapsto 0\}$

Regret Matching: Correctness

For a given play sequence $(\mathbf{s}^t)_{t=1}^T$, and every $\mathbf{s}' \in \mathcal{S}$, define the **relative frequency** of $\mathbf{s}'$ after T rounds via

$$\bar{\varphi}^T(\mathbf{s}') := \frac{1}{T} \cdot \left|\left\{1 \leq t \leq T \mid \mathbf{s}^t = \mathbf{s}'\right\}\right|$$

Theorem [Hart and Mas-Colell, 2000]

Let $G = (P, \mathbf{S}, \mathbf{u})$ be a noncooperative game in normal form.
If every player plays according to regret matching, then $(\bar{\varphi}^t)_{t=1}^T$ converges to the set of correlated equilibria of G as $T \to \infty$.

More precisely: For any $\varepsilon > 0$, there is a $T_0 \geq 0$ such that for all $T > T_0$, there is a correlated equilibrium ψ_T of G whose distance from $\bar{\varphi}^T$ is at most ε.

Note: The result does not say that relative frequencies converge to a *point*.

$\rightsquigarrow$ Since all players must use regret matching, it will be used in self-play.

Regret Matching in Self-Play: Example

(Cat, Dee)	Cinema	Dancing
Cinema	(10, 7)	(2, 2)
Dancing	(0, 0)	(7, 10)

We denote $R_i^T = \left(R_i^T(\texttt{Cinema}), R_i^T(\texttt{Dancing})\right)$ for $i \in \{\texttt{Cat}, \texttt{Dee}\}$.

T	$\mathbf{s}^T = \left(s_{\texttt{Cat}}^T, s_{\texttt{Dee}}^T\right)$	$R_{\texttt{Cat}}^T$	$R_{\texttt{Dee}}^T$	$\pi_{\texttt{Cat}}^{T+1}$	$\pi_{\texttt{Dee}}^{T+1}$
1	(Cinema, Dancing)	(0,5)	(5,0)	$\{\texttt{Cinema} \mapsto 0, \texttt{Dancing} \mapsto 1\}$	$\{\texttt{Cinema} \mapsto 1, \texttt{Dancing} \mapsto 0\}$
2	(Dancing, Cinema)	(10,5)	(5,10)	$\left\{\texttt{Cinema} \mapsto \frac{2}{3}, \texttt{Dancing} \mapsto \frac{1}{3}\right\}$	$\left\{\texttt{Cinema} \mapsto \frac{1}{3}, \texttt{Dancing} \mapsto \frac{2}{3}\right\}$
3	(Cinema, Dancing)	(10,10)	(10,10)	$\left\{\texttt{Cinema} \mapsto \frac{1}{2}, \texttt{Dancing} \mapsto \frac{1}{2}\right\}$	$\left\{\texttt{Cinema} \mapsto \frac{1}{2}, \texttt{Dancing} \mapsto \frac{1}{2}\right\}$
4	(Dancing, Dancing)	(5,10)	(0,10)	$\left\{\texttt{Cinema} \mapsto \frac{1}{3}, \texttt{Dancing} \mapsto \frac{2}{3}\right\}$	$\{\texttt{Cinema} \mapsto 0, \texttt{Dancing} \mapsto 1\}$
5	(Cinema, Dancing)	(5,15)	(5,10)	$\left\{\texttt{Cinema} \mapsto \frac{1}{4}, \texttt{Dancing} \mapsto \frac{3}{4}\right\}$	$\left\{\texttt{Cinema} \mapsto \frac{1}{3}, \texttt{Dancing} \mapsto \frac{2}{3}\right\}$
6	(Dancing, Dancing)	(0,15)	(-5,10)	$\{\texttt{Cinema} \mapsto 0, \texttt{Dancing} \mapsto 1\}$	$\{\texttt{Cinema} \mapsto 0, \texttt{Dancing} \mapsto 1\}$
7	(Dancing, Dancing)	(-5,15)	(-15,10)	$\{\texttt{Cinema} \mapsto 0, \texttt{Dancing} \mapsto 1\}$	$\{\texttt{Cinema} \mapsto 0, \texttt{Dancing} \mapsto 1\}$

Rate of Convergence

For a given sequence $(\boldsymbol{\pi}^t)_{t=1}^T$ of mixed-strategy profiles, define the (external) **overall regret** of player $i \in P$ after T rounds via

$$R_i^T := \max_{\hat{\pi} \in \Pi_i} \left\{ \sum_{t=1}^{T} \left(U_i(\hat{\pi}, \boldsymbol{\pi}_{-i}^t) - U_i(\pi_i^t, \boldsymbol{\pi}_{-i}^t) \right) \right\}$$

Theorem

Let $G = (P, \mathbf{S}, \mathbf{u})$ be a normal-form game and let player $i \in P$ use regret matching in the sequence $(\boldsymbol{\pi}^t)_{t=1}^T$ of mixed-strategy profiles.

