A polynomial algorithm for the deterministic mean payoff game

Bruno Scherrer

March 21, 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 $X = \{1, 2, \ldots, n\}$ 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 $(x_0, x_1, \ldots)$. In the mean payoff game, 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

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

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

$$(1-\gamma) \sum_{t=0}^{\infty} \gamma^t r(x_t)$$

while that of MIN is to minimize this quantity.

LITERATURE

1 Preliminaries

Let M and N be the set of stationary policies for MAX and MIN:

$$M = \{\mu : X_+ \to X \ ; \ \forall x \in X_+,\ \mu(x) \in E(x)\},$$
$$N = \{\nu : X_- \to X \ ; \ \forall x \in X_-,\ \nu(x) \in E(x)\}.$$

For any policies $\mu \in M$ and $\nu \in N$, let us write $P_{\mu,\nu}$ for the transition matrix induced by μ and ν:

$$\forall x \in X_+, \forall y \in X, \quad P_{\mu,\nu}(x, y) = \mathbb{1}_{\mu(x)=y},$$
$$\forall x \in X_-, \forall y \in X, \quad P_{\mu,\nu}(x, y) = \mathbb{1}_{\nu(x)=y}.$$

Seeing the reward $r : X \to 0, 1, \ldots, R$ and any function $v : X \to \mathbb{R}$ as vectors of $\mathbb{R}^n$, consider the following Bellman operators

$$\begin{aligned}
T_{\mu,\nu}v &= (1-\gamma)r + \gamma P_{\mu,\nu}v, \\
T_{\mu}v &= \min_{\nu} T_{\mu,\nu}v, \\
\tilde{T}_{\nu}v &= \max_{\mu} T_{\mu,\nu}v, \\
Tv &= \max_{\mu} T_{\mu}v = \min_{\nu} \tilde{T}_{\nu}v.
\end{aligned}$$

that are γ-contractions with respect to the max-norm $\|\cdot\|$, defined for all $u \in R^n$ as $\|u\| = \max_{x \in X} |u(x)|$. For any policies $\mu \in M$ and $\nu \in N$, the value $v_{\mu,\nu}(x)$ obtained by following policies μ and ν satisfies

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

and is the only fixed point of the operator $T_{\mu,\nu}$. 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 the operator 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 value

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

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

We shall consider policies that are more complicated than usual stationary policies.

2 A local Bellman equation

The system of equations

$$\forall x, \quad v(x) = [Tv](x),$$

that characterizes the optimal value v_* of the game, is *global* in the sense that it involves the values of *all* the vertices. We shall begin by describing and prove an approximate-optimality equation that has the virtue of being *local* in the sense that it involves only *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

$$v_*(x) - [T^n v](x) \leq \frac{\epsilon}{1-\gamma}.$$

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 $n+1$ vertices visited:

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

Since there are n different vertices, by the pigeonhole principle, 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:

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

SIMPLIFY!

The value of this game satisfies for any w,

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

Taking $w = \tilde{T}_{\vec{\nu}_{p'}} v$, we obtain

$$\begin{aligned}
v_{\mu_*,\bar{\nu}}(x) - [\tilde{T}_{\vec{\nu}_{p'}} v](x) &\leq \mathbb{1}_x (\tilde{T}_{\vec{\nu}_p\vec{\nu}_c} \tilde{T}_{\vec{\nu}_{p'}} v - T_{\vec{\nu}_{p'}} v) + \frac{\gamma^j}{1-\gamma^{j-i}} \mathbb{1}_{x_i} (\tilde{T}_{\vec{\nu}_c} \tilde{T}_{\vec{\nu}_{p'}} v - \tilde{T}_{\vec{\nu}_{p'}} v) \\
&= \mathbb{1}_x (\tilde{T}_{\vec{\nu}_p\vec{\nu}_c\vec{\nu}_{p'}} v - T_{\vec{\nu}_{p'}} v) + \frac{\gamma^j}{1-\gamma^{j-i}} \mathbb{1}_{x_i} (\tilde{T}_{\vec{\nu}_c\vec{\nu}_{p'}} v - \tilde{T}_{\vec{\nu}_{p'}} v) \\
&= \mathbb{1}_x (T^n v - T^{n-j} v) + \frac{\gamma^j}{1-\gamma^{j-i}} \mathbb{1}_{x_i} (T^{n-i} v - T^{n-j} v) \\
&\leq \mathbb{1}_x (T^n v - v) + \frac{\gamma^j}{1-\gamma^{j-i}} \mathbb{1}_x (T^n v - v) \\
&\leq \frac{\epsilon}{1-\gamma},
\end{aligned}$$

where we eventually used the facts that $T^n v - v \leq \epsilon$, $j \geq 1$ and $j - i \geq 1$. The result follows by the facts that $v_*(x) = v_{\mu_*,\nu_*}(x) \leq v_{\mu_*,\bar{\nu}}(x)$ and $T^n v \geq T^{n-j} v = \tilde{T}_{\vec{\nu}_{p'}} v$. $\square$

3 A Policy Iteration procedure

Consider the extension of the binary relations $(=, <, \leq, >, \geq)$ to $\{0,1\} \times \mathbb{R}$ by using the lexicographic order. For instance,

$$(a, b) < (a', b') \quad \Leftrightarrow \quad a < a' \text{ or } (a = a' \text{ and } b < b').$$

