A polynomial algorithm for deterministic mean-payoff games

Bruno Scherrer*

March 14, 2022

Abstract

We describe a polynomial algorithm for solving deterministic mean payoff games. Our algorithm solves a mean payoff game with n vertices and integer edge-costs between $-W$ and W in time ... This in particular implies that a parity game with n vertices and d priorities can be solved in time .... This answers positively the long-standing open problem whether these problems are in P.

1 Preliminaries

Consider a mean payoff game played by two players, MAX and MIN, on a graph with n vertices $X = \{1, 2, \ldots, n\} = X_+ \sqcup X_-$ and directed edges E. For any vertex x, we write $E(x) = \{y; (x, y) \in E\}$ the set of vertices that can be reached from x by following one edge. An integer cost $-R \leq r(x) \leq R$ is associated to each node x. The vertices of X_+ (resp. X_-) belong to MAX (resp. MIN). The game starts in some vertex x_0. The player who owns the current vertex chooses a next vertex by following an edge. So on and so forth, these choices generate an infinitely long trajectory $(x_0, x_1, \ldots)$. The goal of MAX is to maximize

$$\lim \inf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} r(x_t),$$

while that of MIN is to minimize

$$\lim \sup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} r(x_t).$$

Transition matrix For any pair of positional strategies $\mu : X_+ \to X$ for MAX and $\nu : X_-$ for MIN (mappings such that for all x, $\mu(x) \in E(()x)$ and $\nu(x) \in E(()x)$), let us write $P_{\mu,\nu}$ for the transition matrix: for all $(i, j) \in \{1, 2, \ldots, n\}^2$, $P_{\mu,\nu}(i, j)$ equals 1 if and only if μ and ν induce a transition $i \to j$ and 0 else.

Discounted value For any $0 < \gamma < 1$, let us introduce the following Bellman operator

$$T^{(\gamma)}_{\mu,\nu} v = r + \gamma P_{\mu,\nu} v,$$

that is a γ-contraction with respect to the max norm. The discounted value $v^{(\gamma)}_{\mu,\nu}$ when MAX and MIN respectively use μ and ν satisfies

$$v^{(\gamma)}_{\mu,\nu} = \sum_{t=0}^{\infty} (\gamma P_{\mu,\nu})^t r = (I - \gamma P_{\mu,\nu})^{-1} r$$

and is the fixed point of $T^{(\gamma)}_{\mu,\nu}$.

*INRIA, bruno.scherrer@inria.fr

Gain, bias, Laurent series expansion of the value Write

$$P^*_{\mu,\nu} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (P_{\mu,\nu})^t.$$

for the Cesaro-limit of $P_{\mu,\nu}$. For any function $g(\cdot)$ of parameter γ, we shall write

$$g(\gamma) = f_\gamma[a, b]$$

when g admits a development when $\gamma \uparrow 1$ of the form :

$$g(\gamma) = \frac{a}{1-\gamma} + b + O(1-\gamma).$$

The Laurent series expansion [3, Appendix A] tells us that:

$$(I - \gamma P_{\mu,\nu})^{-1} = f_\gamma \; (\; P^*_{\mu,\nu} \; , \; (I - P_{\mu,\nu} + P^*_{\mu,\nu})^{-1}(I - \gamma P^*_{\mu,\nu})].$$

We deduce that

$$v^{(\gamma)}_{\mu,\nu} = f_\gamma \; [\; g_{\mu,\nu} \; , \; h_{\mu,\nu} \;],$$

where $g_{\mu,\nu}$ and $h_{\mu,\nu}$ are the gain and the bias defined as

$$\begin{aligned} g_{\mu,\nu} &= P^*_{\mu,\nu} r, \\ h_{\mu,\nu} &= [I - (P_{\mu,\nu} - P^*_{\mu,\nu})]^{-1}(I - P_*)r. \end{aligned}$$

