A Proofs

Lemma 6. *Let π and β be two behavioural strategies, Π and B two mixed strategies that are realization equivalent to π and β, and $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$ with $\lambda_1 + \lambda_2 = 1$. Then for each information state $u \in \mathcal{U}$,*

$$\mu(u) = \pi(u) + \frac{\lambda_2 x_\beta(\sigma_u)}{\lambda_1 x_\pi(\sigma_u) + \lambda_2 x_\beta(\sigma_u)}(\beta(u) - \pi(u))$$

defines a behavioural strategy μ at u and μ is realization equivalent to the mixed strategy $M = \lambda_1 \Pi + \lambda_2 B$.

Proof. The realization plan of $M = \lambda_1 \Pi + \lambda_2 B$ is

$$x_M(\sigma_u) = \lambda_1 x_\Pi(\sigma_u) + \lambda_2 x_B(\sigma_u), \quad \forall u \in \mathcal{U}.$$

and due to realization-equivalence, $x_\Pi(\sigma_u) = x_\pi(\sigma_u)$ and $x_B(\sigma_u) = x_\beta(\sigma_u)\ \forall u \in \mathcal{U}$. This realization plan induces a realization equivalent behavioural strategy

$$\begin{aligned}\mu(u,a) &= \frac{x_M(\sigma_u a)}{x_M(\sigma_u)} \\ &= \frac{\lambda_1 x_\pi(\sigma_u a) + \lambda_2 x_\beta(\sigma_u a)}{\lambda_1 x_\pi(\sigma_u) + \lambda_2 x_\beta(\sigma_u)} \\ &= \frac{\lambda_1 x_\pi(\sigma_u)\pi(u,a) + \lambda_2 x_\beta(\sigma_u)\beta(u,a)}{\lambda_1 x_\pi(\sigma_u) + \lambda_2 x_\beta(\sigma_u)} \\ &= \pi(u,a) + \frac{\lambda_2 x_\beta(\sigma_u)(\beta(u,a) - \pi(u,a))}{\lambda_1 x_\pi(\sigma_u) + \lambda_2 x_\beta(\sigma_u)}.\end{aligned}$$

□

Theorem 7. *Let π_1 be an initial behavioural strategy profile. The extensive-form process*

$$\beta^i_{t+1} \in b^i_{\epsilon_{t+1}}(\pi^{-i}_t),$$

$$\pi^i_{t+1}(u) = \pi^i_t(u) + \frac{\alpha_{t+1} x_{\beta^i_{t+1}}(\sigma_u)\left(\beta^i_{t+1}(u) - \pi^i_t(u)\right)}{(1-\alpha_{t+1})x_{\pi^i_t}(\sigma_u) + \alpha_{t+1} x_{\beta^i_{t+1}}(\sigma_u)}$$

for all players $i \in \mathcal{N}$ and all their information states $u \in \mathcal{U}^i$, with $\alpha_t \to 0$ and $\epsilon_t \to 0$ as $t \to \infty$, and $\sum_{t=1}^{\infty} \alpha_t = \infty$, is realization-equivalent to a generalised weakened fictitious play in the normal-form and therefore the average strategy profile converges to a Nash equilibrium in all games with the fictitious play property.

Proof. By induction. Assume π_t and Π_t are realization equivalent and $\beta_{t+1} \in b_{\epsilon_{t+1}}(\pi_t)$ is an ϵ_{t+1}-best response to π_t. By Kuhn's Theorem, let B_{t+1} be any mixed strategy that is realization equivalent to β_{t+1}. Then B_{t+1} is an ϵ_{t+1}-best response to Π_t in the normal-form. By Lemma 6, the update in behavioural policies, π_{t+1}, is realization equivalent to the following update in mixed strategies

$$\Pi_{t+1} = (1-\alpha_{t+1})\Pi_t + \alpha_{t+1} B_{t+1}$$

and thus follows a generalised weakened fictitious play. □

B Algorithms

Algorithm 3 FSP with FQI and simple counting model

Instantiate functions FICTITIOUSSELFPLAY and GENERATEDATA as in algorithm 2

function UPDATERLMEMORY$(\mathcal{M}^i_{RL}, \mathcal{D}^i)$
  $\mathcal{T} \leftarrow$ Extract from $\mathcal{D}^i$ episodes that consist of transitions $(u_t, a_t, r_{t+1}, u_{t+1})$ from player i's point of view.
  Add $\mathcal{T}$ to $\mathcal{M}^i_{RL}$, replacing oldest data if the memory is full.
  return $\mathcal{M}^i_{RL}$
end function

function UPDATESLMEMORY$(\mathcal{M}^i_{SL}, \mathcal{D}^i)$
  $\mathcal{D}^i_\beta \leftarrow$ Extract all episodes from $\mathcal{D}^i$ where player i chose their approximate best response strategy.
  $\mathcal{B} \leftarrow$ Extract from $\mathcal{D}^i_\beta$ data that consist of pairs (u_t, μ_t), where μ_t is player i's strategy at information state u_t at the time of sampling the respective episode.
  return $\mathcal{B}$
end function

function REINFORCEMENTLEARNING$(\mathcal{M}^i_{RL})$
  Initialize FQI with previous iteration's Q-values.
  $\beta \leftarrow$ FQI$(\mathcal{M}^i_{RL})$
  return β
end function

function SUPERVISEDLEARNING$(\mathcal{M}^i_{SL})$
  Initialize counting model from previous iteration.
  for each (u_t, μ_t) in $\mathcal{M}^i_{SL}$ **do**
    $\forall a \in \mathcal{A}(u_t) : N(u_t, a) \leftarrow N(u_t, a) + \mu_t(a)$
    $\forall a \in \mathcal{A}(u_t) : \pi(u_t, a) \leftarrow \frac{N(u_t,a)}{N(u_t)}$
  end for
  return π
end function

C River Poker

In our experiments, one instance of River poker implements a Texas Hold'em scenario, where the first player called a raise preflop, check/raised on the flop and bet the turn. The community cards were set to KhTc7d5sJh. The players' distributions assume that player 1 likely holds one combination of "K4s-K2s,KTo-K3o,QTo-Q9o,J9o+,T9o,T7o,98o,96o" with probability 0.99 and a uniform random holding with probability 0.01. Similarly, player 2 is likely to hold one combination of "QQ-JJ,99-88,66,AQs-A5s,K6s,K4s-K2s,QTs,Q7s,JTs,J7s,T8s+,T6s-T2s,97s,87s,72s+,AQo-A5o,K6o,K4o-K2o,QTo,Q7o,JTo,J7o,T8o+,T6o-T4o,97o,87o,75o+".