Un algorithme fortement polynomial pour les jeux de Markov déterministes (et les jeux de parité)

Bruno Scherrer

May 26, 2020

1 Introduction

Considérons un jeu séquentiel à 2 joueurs, "max" et "min", sur un graphe orienté ayant n nœuds/états. Cet ensemble d'états X est partitionné en 2 sous-ensembles X_{max} et X_{min} correspondant aux états contrôlés par chacun des 2 joueurs. A chaque état $x \in X$ est associé une récompense $r(x) \in \mathbb{R}$ et on note $f(x) = \{y; (x, y)\}$ l'ensemble des états qui peuvent être atteints à partir de x en suivant une arête (on suppose que la structure des graphe est telle que cet ensemble est toujours non-vide). Le jeu commence dans un état x_0. A chaque instant $t \geq 0$, le joueur qui contrôle l'état courant x_t choisit un état suivant x_{t+1} dans $f(x_t)$; ces choix successifs induisent une trajectoire infinie sur X. La *valeur* de cette trajectoire est la somme γ-actualisée ($\gamma \in [0, 1[$) des récompenses le long de cette trajctoire:

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

Le but du joueur max est de maximiser cette quantité, tandis que celui du joueur min est de la minimiser. Ce jeu est connu dans la littérature sous le nom de *Jeu de Markov déterministe* (γ-actualisé).

Soient Π_{max} et Π_{min} les ensembles de stratégies *stationnaires déterministes* pour les deux joueurs :

$$\Pi_{max} = \{\mu : X_{max} \to X \ ; \ \forall x \in X_{max}, \ \mu(x) \in f(x)\}$$
$$\Pi_{min} = \{\nu : X_{min} \to X \ ; \ \forall x \in X_{min}, \ \nu(x) \in f(x)\}.$$

Soit de plus $\Pi = \Pi_{max} \times \Pi_{min}$ l'ensemble des couples de telles stratégies.

Supposons que les états sont numérotés de 1 à n. Identifions toute fonction de X vers $\mathbb{R}$ à un vecteur de $\mathbb{R}^n = \mathbb{R}^{|X|}$. Prenons un couple de stratégies $(\mu, \nu) \in \Pi$. On notera $P_{\mu,\nu}$ la matrice de transition induite par ce choix de stratégies:

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

On note $T_{\mu,\nu}$ l'opérateur affine de Bellman associé au couple de stratégies (μ, ν):

$$\forall v \in \mathbb{R}^n, \quad T_{\mu,\nu} v = r + \gamma P_{\mu,\nu} v.$$

Il est connu (et assez facile de voir) que cet opérateur a pour unique point fixe une fonction qui à tout état x associe la valeur de la trajectoire induite par (μ, ν), fonction que nous noterons $v_{\mu,\nu}$.

On introduit de plus les trois opérateurs de Bellman suivants:

$$\forall \mu \in \Pi_{max}, \ \forall v \in \mathbb{R}^n, \quad T_\mu v = \min_\nu T_{\mu,\nu} v,$$
$$\forall \nu \in \Pi_{min}, \ \forall v \in \mathbb{R}^n, \quad \tilde{T}_\nu v = \max_\mu T_{\mu,\nu} v,$$
$$\forall v \in \mathbb{R}^n, \quad T v = \max_\mu T_\mu v = \min_\nu \tilde{T}_\nu v,$$

où les min et max de vecteurs sont effectués composante par composante. L'égalité entre les deux formulations du dernier opérateur est une conséquence bien connue du théorème du minimax de Von Neumann.

Il est bien connu que la valeur d'équilibre du jeu pour un état initial x est $v_*(x)$ où v_* est l'unique point fixe de l'opérateur T. Par ailleurs, tout couple de stratégies stationnaires (μ_*, ν_*) satisfaisant

$$v_* = T_{\mu_*} v_* = \tilde{T}_{\nu_*} v_* = T_{\mu_*,\nu_*} v_*.$$

est un couple de stratégies optimales qui atteint cette valeur d'équilibre et constitue ainsi une solution au jeu. En supposant un ordre (arbitraire) sur les états lorsqu'on choisit les argmin et argmax, on peut toujours faire en sorte que le couple de stratégies optimales soit unique.

