A polynomial algorithm for the deterministic mean payoff game

Bruno Scherrer

March 16, 2022

Abstract

...

We consider an infinite-horizon game on a directed graph (X, E) between two players, MAX and MIN. For any vertex x, we write $E(x) = \{y; (x, y) \in E\}$ for the set of vertices that can be reached from x by following one edge and we assume $E(x) \neq \emptyset$. The set of vertices $X = \{1, 2, \dots, n\}$ of the graph is partitionned into the sets X_+ and X_- of nodes respectively controlled by MAX and MIN. The game starts in some vertex x_0. At each time step, the player who controls the current vertex chooses a next vertex by following an edge. So on and so forth, the choices generate an infinitely long trajectory $(x_0, x_1, \dots)$. In the mean payoff game, 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

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

As a proxy to solve the mean payoff game, our technical developments will mainly consider the γ-discounted payoff for some $0 \le \gamma < 1$, where the goal of MAX is to maximize

$$\sum_{t=0}^{\infty} \gamma^t r(x_t)$$

while that of MIN is to minimize this quantity.

SUMMARY

1 Notations and Preliminaries

M N

For any pair of policies $\mu : X_+ \to X$ 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 $(x, y) \in \{1, 2, \dots, n\}^2$, $P_{\mu,\nu}(x, y)$ equals 1 if and only if μ and ν induce a transition $x \to y$ and 0 else. Seeing the reward $r : X \to 0, 1, \dots, R$ and any function $v : X \to \mathbb{R}$ as vectors of $\mathbb{R}^n$, consider the following Bellman operators

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

that are γ-contractions with respect to the max-norm. For all pairs of policies (μ, ν), the value is

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

and is the only fixed point of $T_{\mu,\nu}$. Given any policy μ for MAX, the minimal value that MIN can obtain

$$v_\mu = \min_\nu v_{\mu,\nu}$$

is the fixed point of T_μ, and it is well known that any policy ν_+ for MIN such that $T_{\mu,\nu_+} v_\mu = T_\mu v_\mu = v_\mu$ is optimal. Symmetrically, given any policy ν for MIN, the maximal value that MAX can obtain

$$\tilde{v}_\mu = \max_\nu v_{\mu,\nu}$$

is the fixed point of $\tilde{T}_\nu$, and it is well known that any policy μ_+ for MAX such that $T_{\mu_+,\nu} v_\mu = \tilde{T}_\nu \tilde{v}_\nu = \tilde{v}_\nu$ is optimal. The optimal discounted payoff

$$v_* = \max_\mu \min_\nu v_{\mu,\nu}$$

is the fixed point of T. Let (μ_*, ν_*) be any pair of positional strategies such that $T_{\mu_*,\nu_*} v_* = T v_*$. It is well-known that (μ_*, ν_*) is optimal.

n-periodic policies

Finally, min-attractor

2 A quasi-optimality equation that is local

The equation $v_* = T v_*$, that characterizes the optimal value of the game, is *global* in the sense that it is a system of equations that involves the values of *all* vertices. We shall begin by describing and prove a quasi-optimality equation that has the virtue of being *local* in the sense that it involves only the value of *one* vertex:

Lemma 1. *Let v be any value function that satisfies $v \leq Tv$. If for some x, we have*

$$[T^n v](x) - v(x) \leq \epsilon,$$

Then

$$v(x) \geq v_*(x) - \frac{\epsilon}{1-\gamma}.$$

Proof. First, observe that by the monotonicity of T, and since $v \leq Tv$, we have

$$v \leq Tv \leq T^2 v \leq \cdots \leq T^n v.$$

Let $\vec{\nu} = (\nu_1, \ldots, \nu_n)$ be a policy such that

$$T^n v = \tilde{T}_{\vec{\nu}} v.$$

Assume MIN uses $\vec{\nu}$ to play n steps against the optimal policy μ_* of MAX from x. Consider the n vertices visited:

$$x_0 = x,\ x_1,\ x_2,\ \ldots,\ x_n$$

Since there are n vertices, there necessarily exists $0 \leq i < j \leq n$ such that $x_i = x_j$. Let $\vec{\nu}_p = (\nu_1, \ldots, \nu_{i-1})$, $\vec{\nu}_c = (\nu_i, \ldots, \nu_{j-1})$ and $\vec{\nu}_{p'} = (\nu_j, \ldots, \nu_n)$ so that $\vec{\nu} = \vec{\nu}_p \vec{\nu}_c \vec{\nu}_{p'}$.

