A polynomial algorithm for deterministic mean-payoff games

Bruno Scherrer*

March 16, 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.

Introduction

Consider a mean payoff game played by two players, MAX and MIN, on a graph with n vertices $X = \{1, 2, \dots, n\} = X_+ \sqcup X_-$ 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 $-R \le r(x) \le R$ 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 $(x_0, x_1, \dots)$. The goal of MAX is to maximize

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

while that of MIN is to minimize

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

1 Preliminaries

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 $(i, j) \in \{1, 2, \dots, n\}^2$, $P_{\mu,\nu}(i, j)$ 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 $P_{\mu,\nu}$. The Laurent series expansion [3, Appendix A] tells us that:

$$(I - \gamma P_{\mu,\nu})^{-1} = \frac{P^*_{\mu,\nu}}{1-\gamma} + Q_{\mu,\nu} + O(1-\gamma)$$

with

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

For any policies $\mu \in M$ and $\nu \in N$, let $g_{\mu,\nu}$ and $h_{\mu,\nu}$ be the gain and bias:

$$\begin{aligned} g_{\mu,\nu} &= P^*_{\mu,\nu} r, \\ h_{\mu,\nu} &= Q_{\mu,\nu} r. \end{aligned}$$

We deduce that

$$v^{[\gamma)}_{\mu,\nu} = \frac{g_{\mu,\nu}}{1-\gamma} + h_{\mu,\nu} + O(1-\gamma).$$

Mean payoff operators Consider the following operator

$$H^g_{\mu,\nu} h = r - g + P_{\mu,\nu} h.$$

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

$$\begin{aligned} g &= P_{\mu,\nu} g, \\ h &= H^g_{\mu,\nu} h, \\ w &= h + P_{\mu,\nu} w, \end{aligned}$$

where the extra-variable w ensures that the above system has as a unique solution $g_{\mu,\nu}$ and $h_{\mu,\nu}$ as defined above [3, Theorem 8.2.6 and Corollary 8.2.9].

Consider the following Bellman operators

$$\begin{aligned} G_\mu g &= \min_\nu P_{\mu,\nu} g, \\ G g &= \max_\mu G_\mu g, \\ \mathcal{G} g &= \arg\max_\mu G_\mu g, \\ H^g_\mu h &= \min_\nu H^g_{\mu,\nu} h, \\ \mathcal{H}^g_{\mathcal{M}} h &= \max_{\mu \in \mathcal{M}} H^g_\mu h. \end{aligned}$$

Lemma 1. *For any $\mu \in M$, $\nu \in N$, for any m, $\vec{\mu}' \in M^m$ and $\vec{\nu}' \in N^m$,*

$$\frac{g_{\vec{\mu}',\vec{\nu}'} - g_{\mu,\nu}}{1-\gamma} + h_{\vec{\mu}',\vec{\nu}'} - h_{\mu,\nu} + O(1-\gamma) = (I - \gamma^m P_{\vec{\mu}',\vec{\nu}'})^{-1} \left[\frac{P_{\vec{\mu}',\vec{\nu}'} g_{\mu,\nu} - g_{\mu,\nu}}{1-\gamma} + H^{P_{\vec{\mu}',\vec{\nu}'} g_{\mu,\nu}}_{\vec{\mu}',\vec{\nu}'} h_{\mu,\nu} - h_{\mu,\nu} + O(1-\gamma) \right].$$

Proof. We have for any γ,

$$v^{(\gamma)}_{\vec{\mu}',\vec{\nu}'} - v^{(\gamma)}_{\mu,\nu} = (I - \gamma^m P_{\vec{\mu}',\vec{\nu}'})^{-1}[T^{(\gamma)}_{\vec{\mu}',\vec{\nu}'}0 + (\gamma^m P_{\vec{\mu}',\vec{\nu}'} - I)v^{(\gamma)}_{\mu,\nu}].$$

