Bruno Scherrer, bruno.scherrer@inria.fr, March 16th, 2022.

In this short note, I describe a roadmap in order to prove that a deterministic mean payoff game with n states and rewards bounded by W can be solved in polynomial time (in a polynomial of n and $\log W$), along with an algorithm to do so.

Considering the Bellman equations for discounted games

$$\begin{aligned} T_{\mu,\nu} v &= r + \gamma P_{\mu,\nu} v, \\ T_\mu v &= \min_\nu T_{\mu,\nu} v, \\ T v &= \max_\mu T_\mu v, \end{aligned}$$

Writing $v_*^{(\gamma)}$ the fixed point of T, the optimal gain g_* is the limit of $(1-\gamma)v_*^{(\gamma)}$ when γ tends to 1.

1) A local Bellman certificate of optimality If $v \leq Tv$, prove that if for some state x

$$T^n v(x) - v(x) = 0,$$

(recall that n is the size of the state space), then $T^n v(x) \geq v_*(x)$.

THIS LOOKS FLAWED

In particular, if $v = v_\mu$ is the value of some policy μ (against its optimal opponent), we have $v_\mu \leq T v_\mu$, and this allows to show that $v_\mu(x) = v_*(x)$.

Intuitively, if μ_* is better than policy μ, it should get a (strictly) positive advantage in at least one of the n steps (that allow him to reach any state that is reachable from x).

2) A Policy Iteration algorithm

1. Given a stationary policy μ_k, its value v_k (against its optimal opponent), compute a n-horizon policy $\vec{\mu}$ such that

   $$T_{\vec{\mu}} v_k = T^n v_k$$

2. We compute the set of states that we know have an optimal value (through the local Bellman certificate above):

   $$Z = \{x \ ; \ T^n v_k(x) - v_k(x) = 0\}$$

   If Z is non-empty, we remove from the game the states that belong to the min-attractor set of Z (for these states, we have a policy to get to an optimal cycle through possibly a suboptimal path) and we go back to step 1 (we stop if we have removed all states).

3. Here, since Z is empty, we know that the n-periodic policy $(\vec{\mu})^\infty$ has a (strictly) positive n-step advantage (it can be shown to be bigger than $\frac{1}{n}$ by exploiting the fact that gains are rational number with denominator at most n), and thus its gain is (in all remaining states) stricly bigger than what it was.

4. We build a stationary policy μ_{k+1} that is better than $(\vec{\mu})^\infty$ (we take it greedy to the max of the values of the n rotations of $(\vec{\mu})^\infty$)

In other words, either we have states for which the value has converged (and we remove them, little by litte), or we make a significant step towards the solution.

Since the above algorithm strictly improves the gain in all (remaining) states at each iteration, and since two gains are separated at least by the value $\frac{1}{n(n-1)}$ (cf. Zwick and Paterson), the number of iteration of the above algorithm is bounded by $n(n-1)(g_{\mu_*} - g_{\mu_0})$.

Since $g_{\mu_*} - g_{\mu_0} \leq W$, this gives a pseudopolynomial algorithm. We shall reduce the dependency with respect to W by a scaline approach.

3) A Scaling approach We shall use a scaling approach and iteratively solve $\lceil \log W \rceil$ problems, At each step of this scaling approach, we update the reward as follows

$$r_{i+1} = 2r_i + \Delta r_i$$

where $0 \leq \Delta r_i \leq 1$.

We shall prove that the distance between the optimal gains of problems i and $i+1$ is bounded by 1, therefore the algorithm of the previous section takes at most $n(n-1)$ iterations to make the update. Finally, we have $n(n-1)\lceil \log W \rceil$ Policy-Iteration-like iterations (where each iteration is also polynomial)

Remarks:

- There are 2 possible approaches, either essentially work within the easiest framework of discounted problem, or writing down the algorithm and the analysis in the limit case $\gamma = 1$ (as is done for Policy Iteration for 1-player mean payoff).
- Though the local Bellman certificate can be shown for all discount factor γ, it is not clear how to deal with the general discounted case (the strict and significant positivity of the n-step advantage is only valid when γ tends to 1.)