Quelques majorants de la complexité de l'algorithme itérations sur les politiques ⋆

Bruno Scherrer

Equipe Maia
Inria, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France
bruno.scherrer@inria.fr

Résumé : Étant donné un processus de décision Markovien (PDM) avec n états et m actions, nous étudions le nombre d'étapes requises par l'algorithme itérations sur les politiques (IP) pour converger vers la politique optimale γ actualisée. Nous considérons deux variations d'IP : IP-Howard qui change les actions dans tous les états qui ont un avantage strictement positif, et IP-Simplexe qui change uniquement une action dans l'état qui a l'avantage maximal. Nous montrons que IP-Howard termine après au plus

$$n(m-1)\left\lceil \frac{1}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right\rceil = O\left(\frac{nm}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right)$$

itérations, améliorant d'un facteur $O(\log n)$ un résultat de Hansen *et al.* (2013), tandis que IP-Simplexe termine après au plus

$$n^2(m-1)\left(1+\frac{2}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right) = O\left(\frac{n^2m}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right)$$

itérations, améliorant d'un facteur $O(\log n)$ un résultat de Ye (2011). Sous des hypothèses structurelles du PDM, nous considérons ensuite des majorants qui sont indépendants du facteur d'actualisation γ : étant données une mesure τ_t du temps maximal pour quitter les zones transientes et une mesure τ_r du temps maximal pour revisiter les états dans les classes récurrentes sous l'ensemble des politiques, nous montrons que IP-Simplexe termine après au plus

$$n^2(m-1)\left(\lceil \tau_r \log(n\tau_r)\rceil + \lceil \tau_r \log(n\tau_t)\rceil\right)\left[(m-1)\lceil n\tau_t \log(n\tau_t)\rceil + \lceil n\tau_t \log(n^2\tau_t)\rceil\right] = \tilde{O}\left(n^3m^2\tau_t\tau_r\right)$$

itérations. Ceci généralise un résultat récent de Post & Ye (2012) sur les PDMs déterministes dans lesquels on a $\tau_t = \tau_r = n$. Nous expliquons pourquoi des résultats analogues semblent difficiles à obtenir pour IP-Howard. Enfin, sous l'hypothèse supplémentaire (restrictive) que l'espace d'états est partitionné en deux ensembles, correspondant aux états transients (respectivement récurrents) pour toutes les politiques, nous montrons que IP-Howard termine après au plus

$$n(m-1)\left(\lceil \tau_t \log n\tau_t\rceil + \lceil \tau_r \log n\tau_r\rceil\right) = \tilde{O}(nm(\tau_t+\tau_r))$$

itérations, tandis que IP-Simplexe termine après au plus

$$n(m-1)\left(\lceil n\tau_t \log n\tau_t\rceil + \lceil \tau_r \log n\tau_r\rceil\right) = \tilde{O}(n^2m(\tau_t+\tau_r))$$

itérations.

1 Introduction

Nous considérons un système dynamique contrôlé à temps discret. Nous supposons que ce système vit dans un **espace d'états** X de taille finie n. Dans chaque état $i \in \{1,..,n\}$, le contrôle est choisi dans un **espace d'actions** A de taille finie [1] m. Le contrôle $a \in A$ spécifie la **probabilité de transition** $p_{ij}(a) =$

⋆. Cet article est une version courte française du rapport technique (Scherrer, 2013). Nous renvoyons notamment le lecteur à ce rapport pour les preuves des nouveaux résultats énoncés ici.

1. Dans les travaux de Ye (2011); Post & Ye (2012); Hansen *et al.* (2013) que nous référençons, l'entier "m" correspond au nombre total d'actions, c'est-à-dire à nm avec nos notations. Quand nous énoncerons leurs résultats, nous le ferons avec notre propre notation, soit en remplaçant leur "m" par "nm".