Now observe that

$$\begin{aligned}
& T^{(\gamma)}_{\vec{\mu}',\vec{\nu}'}0 + (\gamma^m P_{\vec{\mu}',\vec{\nu}'} - I)v^{(\gamma)}_{\mu,\nu} \\
= \; & T^{(\gamma)}_{\vec{\mu}',\vec{\nu}'}0 + (\gamma^m P_{\vec{\mu}',\vec{\nu}'} - I)\left(\frac{g_{\mu,\nu}}{1-\gamma} + h_{\mu,\nu} + O(1-\gamma)\right) \\
= \; & T^{(\gamma)}_{\vec{\mu}',\vec{\nu}'}0 + [P_{\vec{\mu}',\vec{\nu}'} - I - (1-\gamma^m)P_{\vec{\mu}',\vec{\nu}'})]\left(\frac{g_{\mu,\nu}}{1-\gamma} + h_{\mu,\nu} + O(1-\gamma)\right) \\
= \; & \frac{P_{\vec{\mu}',\vec{\nu}'}g_{\mu,\nu} - g_{\mu,\nu}}{1-\gamma} + T^{(1)}_{\vec{\mu}',\vec{\nu}'}0 + P_{\vec{\mu}',\vec{\nu}'}(h_{\mu,\nu} - mg_{\mu,\nu}) - h_{\mu,\nu} + O(1-\gamma) \\
= \; & \frac{P_{\vec{\mu}',\vec{\nu}'}g_{\mu,\nu} - g_{\mu,\nu}}{1-\gamma} + T^{(1)}_{\vec{\mu}',\vec{\nu}'}(h_{\mu,\nu} - ng_{\mu,\nu}) - h_{\mu,\nu} + O(1-\gamma) \\
= \; & \frac{P_{\vec{\mu}',\vec{\nu}'}g_{\mu,\nu} - g_{\mu,\nu}}{1-\gamma} + H^{P_{\vec{\mu}',\vec{\nu}'}g_{\mu,\nu}}_{\vec{\mu}',\vec{\nu}'}h_{\mu,\nu} - h_{\mu,\nu} + O(1-\gamma).
\end{aligned}$$

□

Computation of a stationary policy that is better than a non-stationary policy. Compute the values $w^{(n)}_k, \ldots, w^{(1)}_k$ in the 1-player problems for MIN where MAX uses the n-periodic strategies $\sigma^{(n)}_k = (\mu^{(n)}_k, \ldots, \mu^{(1)}_k)$, $\sigma^{(n-1)}_k = (\mu^{(n-1)}_k, \ldots, \mu^{(1)}_k, \mu^{(n)}_k)$, $\ldots$, $\sigma^{(1)}_k = (\mu^{(1)}_k, \mu^{(n)}_k, \ldots, \mu^{(2)}_k)$:

$$\begin{aligned}
w^{(n)}_k &= T_{\mu^{(n)}_k} \ldots T_{\mu^{(1)}_k} w^{(n)}_k, \\
w^{(n-1)}_k &= T_{\mu^{(n-1)}_k} \ldots T_{\mu^{(1)}_k} T_{\mu^{(n)}_k} w^{(n-1)}_k, \\
\vdots & \qquad\qquad \vdots \\
w^{(1)}_k &= T_{\mu^{(1)}_k} T_{\mu^{(n)}_k} \ldots T_{\mu^{(2)}_k} w^{(1)}_k.
\end{aligned}$$

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

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

2 A Policy Iteration algorithm

We consider the following iterative algorithm:

1. (Initialization) Set $k = 0$ and take an arbitrary policy $\mu_0 \in M$ for MAX.
2. (Evaluation) Compute the optimal gain g_k and bias h_k for MIN in the 1-player problem where MAX uses μ_k by solving the system:

$$\begin{aligned}
g_k &= G_{\mu_k} g_k, \\
h_k &= H^{g_k}_{\mu_k} h_k = H^{g_k}_{\mu_k,\nu_k} h_k, \\
w_k &= h_k + P_{\mu_k,\nu_k} w_k.
\end{aligned}$$

3. (Optimisation of the n-step policy) Setting $\tilde{g}_{k,n} = g_k$, compute for $i = n-1, n-2, \ldots, 0$

$$\begin{aligned}
\tilde{g}_{k,i} &= G\tilde{g}_{k,i+1}, \\
\mathcal{M}_{k,i} &= \mathcal{G}\tilde{g}_{k,i+1}.
\end{aligned}$$

Then, compute for $i = n-1, n-2, \ldots, 0$ and identify a sequence of policies $(\tilde{\mu}_{k,n-1}, \tilde{\mu}_{k,n-2}, \ldots, \tilde{\mu}_{k,0}) \in \mathcal{M}_{k,n-1} \times \mathcal{M}_{k,n-2} \times \cdots \times \mathcal{M}_{k,0}$ such that

$$\tilde{h}_{k,i} = \mathcal{H}^{\tilde{g}_{k,0}}_{\mathcal{M}_{k,i}} \tilde{h}_{k,i+1} = H^{\tilde{g}_{k,0}}_{\tilde{\mu}_{k,i}} \tilde{h}_{k,i+1}.$$

4. (Identification of nodes with optimal gain) Compute the set

$$Z_k = \{x \in X \ ; \ (\tilde{g}_{k,0}(x), \tilde{h}_{k,0}(x)) = (g_k(x), h_k(x))\}.$$

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). 2) If the game still has nodes, increment k by 1 and go to step 2 (otherwise stop)

5. (Computation of the next stationary policy) Set

$$\mu_{k+1} = \mathit{Stationary}(\tilde{\mu}_{k,0}, \tilde{\mu}_{k,1}, \ldots, \tilde{\mu}_{k,n-1}).$$

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.