Then $R_i^T \leq \omega \cdot \sqrt{T}$, where the constant $\omega \in \mathbb{R}$ depends only on $\mathbf{u}$.

The **average overall regret** is then $\bar{R}_i^T := \frac{1}{T} \cdot R_i^T$.

Proposition

$\bar{R}_i^T$ tends to zero as $T \to \infty$ iff $\bar{\varphi}^T$ tends to the set of correlated equilibria.

The Case of Two-Player Zero-Sum Games

For a given sequence $(\boldsymbol{\pi}^t)_{t=1}^T$ of mixed-strategy profiles, define the **average mixed strategy** $\bar{\pi}_i^T$ of player $i \in P$ after T rounds via

$$\bar{\pi}_i^T(s_j) := \frac{1}{T} \cdot \sum_{t=1}^{T} \pi_i^t(s_j) \qquad \text{for } s_j \in S_i$$

Theorem

Let $G = (P, \mathbf{S}, \mathbf{u})$ be a two-player, zero-sum normal-form game, i.e. $P = \{1, 2\}$, and let $(\boldsymbol{\pi}^t)_{t=1}^T$ be obtained from both players using regret matching.
Then as $T \to \infty$, the pair $\left(\bar{\pi}_1^T, \bar{\pi}_2^T\right)$ converges to the set of Nash equilibria of G.

Regret Matching$^+$

The computation of the accumulated (possibly negative) regret of a strategy $s_k \in S_i$ can be rewritten as:

$$R_i^T(s_k) := R_i^{T-1}(s_k) + r_i^T(s_k) \qquad \text{with } R_i^0(s_k) := 0$$

Tammelin [2014] observed a better convergence when this is replaced by

$$R_i^{T,+}(s_k) := \left[R_i^{T-1,+}(s_k) + r_i^T(s_k)\right]^+$$

The probabilities at $T+1$ are again set to be proportional to positive regret:

$$\pi_i^{T+1}(s_j) := \begin{cases} \dfrac{R_i^{T,+}(s_j)}{R_i^{T,+}} & \text{if } R_i^{T,+} > 0, \quad \text{where} \quad R_i^{T,+} := \displaystyle\sum_{s_k \in S_i} R_i^{T,+}(s_k), \\ \dfrac{1}{|S_i|} & \text{otherwise.} \end{cases} \qquad \text{for } s_j \in S_i$$

RM$^+$ reacts more quickly when a previously poor action improves over time.

Counterfactual Regret Minimisation

From Normal Form to Extensive Form

Solving Imperfect-Information Games: Main Ideas

- Traverse the game tree in a backward induction-like fashion.
- Apply regret matching at each decision point (information set).

Problem

Optimal moves depend on probabilities of moves in other information sets.

Solution of Zinkevich, Johanson, Bowling, and Piccione [2007]

- Define new notion of counterfactual regret:

  Assume the player played to deliberately reach a certain information set.
- Then for games with perfect recall:
  - Regret matching can be applied to each information set independently.
  - Counterfactual regret is an upper bound for actual regret (main theorem).
  - Thus minimising counterfactual regret minimises actual regret.

Remember, Remember

Recall

$P(h' \mid h, \boldsymbol{\sigma})$ is the probability that h' is reached when playing $\boldsymbol{\sigma}$ from h on:

- $P(h \mid h, \boldsymbol{\sigma}) = 1$ for all $h \in H$,
- $P([] \mid h, \boldsymbol{\sigma}) = 0$ for all $h \neq []$, and
- $P([h'; m] \mid h, \boldsymbol{\sigma}) = \sigma_{p(\mathcal{I}_{h'})}(m \mid \mathcal{I}_{h'}) \cdot P(h' \mid h, \boldsymbol{\sigma})$.

Recall

The probability of reaching information set $\mathcal{I}_j$ when playing $\boldsymbol{\sigma}$ is thus

$$P(\mathcal{I}_j \mid \boldsymbol{\sigma}) := \sum_{h \in \mathcal{I}_j} P(h \mid \boldsymbol{\sigma}) \qquad \text{where } P(h \mid \boldsymbol{\sigma}) \text{ denotes } P(h \mid [], \boldsymbol{\sigma})$$

Recall

Player i's expected utility of playing $\boldsymbol{\sigma}$ when history h has been reached is

$$U_i(\boldsymbol{\sigma} \mid h) := \sum_{z \in Z} P(z \mid h, \boldsymbol{\sigma}) \cdot u_i(z)$$

Towards Counterfactual Regret

Definition

Consider an extensive-form game with player $i \in P$ and information sets $\mathcal{I}$.