$\mathbb{P}(i_{t+1} = j | i_t = i, a_t = a)$ vers un nouvel état j. A chaque transition, le système reçoit une récompense $r(i, a, j)$, où $r : X \to \mathbb{R}$ est la **fonction de récompense** instantanée. Dans ce contexte, nous cherchons une politique déterministe et stationnaire (une fonction $\pi : X \to A$ qui associe un contrôle à chaque état [2]) qui maximise l'espérance de la somme actualisée des récompenses à partir de tout état i, quantité appelée la **fonction valeur de la politique** π en l'état i :

$$v_\pi(i) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(i_k, a_k, i_{k+1}) \middle| i_0 = i,\ \forall k \geq 0,\ a_k = \pi(i_k),\ i_{k+1} \sim \mathbb{P}(\cdot|i_k, a_k)\right]$$

où $\gamma \in [0, 1]$ est un facteur d'actualisation. Le quintuplé $\langle X, A, p, r, \gamma \rangle$ est appelé **processus de décision Markovien (PDM)** (Puterman, 1994; Bertsekas & Tsitsiklis, 1996), et le problème que nous avons décrit est communément appelé **contrôle optimal**.

La **fonction valeur optimale** à partir de l'état i est définie par

$$v_*(i) := \max_\pi v_\pi(i).$$

Pour toute politique π, nous notons P_π la matrice stochastique de dimension $n \times n$ dont les éléments sont $p_{ij}(\pi(i))$, et r_π le vecteur (colonne) de taille n dont les éléments sont $\sum_j p_{ij}(\pi(i)) r(i, \pi(i), j)$. Les fonctions valeur v_π et v_* peuvent être vues comme des vecteurs de taille n. Il est connu que la fonction valeur v_π d'une politique π est la solution de l'équation de Bellman linéaire suivante :

$$v_\pi = r_\pi + \gamma P_\pi v_\pi,$$

c'est-à-dire que v_π est le point fixe de l'opérateur affine $T_\pi : v \mapsto r_\pi + \gamma P_\pi v$. Il est également connu que la fonction valeur optimale v_* satisfait l'équation de Bellman non linéaire suivante

$$v_* = \max_\pi (r_\pi + \gamma P_\pi v_*) = \max_\pi T_\pi v_*$$

où l'opérateur max s'applique à chaque composante. En d'autres termes, v_* est le point fixe de l'opérateur non linéaire $T : v \mapsto \max_\pi T_\pi v$. Pour toute fonction $v : X \to \mathbb{R}$, on dit d'une politique π qu'elle est **gourmande par rapport à** v si elle satisfait

$$\pi \in \arg\max_{\pi'} T_{\pi'} v$$

ou de manière équivalente, $T_\pi v = Tv$. On note $\mathcal{G}(v)$—de manière légèrement abusive—n'importe quelle politique qui est gourmande par rapport à v. Les notions de *fonction valeur optimale* et de *politique gourmande* sont fondamentales pour le contrôle optimal : en effet, n'importe quelle politique π_* qui est gourmande par rapport à la fonction valeur optimale v_* est une politique optimale et sa valeur v_{π_*} est égale à v_*.

Soit π une politique. On appelle **avantage par rapport à la politique** π la quantité :

$$a_\pi = \max_{\pi'} T_{\pi'} v_\pi - v_\pi = Tv_\pi - v_\pi.$$

On appelle de plus **ensemble des états modifiables** (en anglais, *switchable*) l'ensemble

$$S_\pi = \{i,\ a_\pi(i) > 0\}.$$

Supposons maintenant que π n'est pas optimale (cela implique que S_π est un ensemble non vide). Pour n'importe quel sous-ensemble Y de S_π, notons switch(π, Y) une politique qui satisfait :

$$\forall i,\ \text{switch}(\pi, Y)(i) = \begin{cases} \mathcal{G}(v_\pi)(i) & \text{si } i \in Y \\ \pi(i) & \text{si } i \notin Y. \end{cases}$$

Le résultat suivant est standard (voir par exemple (Puterman, 1994)).

Lemme 1
Soit π une politique non optimale. Si $\pi' = \text{switch}(\pi, Y)$ pour un sous-ensemble non vide Y de S_π, alors $v'_\pi \geq v_\pi$ et il existe au moins un état i tel que $v'_\pi(i) > v_\pi(i)$.