For any $c \in \{0,1\}^n$, consider the following operators on $\{0,1\}^n \times \mathbb{R}^n$

$$\begin{aligned} U_{c,\mu,\nu}(b, v) &= (\ \min(c, P_{\mu,\nu} b),\ T_{\mu,\nu} v\ ), \\ U_{c,\mu}(b, v) &= \min_{\nu} U_{c,\mu,\nu}(b, v), \\ U_c(b, v) &= \max_{\mu} U_{c,\mu}(b, v). \end{aligned}$$

When solving the n-step problem by computing $(U_c)^n(1, v)$, the primary objective of MAX is to only visit nodes x with value $c(x) = 1$ while that of MIN is to visit at least one node x with value $c(x) = 0$; the second objective is the usual discounted value with terminal value v (that MAX wants to maximize, and MIN to minimize).

We shall consider a Policy Improvement step that takes as parameters a threshold ρ and a stationary policy μ, and that returns a non-stationary policy $(\mu_1, \ldots, \mu_n)$:

1. (Initialization) Set $C = \emptyset$. Compute the value v when MIN plays optimally against μ:

$$v = T_\mu v.$$

2. (n-step advantage) For all x, set $c(x) = \mathbb{1}_{x \notin C}$. Solve the n-step decision problem and identify $(\mu_1, \ldots, \mu_n)$ such that:

$$(U_c)^n(1, v) = U_{c,\mu_1} \ldots U_{c,\mu_n}(1, v).$$

3. (Identification of converged states) Compute the set

$$D = \{x \in X \ ;\ [T_{\mu_1} \ldots T_{\mu_n} v](x) - v(x) \leq (1-\gamma)\rho\},$$

   If D is non empty, set $C = C \cup D$. If $C \neq X$, go to step 2. Otherwise stop.

Since the size of the set C increases at each iteration, it is clear that this procedure halts after at most n iterations.

Lemma 2. *The policy $\mu_1, \ldots, \mu_n$ is such that:*

$$\begin{aligned} \forall x \in X \backslash Y, \quad v_{\mu_1,\ldots,\mu_n}(x) > v_\mu(x) + \frac{\rho}{n} \\ \forall x \in Y, \quad v_{\mu',\nu'}(x) \geq -n(1-\gamma)R + \rho. \end{aligned}$$

Proof. Consider the first inequality. When one gets to step 4 of the procedure, we know that C is empty. In other words, for all $x \in X \backslash Y$, we know that

$$\delta(x) > (1-\gamma)\rho.$$

Now, for all $x \in X \backslash Y$,

$$\begin{aligned} \mathbb{1}_x(v_{\mu',\nu'} - v_\mu) &= \mathbb{1}_x(I - \gamma^n P_{\mu_1,\nu'_1(x)} \ldots P_{\mu_n,\nu'_n(x)})^{-1}(T'_{\mu_1,\nu'_1(x)} \ldots T'_{\mu_n,\nu'_n(x)} v_\mu - v_\mu) \\ &\geq \mathbb{1}_x(I - \gamma^n P_{\mu_1,\nu'_1(x)} \ldots P_{\mu_n,\nu'_n(x)})^{-1}(T'_{\mu_1} \ldots T'_{\mu_n} v_\mu - v_\mu) \\ &= \mathbb{1}_x(I - \gamma^n P_{\mu_1,\nu'_1(x)} \ldots P_{\mu_n,\nu'_n(x)})^{-1}((T')^n v_\mu - v_\mu) \\ &= \mathbb{1}_x(I - \gamma^n P_{\mu_1,\nu'_1(x)} \ldots P_{\mu_n,\nu'_n(x)})^{-1}(\delta) \\ &> \frac{1-\gamma}{1-\gamma^n}\rho \\ &\geq \frac{\rho}{n}. \end{aligned}$$

Now consider the second inequality.

□

Theorem 1. *The Policy Iteration procedure stops after at most* $n + \frac{n(v_{\mu_*} - v_{\mu_0})}{\rho}$ *iterations, the algorithm stops and returns a policy* μ *such that*

$$v_*(x) - v_{\vec{\mu}_x}(x) \leq \rho + n(1-\gamma)R$$

Starting from $\rho_0 = \frac{W}{2}$, let us choose the sequence of parameters

$$\rho_{k+1} = \frac{\rho_k + n(1-\gamma)R}{2}$$

so that each call to the procedure lasts at most $2n$ iterations.

Then after k iterations, we have

$$\begin{aligned} v_* - v_{\mu_k} &\leq \frac{W}{2^k} + \sum_{i=0}^{k-1} \frac{1}{2^i}(1-\gamma)nR \\ &\leq \frac{W}{2^k} + 2(1-\gamma)nR \end{aligned}$$

For the mean payoff game, we have

$$\begin{aligned} \|g_* - g_{\mu_k}\| &\leq 4n(1-\gamma)R + \|v_* - v_{\mu_k}\| \\ &\leq 6n(1-\gamma)R + \frac{W}{2^k} \end{aligned}$$

When this is smaller thant $\frac{1}{n^2}$, we are done!