1. The **counterfactual probability** of playing to reach $h \in H$ is given by

$$P([]\,|\,\boldsymbol{\sigma}_{-i}) = 1 \text{ and } P([h';m]\,|\,\boldsymbol{\sigma}_{-i}) := \begin{cases} \sigma_k(m\,|\,h') \cdot P(h'\,|\,\boldsymbol{\sigma}_{-i}) & \text{if } p(h') = k \neq i, \\ P(h'\,|\,\boldsymbol{\sigma}_{-i}) & \text{otherwise.} \end{cases}$$

2. The **counterfactual probability** of playing to reach $\mathcal{I}_j \in \mathcal{I}$ is

$$P(\mathcal{I}_j\,|\,\boldsymbol{\sigma}_{-i}) := \sum_{h \in \mathcal{I}_j} P(h\,|\,\boldsymbol{\sigma}_{-i})$$

3. The **counterfactual utility** of playing to reach $\mathcal{I}_j$ and then playing $\boldsymbol{\sigma}$ is

$$U_i(\boldsymbol{\sigma}\,|\,\mathcal{I}_j) = \frac{\sum_{h \in \mathcal{I}_j} P(h\,|\,\boldsymbol{\sigma}_{-i}) \cdot U_i(\boldsymbol{\sigma}\,|\,h)}{P(\mathcal{I}_j\,|\,\boldsymbol{\sigma}_{-i})} = \frac{\sum_{h \in \mathcal{I}_j} P(h\,|\,\boldsymbol{\sigma}_{-i}) \cdot \sum_{z \in Z} P(z\,|\,h,\boldsymbol{\sigma}) \cdot u_i(z)}{P(\mathcal{I}_j\,|\,\boldsymbol{\sigma}_{-i})}$$

We counterfactually assume that i intentionally played to reach $\mathcal{I}_j$.

Counterfactual Regret

Definition

Consider $i \in P$ and $\mathcal{I}_j \in \mathcal{I}$ with $p(\mathcal{I}_j) = i$.

1. Recall: The possible moves of i in $\mathcal{I}_j$ are
$$M_i(\mathcal{I}_j) := \{m \in M_i \mid [h; m] \in H \text{ for some } h \in \mathcal{I}_j\}$$
2. For behaviour strategy profile $\boldsymbol{\sigma}$ and move $m \in M_i(\mathcal{I}_j)$, define modified profile $\langle \boldsymbol{\sigma} \rangle_m^{\mathcal{I}_j}$ to be just like $\boldsymbol{\sigma}$, except that in $\mathcal{I}_j$ it always chooses m.
3. The **immediate counterfactual regret** at time T is then defined by
$$r_i^T(\mathcal{I}_j) := \max_{m^* \in M_i(\mathcal{I}_j)} \sum_{t=1}^{T} P(\mathcal{I}_j \mid \boldsymbol{\sigma}_{-i}^t) \cdot \left(U_i\left(\langle \boldsymbol{\sigma}^t \rangle_{m^*}^{\mathcal{I}_j} \,\middle|\, \mathcal{I}_j \right) - U_i(\boldsymbol{\sigma}^t \mid \mathcal{I}_j) \right)$$
for any sequence $(\boldsymbol{\sigma}^t)_{t=1}^T$ of behaviour strategy profiles.

Key Feature: $r_i^T(\mathcal{I}_j)$ can be minimised by controlling only $\sigma_i(\mathcal{I}_j) \colon M_i(\mathcal{I}_j) \to [0, 1]$.

Overall Regret $\leq$ Immediate Regret

Given sequence $(\boldsymbol{\sigma}^t)_{t=1}^T$, the (external) **overall regret** of player i at time T is:

$$R_i^T = \max_{\sigma_i^* \in \Sigma_i} \sum_{t=1}^{T} \left(U_i(\sigma_i^*, \boldsymbol{\sigma}_{-i}^t) - U_i(\boldsymbol{\sigma}^t) \right)$$

where $U_i(\boldsymbol{\sigma})$ denotes $U_i(\boldsymbol{\sigma} \mid [])$.

Theorem [Zinkevich, Johanson, Bowling, and Piccione, 2007]

In any extensive-form game with perfect recall, for any player $i \in P$ and any sequence $(\boldsymbol{\sigma}^t)_{t=1}^T$ of behaviour strategy profiles:

$$R_i^T \leq \sum_{\mathcal{I}_j \in p^{-1}(i)} \left[r_i^T(\mathcal{I}_j) \right]^+$$

Thus: Minimising immediate regret in each $\mathcal{I}_j$ minimises overall regret.