Now, assume that against μ_*, MIN uses the non-stationary policy $\vec{\nu}' = \vec{\nu}_p(\vec{\nu}_c)^\infty$. The trajectory is made of a path followed by a cycle of length $j-i$ that is repeated infinitely often:

$$\underbrace{x_0 = x,\ x_1,\ x_2,\ \ldots, x_{i-1}}_{\text{path}},\ \underbrace{x_i,\ x_{i+1},\ \ldots,\ x_{j-1}}_{\text{cycle}},\ \underbrace{x_i,\ x_{i+1},\ \ldots, x_{j-1}}_{\text{cycle}},\ \ldots$$

The value of this game satisfies for any w,

$$\begin{aligned}
v_{\mu_*,\bar{\nu}}(x) &= \mathbb{1}_x T_{\mu_*,\vec{\nu}_p} (T_{\mu_*,\vec{\nu}_c})^\infty w \\
&= \mathbb{1}_x T_{\mu_*,\vec{\nu}_p} 0 + \gamma^i \mathbb{1}_{x_i} \sum_{k=0}^{\infty} [(T_{\mu_*,\vec{\nu}_c})^{k+1} w - T_{\mu_*,\vec{\nu}_c})^k w] + \gamma^i \mathbb{1}_{x_i} w \\
&= \mathbb{1}_x T_{\mu_*,\vec{\nu}_p} w + \gamma^i \mathbb{1}_{x_i} \sum_{k=0}^{\infty} \gamma^{(j-i)k} (P_{\mu_*,\vec{\nu}_c})^k (T_{\mu_*,\vec{\nu}_c} w - w) \\
&= \mathbb{1}_x T_{\mu_*,\vec{\nu}_p} w + \gamma^i \mathbb{1}_{x_i} (I - \gamma^{j-i} P_{\mu_*,\vec{\nu}_c})^{-1} (T_{\mu_*,\vec{\nu}_c} w - w) \\
&= \mathbb{1}_x T_{\mu_*,\vec{\nu}_p} w + \frac{\gamma^i \mathbb{1}_{x_i}}{1-\gamma^{j-i}} (T_{\mu_*,\vec{\nu}_c} w - w)
\end{aligned}$$

Take $w = T^{n-j} v$ and substract $v(x)$, we obtain:

$$\begin{aligned}
v_{\mu_*,\bar{\nu}}(x) - v(x) &= \mathbb{1}_x (T_{\mu_*,\vec{\nu}_p} T^{n-j} v - v) + \frac{\gamma^i \mathbb{1}_{x_i}}{1-\gamma^{j-i}} (T_{\mu_*,\vec{\nu}_c} T^{n-i} v - T^{n-i} v) \\
&\leq \mathbb{1}_x (\tilde{T}_{\vec{\nu}_p} T^{n-j} v - v) + \frac{\gamma^i \mathbb{1}_{x_i}}{1-\gamma^{j-i}} (\tilde{T}_{\vec{\nu}_c} T^{n-j} v - T^{n-j} v) \\
&\leq \mathbb{1}_x (\tilde{T}_{\vec{\nu}_p} T^{n-i} v - v) + \frac{\gamma^i \mathbb{1}_{x_i}}{1-\gamma^{j-i}} (\tilde{T}_{\vec{\nu}_c} T^{n-j} v - T^{n-j} v) \\
&= \mathbb{1}_x (T^n v - v) + \frac{\gamma^i}{1-\gamma^{j-i}} \mathbb{1}_{x_i} (T^{n-i} v - T^{n-j} v) \\
&\leq \mathbb{1}_x (T^n v - v) + \frac{\gamma^i}{1-\gamma^{j-i}} \mathbb{1}_x (T^n v - v) \\
&\leq \frac{\epsilon}{1-\gamma}.
\end{aligned}$$

The result follows by the fact that $v_{\mu_*,\nu_*}(x) \leq v_{\mu_*,\bar{\nu}}(x)$. □

3 A non-stationary Policy Iteration algorithm

We consider the following algorithm that takes as main parameter a threshold ϵ.

$\vec{\mu}$

Lemma 2. *After at most* $\frac{n(1-\gamma)(v_{\mu_*} - v_{\mu_0})}{\epsilon}$ *iterations, the algorithm stops and return a policy* μ *such that*

$$v_{\mu_*} - v_\mu \leq nR + \frac{\epsilon}{1-\gamma}$$

If one chooses γ sufficiently high and ϵ sufficiently small of order $\frac{1}{n^2}$

By Zwick we know that

$$\|g_{\mu_*} - (1-\gamma) v^\gamma_{\mu_*}\| \leq 2n(1-\gamma)R.$$

Therefore

$$\|g_{\mu_*} - (1-\gamma)v_\mu^{(\gamma)}\| \leq 3n(1-\gamma)R + \epsilon.$$

If we choose $\gamma = 1 - \frac{1}{12n^3R}$ and $\epsilon = \frac{1}{Rn^2}$, we get

$$\|g_{\mu_*} - (1-\gamma)v_\mu^{(\gamma)}\| \leq \frac{1}{2n^2}.$$

and we can thus deduce the value g_{μ_*}.

4 A scaling variant

Theorem 1. *The total number of iterations of the Policy Iteration algorithm is bounded by $n^3 \log R$.*