2. Se restreindre aux politiques déterministes et stationnaires n'est pas une limitation, dans le sens où, pour le critère d'optimalité que nous allons considérer, on peut montrer qu'il existe au moins une politique déterministe et stationnaire qui est optimale.

Ce lemme constitue le fondement de l'algorithme itératif du nom d'*itérations sur les politiques* (IP), qui génère une séquence de politiques (π_k) comme suit.

$$\pi_{k+1} \leftarrow \text{switch}(\pi_k, Y_k) \text{ pour un } Y_k \text{ tel que } \emptyset \subsetneq Y_k \subseteq S_{\pi_k}.$$

Selon la façon dont on choisit l'ensemble Y_k, on obtient différentes variations de IP. Dans cet article, nous considérons les deux variantes suivantes :

- Quand pour tout k, $Y_k = S_{\pi_k}$, c'est-à-dire qu'on change les actions dans tous les états où l'avantage est strictement positif, l'algorithme est connu sous le nom d'IP de Howard (IP-Howard dans la suite) ; on peut notamment remarquer dans ce cas qu'à chaque étape, la politique est gourmande par rapport à la valeur de la politique précédente, soit formellement $\pi_{k+1} \in \mathcal{G}(v_{\pi_k})$.
- Quand pour tout k, Y_k est un singleton qui contient l'état $i_k \in \arg\max_i a_{\pi_k}(i)$, c'est-à-dire qu'on change une seule action dans l'état qui a un avantage maximal par rapport à π_k, on appellera cet algorithme IP-Simplexe [3].

Comme ces algorithmes génèrent une séquence de politiques dont les valeurs forment une suite strictement croissante (par le Lemme 1), toute variation d'IP converge vers la politique et la valeur optimales en un nombre d'itérations qui est plus petit que m^n, le nombre total de politiques du problème. En pratique, IP converge en très peu d'itérations. Sur des PDM aléatoires, la convergence a lieu en un temps qui est souvent sous-linéaire en n. Le but de cet article est d'exhiber des majorants du nombre d'itérations requis par IP-Howard et IP-Simplexe qui sont beaucoup plus fins que m^n.

2 Résultats

Dans cette section, nous décrivons quelques résultats de la littérature—voir (Ye, 2011) pour un état de l'art plus étoffé—sur le nombre d'itérations requis par IP-Howard et IP-Simplexe, ainsi qu'un certain nombre d'améliorations et d'extensions originales. Nous renverrons le lecteur à (Scherrer, 2013) pour les démonstrations des différents résultats énoncés.

Une observation importante concernant les deux algorithmes, qui sera centrale dans les résultats que nous allons décrire, est que la séquence qu'ils génèrent satisfait une certaine propriété de contraction [4]. Pour tout vecteur u de $\mathbb{R}^n$, soit $\|u\|_\infty = max_{1\leq i\leq n}|u(i)|$ la norme infinie de u. Soit $\mathbb{1}$ le vecteur dont toutes les composantes sont égales à 1.

Lemme 2 (Preuve dans la Section 5 de Scherrer (2013))
La séquence $(\|v_* - v_{\pi_k}\|_\infty)_{k\geq 0}$ générée par IP-Howard est contractante de coefficient γ.

Lemme 3 (Preuve dans la Section 6 de Scherrer (2013))
La séquence $(\mathbb{1}^T(v_* - v_{\pi_k}))_{k\geq 0}$ générée par IP-Simplexe est contractante de coefficient $1 - \frac{1-\gamma}{n}$.

Si cette observation est bien connue pour IP-Howard, elle n'a à notre connaissance jamais été mentionnée explicitement dans la littérature pour IP-Simplexe. Ces propriétés de contractions ont pour immédiate conséquence le résultat suivant [5].

Corollaire 1
Soit $V_{\max} = \frac{\max_\pi \|r_\pi\|_\infty}{1-\gamma}$ un majorant de $\|v_\pi\|_\infty$ pour toute politique π. Afin d'obtenir une politique ϵ-optimale, c'est-à-dire une politique π_k satisfaisant $\|v_* - v_{\pi_k}\|_\infty \leq \epsilon$, IP-Howard a besoin d'au plus $\left\lceil \frac{\log \frac{V_{\max}}{\epsilon}}{1-\gamma} \right\rceil$ itérations, tandis que IP-Simplexe a besoin d'au plus $\left\lceil \frac{n \log \frac{nV_{\max}}{\epsilon}}{1-\gamma} \right\rceil$ itérations.