Notations:

$$v_{\mu,\nu} = (T_{\mu,\nu})^\infty$$

Algo PI: on choisit μ_{k+1} tel que

$$T_{\mu_{k+1}}(T_{\mu_k})^\infty = T(T_{\mu_k})^\infty$$

Alors, la preuve de la croissance de PI est:

$$\begin{aligned}(T_{\mu_{k+1}})^\infty - (T_{\mu_k})^\infty &= \sum_{i=0}^{\infty} (T_{\mu_{k+1}})^{i+1}(T_{\mu_k})^\infty - (T_{\mu_{k+1}})^i (T_{\mu_k})^\infty \\ &= \sum_{i=0}^{\infty} (T_{\mu_{k+1}})^i T(T_{\mu_k})^\infty - (T_{\mu_{k+1}})^i T_{\mu_k}(T_{\mu_k})^\infty \\ &\geq 0\end{aligned}$$

par monotonicité de $(T_{\mu_{k+1}})^i$.

Idée de l'algo: On a μ politique du joueur MAX, ν meilleur réponse à μ:

$$(T_{\mu,\nu})^\infty = T_\mu^\infty$$

On cherche des séquences de politiques $\mu_1, \mu_2, \ldots, \mu_k$ et $\nu_1, \nu_2, \ldots, \nu_k$, un état x et $c \leq k$, tels que:

$$\begin{aligned}T_{\mu^c,\nu^c} T^{k-c}(T_\mu)^\infty &= T^c T^{k-c}(T_\mu)^\infty, \\ \mathbb{1}_x P_{\mu^c,\nu^c} &= \mathbb{1}_x\end{aligned}$$

Alors, par monotonicité,

$$(T_{\mu^c,\nu^c})^\infty \geq T^{k-c}(T_\mu)^\infty$$

2 Algorithme

L'algorithme que nous allons considérer peut être vu comme une variation de l'algorithme *itérations sur les politiques* de Howard. On se place du point de vue du joueur max, et on va itérer dans l'espace des politiques périodiques:

$$\mu^{(k)} = (\mu_0^{(k)}, \mu_1^{(k)}, \ldots, \mu_{p_k-1}^{(k)}).$$

Initialiement, on choisit une politique quelconque (on peut par exemple prendre n'importe quelle politique stationnaire).

A chaque itération k, on effectue successivement les deux étapes suivantes.

- **Evaluation de la politique:** On calcule la valeur de la politique $\mu^{(k)}$ lorsqu'il joue contre son meilleur adversaire. C'est un problème de décision déterministe 1-joueur dont l'unique solution de l'équation *point-fixe*,

$$v_k = T_{\mu_0^{(k)}} T_{\mu_1^{(k)}} \ldots T_{\mu_{p_k-1}^{(k)}} v_k.$$

- **Détermination d'une nouvelle politique:** On considère un jeu auxiliaire à horizon fini n, de valeur terminale v_k, dont on calcule la valeur optimale w_0 et le couple de stratégies optimales

$$\big((\mu_0, \mu_1, \ldots, \mu_{n-1}), (\nu_0, \nu_1, \ldots, \nu_{n-1})\big)$$

avec n étapes de l'algorithme *itération sur les valeurs*:

$$w_n = v_k,$$
$$\forall j \in \{0, 1, \ldots, n-1\}, \quad w_j = T w_{j+1} = T_{\mu_j} w_{j+1} = \tilde{T}_{\nu_j} w_{j+1} = T_{\mu_j, \nu_j} w_{j+1}.$$

Pour chaque état initial x, le couple de stratégies optimales induit un chemin min-max optimal ($x = y_0, y_1, y_2, \ldots y_n$) pour le problème auxiliaire. Considérons l'ensemble des "boucles" des chemins partant de x:

$$B_x = \{(y, i, j) \ ; \ 0 \leq i < j \leq n \text{ tels que } y_i = y_j = y\}.$$

Par le principe des tiroirs cet ensemble contient toujours au moins un élément. A chaque boucle $(y, i, j) \in B_x$, on associe le score:

$$w_i(y) - w_j(y).$$

Comme nous le verrons plus loin, ce score est nécessairement positif ou nul. S'il existe un état x dont une boucle (y, i, j) a un score *strictement* positif, alors on peut prendre comme prochaine politique non-stationnaire à évaluer

$$\mu^{(k+1)} = (\mu_i, \mu_{i+1}, \ldots, \mu_{j-1}).$$

Sinon, l'algorithme est terminé et on renvoie la valeur w_0 qui est, comme nous allons le prouver, égale à la valeur optimale v_*.