For any $(g, h) \in \mathbb{R}^2$, consider the following Bellman operators

$$T_{\mu,\nu}(g, h) = (\; P_{\mu,\nu} g, \; r + P_{\mu,\nu}(h - g) \;).$$

Given some policies (μ, ν), the gain $g_{\mu,\nu}$ and the bias $h_{\mu,\nu}$ are solutions to the following system of linear equations

$$\begin{aligned} (g, h) &= T_{\mu,\nu}(g, h), \\ w &= h + P_{\mu,\nu} w, \end{aligned}$$

where the extra-variable w ensures that the above system has a unique solution [3, Theorem 8.2.6 and Corollary 8.2.9].

Order relation and Optimality Bellman operator Consider the lexicographic order relation $\prec$ on $\mathbb{R}^2$:

$$(g, h) \prec (g', h') \quad \Leftrightarrow \quad g < g' \text{ or } (g = g' \text{ and } h < h')$$

Consider the following Bellman operators

$$\begin{aligned} T_\mu(g, h) &= \min_\nu T_{\mu,\nu}(g, h), \\ T(g, h) &= \max_\mu T_\mu(g, h), \end{aligned}$$

where the max and min operators are based on the order relation $\prec$.

The standard Policy Iteration for mean payoff games is based on the following observations:

Lemma 1. *Let μ be some policy for MAX. Let ν be an optimal counter policy for MIN and $v_{\mu,\nu} = (g_{\mu,\nu}, h_{\mu,\nu})$ be the value (gain and bias) of the resulting game. Let $\bar{\mu}$ be any policy that satisfies $T_{\bar{\mu}} v = T v$. If for some x, $v_{\mu,\nu}(x) \prec T_{\bar{\mu}} v_{\mu,\nu}(x)$, then $(g_{\mu,\nu}, h_{\mu,\nu}) \prec (g_{\bar{\mu},\bar{\nu}}, h_{\bar{\mu},\bar{\nu}})$; otherwise, μ is an optimal policy.*

For the sake of completeness we give a proof.

Proof. For any policies (μ', ν'),

$$v^{(\gamma)}_{\mu',\nu'} - v^{(\gamma)}_{\mu,\nu} = (I - \gamma P_{\mu',\nu'})^{-1}[r + (\gamma P_{\mu',\nu'} - I)v^{(\gamma)}_{\mu,\nu}].$$

Now observe that

$$\begin{aligned}
& r + (\gamma P_{\mu',\nu'} - I)\left(\frac{g_{\mu,\nu}}{1-\gamma} + h_{\mu,\nu} + O(1-\gamma)\right) \\
= \;& r + [P_{\mu',\nu'} - I - (1-\gamma)P_{\mu',\nu'})]\left(\frac{g_{\mu,\nu}}{1-\gamma} + h_{\mu,\nu} + O(1-\gamma)\right) \\
= \;& \frac{P_{\mu',\nu'}g_{\mu,\nu} - g_{\mu,\nu}}{1-\gamma} + r + P_{\mu',\nu'}(h_{\mu,\nu} - g_{\mu,\nu}) - h_{\mu,\nu} + O(1-\gamma) \\
= \;& \frac{P_{\mu',\nu'}g_{\mu,\nu} - P_{\mu,\nu}g_{\mu,\nu}}{1-\gamma} + P_{\mu',\nu'}(h_{\mu,\nu} - g_{\mu,\nu}) - P_{\mu,\nu}(h_{\mu,\nu} - g_{\mu,\nu}) + O(1-\gamma)
\end{aligned}$$

□

By taking $(\mu', \nu') = (\bar{\mu}, \bar{\nu})$, we get

2 A non-stationary Policy Iteration algorithm

We consider the following iterative algorithm:

1. (Initialization) Set $k = 0$ and take an arbitrary policy μ_0 for MAX.
2. (Evaluation) Compute the value v_k and an optimal counter-policy ν_k of MIN in the 1-player problem where MAX uses μ_k:

$$v_k = Tv_k = T_{\mu_k}v_k = T_{\mu_k,\nu_k}v_k$$

3. (Computation of the n-step advantage) Compute the advantage δ_n and a pair of n-horizon strategies $(\mu_k^{(n)}, \dots, \mu_k^{(1)})$ and $(\nu_k^{(n)}, \dots, \nu_k^{(1)})$ such that:

$$\delta_k = T^n v_k - v_k = T_{\mu_k^{(n)},\nu_k^{(n)}} T_{\mu_k^{(n-1)},\nu_k^{(n-1)}} \dots T_{\mu_k^{(1)},\nu_k^{(1)}} v_k - v_k.$$

4. (Identification of converged nodes) Compute the set

$$Z_k = \{x \in X \; ; \; \delta_k(x) = 0\}.$$

   If $Z_k \neq \emptyset$: 1) remove the nodes of the MIN-attractor A_k set of Z_k from the game (along with the transitions that go to A_k). If the game still has nodes, increment k by 1 and go to step 2 (otherwise stop).

5. (Computation of a stationary policy) Compute the values $w_k^{(n)}, \dots, w_k^{(1)}$ in the 1-player problems for MIN where MAX uses the n-periodic strategies $\sigma_k^{(n)} = (\mu_k^{(n)}, \dots, \mu_k^{(1)})$, $\sigma_k^{(n-1)} = (\mu_k^{(n-1)}, \dots, \mu_k^{(1)}, \mu_k^{(n)})$, $\dots$, $\sigma_k^{(1)} = (\mu_k^{(1)}, \mu_k^{(n)}, \dots, \mu_k^{(2)})$:

$$\begin{aligned}
w_k^{(n)} &= T_{\mu_k^{(n)}} \dots T_{\mu_k^{(1)}} w_k^{(n)}, \\
w_k^{(n-1)} &= T_{\mu_k^{(n-1)}} \dots T_{\mu_k^{(1)}} T_{\mu_k^{(n)}} w_k^{(n-1)}, \\
&\vdots \\
w_k^{(1)} &= T_{\mu_k^{(1)}} T_{\mu_k^{(n)}} \dots T_{\mu_k^{(2)}} w_k^{(1)}.
\end{aligned}$$

Compute the pointwise maximum $w_k = \max_i w_k(i)$, and identify a policy μ_{k+1} that satisfies:

$$T_{\mu_k} w_k = T w_k$$

Increment k by 1 and go to step 2.

3 A scaling approach

4 Application to the Mean Payoff game

5 Conclusion

We have shown that the problem "Mean Payoff Game" is in P. To our knowledge, this problem was only previously known to be in $NP \cap co - NP$ [4].

It was shown in [2] that any parity game (a game that is central to μ-calculus model checking) on a graph with n vertices and d priorities can be reduced to a mean payoff game on the same graph with edge costs bounded in absolute value by $W = n^d$. As a consequence, Theorem **??** implies that:

Theorem 1. *A parity game with n vertices and d priorities can be solved in time ...*

Though "Parity Game" was long thought to be in $NP \cap co - NP$ and has recently be shown to be quasi-polynomial [1], it is in fact in P.

References

[1] Cristian S. Calude, Sanjay Jain, Bakhadyr Khoussainov, Wei Li, and Frank Stephan. Deciding parity games in quasipolynomial time. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 252–263, 2017.

[2] Anuj Puri. *Theory of Hybrid Systems and Discrete Event Systems*. PhD thesis, Berkeley, CA, USA, 1995. UMI Order No. GAX96-21326.

[3] M. Puterman. *Markov Decision Processes*. Wiley, New York, 1994.

[4] Uri Zwick and Mike Paterson. The complexity of mean payoff games on graphs. *Theor. Comput. Sci.*, 158(1&2):343–359, 1996.