3. IP-Simplexe est équivalent à l'algorithme du simplexe (utilisant la règle du pivot le plus grand) appliqué à une formulation de type programmation linéaire du problème (Ye, 2011)

4. Une séquence $(x_k)_{k\geq 0}$ de réels positifs est dite contractant de coefficient $\alpha < 1$ si et seulement si pour tout $k \geq 0$, $x_{k+1} \leq \alpha x_k$.

5. Pour IP-Howard, nous avons : $\|v_* - v_{\pi_k}\|_\infty \leq \gamma^k \|v_* - v_{\pi_0}\|_\infty \leq \gamma^k V_{\max}$. Ainsi, une condition suffisante pour avoir $\|v_* - v_{\pi_k}\|_\infty < \epsilon$ est $\gamma^k V_{\max} < \epsilon$, ce qui est vrai dès lors que $k \geq \frac{\log \frac{V_{\max}}{\epsilon}}{1-\gamma} > \frac{\log \frac{V_{\max}}{\epsilon}}{\log \frac{1}{\gamma}}$. Pour IP-Simplexe, nous avons $\|v_* - v_{\pi_k}\|_\infty \leq \mathbb{1}^T(v_* - v_{\pi_k}) \leq \gamma^k \mathbb{1}^T(v_* - v_{\pi_0}) \leq \gamma^k n V_{\max}$, et on conclue de manière similaire à IP-Howard.

Comme ces majorants dépendent d'un terme de précision ϵ, cela implique que IP-Howard et IP-Simplexe sont des algorithmes *faiblement polynomiaux* pour un facteur d'actualisation γ fixé. Un résultat particulièrement important a récemment été obtenu par Ye (2011), qui a fourni des majorants qui ne dépendent plus de la précision ϵ, ce qui implique que IP-Howard et IP-Simplexe sont *fortement polynomiaux* pour un facteur d'actualisation γ fixé.

Theorème 1 (Ye (2011))
IP-Simplexe et IP-Howard terminent après au plus $n(m-1)\left\lceil \frac{n}{1-\gamma} \log\left(\frac{n^2}{1-\gamma}\right)\right\rceil$ *itérations.*

La démonstration se base sur le fait que ces algorithmes sont des instances particulières de l'algorithme du simplexe appliqué à une formulation de type programme linéaire du problème. En utilisant des arguments plus directs, Hansen *et al.* (2013) ont récemment amélioré le majorant d'un facteur $O(n)$ pour IP-Howard.

Theorème 2 (Hansen *et al.* (2013))
IP-Howard termine après au plus $(nm+1)\left\lceil \frac{1}{1-\gamma} \log\left(\frac{n}{1-\gamma}\right)\right\rceil$ *itérations.*

Nos deux premiers résultats, qui sont des conséquences des propriétés de contractions décrites plus haut (Lemmes 2 et 3) sont énoncés dans les théorèmes ci-dessous.

Theorème 3 (Preuve dans la Section 7 de Scherrer (2013))
IP-Howard termine après au plus $n(m-1)\left\lceil \frac{1}{1-\gamma} \log\left(\frac{1}{1-\gamma}\right)\right\rceil$ *itérations.*

Theorème 4 (Preuve dans la Section 8 de Scherrer (2013))
IP-Simplexe termine après au plus $n(m-1)\left\lceil \frac{n}{1-\gamma} \log\left(\frac{n}{1-\gamma}\right)\right\rceil$ *itérations.*

Notre résultat pour IP-Howard est un facteur $O(\log n)$ plus fin que celui de Hansen *et al.* (2013), qui est à notre connaissance le meilleur résultat publié dans la littérature. Notre résultat pour IP-Simplexe est très légèrement meilleur (d'un facteur 2) que celui de Ye (2011), et utilise des arguments plus directs. En utilisant une preuve un peu plus complexe, on peut améliorer le résultat pour IP-Simplexe d'un facteur $O(\log n)$:

Theorème 5 (Preuve dans la Section 9 de Scherrer (2013))
IP-Simplexe termine après au plus $n^2(m-1)\left(1+\frac{2}{1-\gamma}\log\left(\frac{1}{1-\gamma}\right)\right)$ *itérations.*

Par rapport à IP-Howard, le nombre d'itérations requis par IP-Simplexe est ainsi un facteur $O(n)$ plus grand. Cependant, chaque itération de IP-Simplexe (qui change seulement une action) a en général une complexité moindre qu'une itération de IP-Howard : en effet, la mise à jour peut-être réalisée en temps $O(n^2)$ à l'aide de la formule de Sherman-Morrisson, alors qu'une itération de IP-Howard, qui nécessite de calculer la valeur d'une politique qui peut être arbitrairement différente de la politique précédente, peut demander un temps en $O(n^3)$. Globalement, on pourra retenir que les deux algorithmes ont une complexité similaire.

Il est facile de voir que la dépendance linéaire en n du majorant pour IP-Howard est optimale. Nous conjecturons que la dépendance linéaire en m pour les deux bornes est également optimale. Il est peut-être possible d'améliorer la dépendance vis-à-vis du terme $\frac{1}{1-\gamma}$, mais la supprimer est impossible pour IP-Howard et peu vraisemblable pour IP-Simplexe. Fearnley (2010) a décrit un PDM pour lequel IP-Howard met un temps exponentiel en n pour $\gamma = 1$ et Hollanders *et al.* (2012) ont ensuite expliqué que cette complexité était la même lorsque γ est dans le voisinage de 1. Bien que des résultats similaires pour IP-Simplexe ne sont à notre connaissance pas documentés dans la littérature, Melekopoglou & Condon (1994) ont considéré quatre variations d'IP qui changent seulement une action par itération, et exhibent des PDM sur lesquels ces algorithmes terminent en un temps exponentiel en n lorsque $\gamma = 1$.

Dans la suite de cet article, nous décrivons des majorants qui ne dépendent pas du facteur d'actualisation γ, mais qui seront obtenus sous des hypothèses structurelles du PDM. Sur ce sujet, Post & Ye (2012) ont récemment dérivé une borne pour les PDM déterministes.

Theorème 6 (Post & Ye (2012))
Si le PDM est déterministe, alors IP-Simplexe termine après au plus $O(n^5 m^2 \log^2 n)$ *itérations.*

Étant donnée une politique π d'un PDM déterministe, les états appartiennent soit à un cycle, soit à un chemin induit par π. Le cœur de l'analyse repose sur les lemmes que nous énonçons ci-après, qui montrent qu'au cours des itérations, des cycles sont créés régulièrement et qu'à chaque fois qu'un nouveau cycle apparaît, un progrès significatif vers la solution est effectué ; en conséquence, un progrès significatif a lieu régulièrement.

Lemme 4
Supposons que le PDM est déterministe. Après au plus $O(n^2 m \log n)$ itérations, soit IP-Simplexe termine soit un nouveau cycle apparaît.

Lemme 5
Supposons que le PDM est déterministe. Lorsque IP-Simplexe passe de π à π' où π' implique un nouveau cycle, on a

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{n}\right) \mathbb{1}^T(v_{\pi_*} - v_{\pi}).$$

Ces observations suffisent à prouver [6] que IP-Simplexe termine après au plus $O(n^4 m^2 \log \frac{n}{1-\gamma}) = \tilde{O}(n^4 m^2)$ itérations. Éliminer complètement la dépendance en le facteur d'actualisation γ—le terme en $O(\log \frac{1}{1-\gamma})$—requiert un travail minutieux supplémentaire (voir Post & Ye (2012)), et se traduit par une dépendance supplémentaire en $O(n \log(n))$.