Regret Matching at Information Sets

Definition

Consider the sequence $(\boldsymbol{\sigma}^t)_{t=1}^T$ of behaviour strategy profiles of past play.

1. Let $\mathfrak{I}_j \in p^{-1}(i)$ and $m \in M_i(\mathfrak{I}_j)$. The **accumulated regret** of move m is

$$R_i^T(\mathfrak{I}_j, m) := \sum_{t=1}^{T} P(\mathfrak{I}_j \mid \boldsymbol{\sigma}_{-i}^t) \cdot \left(U_i\left(\langle \boldsymbol{\sigma}^t \rangle_m^{\mathfrak{I}_j} \,\middle|\, \mathfrak{I}_j \right) - U_i\left(\boldsymbol{\sigma}^t \mid \mathfrak{I}_j \right) \right)$$

2. The probability of playing m at $\mathfrak{I}_j$ at time $T+1$ is set to

$$\sigma_i^{T+1}(\mathfrak{I}_j)(m) := \begin{cases} \dfrac{\left[R_i^T(\mathfrak{I}_j, m)\right]^+}{R_i^{T,+}} & \text{if } R_i^{T,+} > 0, \quad \text{where } R_i^{T,+} := \displaystyle\sum_{m \in M_i(\mathfrak{I}_j)} \left[R_i^T(\mathfrak{I}_j, m)\right]^+ \\ \dfrac{1}{\left|M_i(\mathfrak{I}_j)\right|} & \text{otherwise.} \end{cases}$$

CFR: Algorithm (1)

Initialisation of global variables:

```
function init() {
    foreach i ∈ {1, 2} do {
        foreach 𝓘_j ∈ 𝓘 with p(𝓘_j) = i do {
            foreach m ∈ M_i(𝓘_j) do {
                regret[j][m] := 0                          // accumulated regret table
                strategy[j][m] := 0                        // accumulated strategy table
                profile[1][j][m] := 1/|M_i(𝓘_j)|           // move distribution for 𝓘_j at t = 1
            } } } }
```

Main Loop:

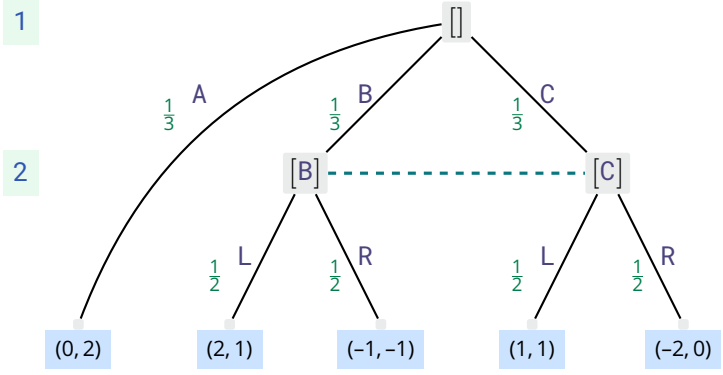
```
function solve(T) {
    foreach t ∈ {1, 2, ..., T} do {
        foreach i ∈ {1, 2} do {
            cfr([], i, t, 1, 1)
        } } }
```

CFR: Algorithm (2)

function **cfr**(h, i, t, p_1, p_2) { // history, player, time point, reach probabilities
  if IS-TERMINAL(h) **then return** UTILITY$_i$(s)
  $v_h := 0$ // initialise expected payoff at $h \in \mathcal{I}_j$
  foreach $m \in M_{p(\mathcal{I}_j)}(\mathcal{I}_j)$ **do** { $v'_h[j][m] := 0$ } // initialise payoffs of single moves
  foreach $m \in M_{p(\mathcal{I}_j)}(\mathcal{I}_j)$ **do** {
    if TURN$(h) = 1$ **then** { $v'_h[j][m] :=$ **cfr**$([h; m], i, t, profile[t][j][m] \cdot p_1, p_2)$ }
      else { $v'_h[j][m] :=$ **cfr**$([h; m], i, t, p_1, profile[t][j][m] \cdot p_2)$ }
    $v_h := v_h + profile[t][j][m] \cdot v'_h[j][m]$ } // accumulate currently expected payoff
  if TURN$(h) = i$ **then** { // players minimise immediate regret of own moves
    $r^+ := 0$ // initialise sum of positive regrets
    for $m \in M_i(\mathcal{I}_j)$ **do** { // update values needed for regret matching
      $regret[j][m] := regret[j][m] + p_{3-i} \cdot (v'_h[m] - v_h)$ // update accumulated cf regret
      $strategy[j][m] := strategy[j][m] + p_i \cdot profile[t][j][m]$ // update "frequency" of move
      $r^+ := r^+ + [regret[j][m]]^+$ } // accumulate positive regret sum for normalisation
    if $r^+ > 0$ **then** { **foreach** $m \in M_i(\mathcal{I}_j)$ **do** { // apply regret matching at $\mathcal{I}_j$
      $profile[t+1][j][m] := [regret[j][m]]^+ / r^+$ } }
    else { **foreach** $m \in M_i(\mathcal{I}_j)$ **do** {
      $profile[t+1][j][m] := 1 / |M_i(\mathcal{I}_j)|$ } } }
  return v_h }