3 Exactitude de l'algorithme

Lemme 1. *A chaque itération k, pour tout état initial x, pour toute boucle $(y, i, j) \in B_x$, le score $w_i(y) - w_j(y)$ d'une boucle est positif.*

Proof. Pour tout $j \in 0, 1, \ldots, n-1$,

$$\begin{aligned} w_j &= T^{n-j} v_j \\ &= T_{\mu_j} T_{\mu_{j+1}} \ldots T_{\mu_{n-1}} (T_{\mu^{(k)}})^{\infty} 0. \end{aligned}$$

□

Lemme 2. *La séquence de fonctions (v_k) est croissante (à chaque étape, la croissance est stricte pour au moins un état x).*

Lemme 3. *score nul équivaut à satisfaction de l'équation de Bellman*

4 Nombre d'itérations

$$\Pi = \cup_c \Pi_c = \cup_{x,c} \Pi_c(x)$$

On considère des politiques non-stationnaires:

$$\mathfrak{M}_c(x) = \{\mu; (\mu, M(\mu)) \in \Pi_c(x)\}$$
$$\mathfrak{N}_c(x) = \{\nu; (N(\nu), \nu) \in \Pi_c(x)\}$$

$$v_c(x) = \max_{\mu \in \mathfrak{M}_c(x)} v_\mu(x)$$
$$\tilde{v}_c(x) = \min_{\nu \in \mathfrak{N}_c(x)} \tilde{v}_\nu(x)$$

Par construction, il est clair que pour tout c,

$$\begin{aligned} \forall x, \quad & v_c(x) \leq \tilde{v}_c(x), \\ \forall x \in \mathcal{C}(\mu^*, \nu^*), \quad & v_c(x) \leq v^*(x) \leq \tilde{v}_c(x). \end{aligned}$$

Lemme 4. *Pour tout v, pour tout $(\mu, \nu) \in \Pi_c$,*

$$v_{\mu,\nu} - v = [I - (\gamma P_{\mu,\nu})^c]^{-1}(T_{\mu,\nu} v - v).$$

Pour tout v, pour tout $(\mu, \nu) \in \Pi_c(x)$,

$$\mathbb{1}'_x(v_{\mu,\nu} - v) = \frac{1}{1-\gamma^c} \mathbb{1}'_x(T_{\mu,\nu} v - v).$$

Lemme 5 (Le super pouvoir de Value Iteration)**.** *Soient v et $(\mu, \nu) \in \Pi_c(x)$ tels que $T_{\mu,\nu} v = T^c v$. Alors:*

$$v_c(x) = \tilde{v}_c(x).$$

Proof.

$$\begin{aligned} \max_{\mu' \in \mathfrak{M}_c(x)} \mathbb{1}'_x(v_{\mu'} - v) &= \frac{1}{1-\gamma^c} \max_{\mu' \in \mathfrak{M}_c(x)} \mathbb{1}'_x(T_{\mu'} v - v) \\ &= \frac{1}{1-\gamma^c} \mathbb{1}'_x(\max_{\mu' \in \mathfrak{M}_c(x)} T_{\mu'} v - v) \\ &= \frac{1}{1-\gamma^c} \mathbb{1}'_x(T_\mu v - v) \\ &= \frac{1}{1-\gamma^c} \mathbb{1}'_x(T^c v - v), \end{aligned}$$

et symmétriquement pour

$$\min_{\nu' \in \mathfrak{N}_c(x)} \mathbb{1}'_x(\tilde{v}_{\nu'} - v),$$

d'où le résultat. □

5 Conséquences