D'un point de vue plus technique, la preuve de Post & Ye (2012) se fonde singulièrement sur des propriétés du vecteur $x_\pi = (I - \gamma P_\pi^T)^{-1}\mathbb{1}$, qui représente une mesure actualisée de la présence en chacun des états lorsqu'on considère une trajectoire induite par la politique π à partir d'un état initial aléatoire uniforme :

$$\forall i \in X, \;\; x_\pi(i) = n \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(i_t = i \mid i_0 \sim U, \; a_t = \pi(i_t)),$$

où nous avons noté U la distribution uniforme sur X. Pour toute politique π et tout état i, on a trivialement $x_\pi(i) \in \left[1, \frac{n}{1-\gamma}\right]$. La démonstration exploite de plus le fait que $x_\pi(i)$ appartient à l'intervalle $[1, n]$ lorsque i est sur un chemin de π, tandis que $x_\pi(i)$ appartient à l'intervalle $\left[\frac{1}{1-\gamma}, \frac{n}{1-\gamma}\right]$ lorsque i est sur un cycle de π. Il est possible, c'est ce que nous montrons ci-après, de généraliser l'approche de (Post & Ye, 2012) aux PDM stochastiques. Étant donnée une politique π d'un PDM stochastique, les états sont soit *récurrents*, soit *transients* (ces deux catégories généralisant respectivement celles de cycles et de chemins pour les PDM déterministes). Ceci nous amène à formuler l'hypothèse suivante.

Hypothèse 1
Soit $\tau_t \geq 1$ et $\tau_r \geq 1$ les plus petits coefficients tels que pour toute politique π, pour tout état i transient pour π,

$$(1 \leq) x_\pi(i) \leq \tau_t,$$

et pour tout état i récurrent pour π,

$$\frac{n}{(1-\gamma)\tau_r} \leq x_\pi(i) \left(\leq \frac{n}{1-\gamma}\right).$$

La constante τ_t (respectivement τ_r) peut être vue comme une mesure du temps nécessaire pour quitter les états transients (respectivement pour revisiter les états dans les classes récurrentes). En particulier, quand γ tend vers 1, on peut voir que τ_t est un majorant de $\mathcal{L}$, le temps aléatoire nécessaire pour quitter les états transients, car pour toute politique π,

$$\begin{aligned}\lim_{\gamma \to 1} \tau_r \geq \frac{1}{n} \lim_{\gamma \to 1} \sum_{i \text{ transient pour } \pi} x_\pi(i) &= \sum_{t=0}^{\infty} \mathbb{P}(i_t \text{ transient pour } \pi \mid i_0 \sim U, \; a_t = \pi(i_t)) \\ &= \mathbb{E}\left[\, \mathcal{L} \mid i_0 \sim U, \; a_t = \pi(i_t)\right].\end{aligned}$$

6. Ceci peut être fait par des arguments similaires à la démonstration du Théorème 5 (voir la Section 6 de Scherrer (2013)).

De manière similaire, lorsque γ est dans le voisinage de 1, $\frac{1}{\tau_r}$ est la fréquence asymptotique minimale [7] dans les états récurrents étant donné que la chaîne est initialisée uniformément sur X, car pour toute politique π et tout état récurrent i,

$$\lim_{\gamma \to 1} \frac{1-\gamma}{n} x_\pi(i) = \lim_{\gamma \to 1} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(i_t = i \mid i_0 \sim U,\ a_t = \pi(i_t))$$
$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{P}(i_t = i \mid i_0 \sim U,\ a_t = \pi(i_t)).$$

Une fois l'hypothèse 1 formulée, nous pouvons généraliser les Lemmes 4-5 comme suit.

Lemme 6
Supposons que le PDM satisfait l'hypothèse 1. Après au plus $n^2 m \lceil \tau_t \log(n(\tau_t + 1)) \rceil$ itérations, soit IP-Simplexe termine soit une nouvelle classe récurrente apparaît.

Lemme 7
Supposons que le PDM satisfait l'hypothèse 1. Lorsque IP-Simplexe passe de π à π' où π' implique une nouvelle classe récurrente, on a

$$\mathbb{1}^T (v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{\tau_r}\right) \mathbb{1}^T (v_{\pi_*} - v_\pi).$$

De ces observations générales découle le résultat original suivant.