CFR: Example

Recall the following extensive-form game G_4:



(1) Initialise move probabilities by uniform distributions

(2) Traverse game tree for $T = 1$, $i = 1$

(3) Traverse game tree for $T = 1$, $i = 2$

(4) Update move probabilities according to regret matching

CFR: Convergence

Theorem [Zinkevich, Johanson, Bowling, and Piccione, 2007]

For any extensive-form game with perfect recall, if player i selects actions according to regret matching at information sets, then

$$r_i^T(\mathcal{I}_j) \leq \omega \cdot \sqrt{\mu_i} \cdot \sqrt{T} \quad \text{whence} \quad R_i^T \leq \omega \cdot \sqrt{\mu_i} \cdot \sqrt{T} \cdot \left|p^{-1}(i)\right|$$

where $\omega \in \mathbb{R}$ only depends on $\mathbf{u}$, and $\mu_i := \max_{\mathcal{I}_j \in p^{-1}(i)} \left|M_i(\mathcal{I}_j)\right|$.

- The bound on overall regret is linear in the number of information sets.
- The overall regret grows sublinearly in T, thus we obtain:

Corollary

For the **average overall regret** $\bar{R}_i^T := \frac{1}{T} \cdot R_i$, we have $\lim_{T \to \infty} \bar{R}_i^T = 0$.

CFR: Correctness

Theorem

In any two-player, zero-sum extensive-form game with perfect recall, if both players select actions according to regret matching at information sets, then the average strategy profiles tend to the set of Nash equilibria as $T \to \infty$.

Why Nash equilibria and not sequential equilibria?

- In zero-sum games, the minimax value of the game is unique (even when there are different equilibria).
- Off-equilibrium behaviour can be adjusted without changing the outcome.
- In any case, regret matching at every information set already strives for sequential rationality …
- …but there are no known results on convergence to sequential equilibria.
- For general-sum games, convergence is only guaranteed for more permissive equilibrium concepts (e.g. correlated equilibria for extensive-form games).

CFR Algorithm: Remarks

- Histories/information sets of `Nature` can be treated in the algorithm via sampling a move from $M_{\texttt{Nature}}(\mathcal{I}_j)$ with the specified distribution.
- At each time step $t = 1, 2, \ldots, T$ (and for each $i \in P$), the call to **cfr**$([], i, t, 1, 1)$ leads to a full traversal of the game tree.
- After **solve**(T), the final values of $strategy[j][m]$ can be normalised to obtain the behaviour strategies tending towards Nash equilibria.
- Additional techniques, e.g. game abstraction, are used in practice to reduce the number of information sets (per player) to a manageable size.
- By using regret matching$^+$ in place of regret matching, we obtain CFR$^+$.
- CFR$^+$ also uses linear weighting to compute average strategies:

$$\bar{\pi}_i^{T,+}(s_j) := \frac{2}{T^2+T} \cdot \sum_{t=1}^{T} \left(t \cdot \sigma^t(s_j)\right)$$

- Bowling et al. [2015] used CFR$^+$ (with additional optimisations) to "essentially weakly solve" heads-up limit hold'em poker.

Conclusion

Summary

- The **regret** is the difference between a player's best possible strategy and their actual strategy.
- The **regret matching** algorithm uses self-play to steer average play towards the set of correlated equilibria.
- In the case of two-player zero-sum games, regret matching of both players tends towards the set of (mixed) Nash equilibria.
- The **counterfactual regret minimisation** algorithm applies regret matching to every information set of an extensive-form game with perfect recall.
- For two-player zero-sum games, the counterfactual regret minimisation (CFR) algorithm's computed strategies tend to the set of Nash equilibria.

Action Point

Implement CFR$^{(+)}$ and use it to solve Simplified Poker.