Theorème 7 (Preuve dans la Section 10 de Scherrer (2013))
Si le PDM satisfait l'hypothèse 1, alors IP-Simplexe termine après au plus

$$n^2(m-1)\left(\lceil \tau_r \log(n\tau_r) \rceil + \lceil \tau_r \log(n\tau_t) \rceil\right)\left[(m-1)\lceil n\tau_t \log(n\tau_t) \rceil + \lceil n\tau_t \log(n^2\tau_t) \rceil\right] = \tilde{O}\left(n^3 m^2 \tau_t \tau_r\right)$$

itérations.

Remarque 1
Ce nouveau résultat est une généralisation stricte de celui énoncé pour les PDM déterministes. En effet, dans le cas déterministe, on a $\tau_t = \tau_r = n$ et il est facile de voir que les Lemmes 6, 7 et le Théorème 7 impliquent les Lemmes 4, 5 et le Théorème 6.

Une conséquence immédiate de ce résultat est que IP-Simplexe est *fortement polynomial* pour des ensembles de PDM qui sont significativement plus grand que les PDM déterministes considérés au Théorème 6.

Corollaire 2
Pour toute famille de PDM indexée par n et m telle que τ_t et τ_r sont polynomiaux en n et m, IP-Simplexe termine en un temps polynomial en n et m.

On pourrait se demander si un résultat analogue peut être dérivé pour IP-Howard. Malheureusement, et comme Post & Ye (2012) le mentionnent brièvement, la technique de preuve développée pour IP-Simplexe ne semble pas s'adapter aisément à IP-Howard, parce que les changements de plusieurs actions lors d'une itération peuvent interférer les uns avec les autres de sorte à rendre l'amélioration de la politique très petite. Nous pouvons détailler un peu plus précisément ce qui, dans l'approche décrite plus haut, pose problème. D'un côté, il est possible d'écrire des variations des Lemmes 4 et 6 pour IP-Howard.

Lemme 8 (Preuve dans la Section 11 de Scherrer (2013))
Supposons que le PDM est déterministe. Après au plus n itérations, soit IP-Howard termine soit un nouveau cycle apparaît.

7. Si le PDM est apériodique et irréductible, de sorte qu'il admet une politique stationnaire ν_π pour toute politique π, on peut voir que

$$\frac{1}{\tau_r} = \min_{\pi,\ i \text{ récurrent pour } \pi} \nu_\pi(i).$$

Lemme 9 (Preuve dans la Section 11 de Scherrer (2013))
Supposons que le PDM satisfait l'hypothèse 1. Après au plus $nm\lceil \tau_t \log n\tau_t \rceil$ *itérations, soit IP-Howard termine soit une nouvelle classe récurrente apparaît.*

Cependant, de l'autre côté, nous n'avons pas réussi à adapter les Lemmes 5 ou 7. De fait, il semble peu vraisemblable qu'un résultat similaire à celui énoncé dans le Lemme 5 soit vrai pour IP-Howard. Dans un exemple déterministe récemment décrit par Hansen & Zwick (2010) pour montrer que IP-Howard pouvait requérir au moins $O(n^2)$ itérations, des cycles apparaissent à chaque itération mais la séquence des fonctions valeur satisfait [8] pour toute itération $k < \frac{n^2}{4} + \frac{n}{4}$ et état i,

$$v_*(i) - v_{\pi_{k+1}}(i) \geq \left[1 - \left(\frac{2}{n}\right)^k\right](v_*(i) - v_{\pi_k}(i)).$$

Contrairement au Lemme 5, on voit ici que lorsque k grandit, la contraction est (exponentiellement rapidement) de moins en moins forte. Par rapport à IP-Simplexe, cela suggère notamment que IP-Howard pourrait (dans le pire cas) souffrir de subtiles pathologies. Plus largement, déterminer le nombre d'itérations nécessaire pour IP-Howard est un problème qui résiste aux efforts des chercheurs depuis maintenant presque 30 ans : il a été originellement énoncé par Schmitz (1985). Dans le cas le plus simple (déterministe), le problème est aujourd'hui encore ouvert : le meilleur minorant connu est la borne en $O(n^2)$ de Hansen & Zwick (2010) que nous venons de mentionner, alors que le meilleur majorant est $O(\frac{m^n}{n})$—valable pour les PDM en général—prouvé par Mansour & Singh (1999).

Du côté positif, nous avons été à même d'adapter l'analyse décrite plus haut sous une hypothèse structurelle supplémentaire.

Hypothèse 2
L'espace d'états X *peut être partitionné en deux ensembles* $\mathcal{T}$ *et* $\mathcal{R}$ *tels que pour toute politique* π*, les états de* $\mathcal{T}$ *sont transients et ceux de* $\mathcal{R}$ *sont récurrents.*

En effet, sous cette hypothèse, il est possible de prouver pour IP-Howard une variation du Lemme 7 introduit pour IP-Simplexe.

Lemme 10
Supposons que le PDM satisfait les Hypothèses 1 et 2. Lorsque IP-Howard passe de π *à* π' *où* π' *implique une nouvelle classe récurrente, on a*

$$\mathbb{1}^T(v_{\pi_*} - v_{\pi'}) \leq \left(1 - \frac{1}{\tau_r}\right)\mathbb{1}^T(v_{\pi_*} - v_\pi).$$

Et on peut en déduire le résultat suivant original (qui s'applique d'ailleurs également à IP-Simplexe).

Theorème 8 (Preuve dans la Section 12 de Scherrer (2013))
Si le PDM satisfait les Hypothèses 1 et 2, alors IP-Howard termine après au plus $n(m-1)\left(\lceil \tau_t \log n\tau_t \rceil + \lceil \tau_r \log n\tau_r \rceil\right)$ *itérations tandis que IP-Simplexe termine après au plus* $n(m-1)\left(\lceil n\tau_t \log n\tau_t \rceil + \lceil \tau_r \log n\tau_r \rceil\right)$ *itérations.*

Il est important de noter que l'hypothèse 2 est plutôt restrictive. Elle implique que les deux algorithmes convergent sur les états récurrents indépendemment de ce qui se passe sur les états transients, ce qui permet de réduire l'analyse à deux étapes : 1) l'analyse de la convergence sur les états récurrents ; 2) l'analyse de la convergence sur les états transients (sachant que la convergence a eu lieu sur les états récurrents). L'étude de la première phase (convergence sur les états récurrents) est grandement facilitée par le fait que, dans ce cas, une classe récurrente apparaît à chaque itération (ceci est à contraster avec les Lemmes 4, 6, 8 et 9 dont la fonction est de montrer que des cycles ou des classes récurrentes apparaissent en un temps raisonnable). De plus, l'analyse de la deuxième phase (convergence sur les états transients) est similaire à celle du cas avec facteur d'actualisation γ (Théorèmes 3 et 5). En d'autres termes, si ce dernier résultat contribue à nous éclairer sur l'efficacité pratique usuelle de IP-Howard et IP-Simplexe, une analyse plus générale de IP-Howard est encore à faire, et constitue la principale perspective de cette étude.

8. Ce PDM a un nombre pair d'états $n = 2p$. Dans (Hansen & Zwick, 2010), le but est de minimiser l'espérance de la somme actualisée des coûts. La fonction valeur optimale satisfait $v_*(i) = -p^N$ pour tout i, avec $N = p^2 + p$. Les politiques générées par IP-Howard ont des fonctions valeur satisfaisant $v_{\pi_k}(i) \in [p^{N-k-1}, p^{N-k}]$. On en déduit que pour toute itération k et tout état i, $\frac{v_*(i) - v_{\pi_{k+1}}(i)}{v_*(i) - v_{\pi_k}(i)} \geq \frac{1+p^{-k-2}}{1+p^{-k}} = 1 - \frac{p^{-k} - p^{-k-2}}{1+p^{-k}} \geq 1 - p^{-k}(1-p^{-2}) \geq 1 - p^{-k}$.

Références

BERTSEKAS D. & TSITSIKLIS J. (1996). *Neurodynamic Programming*. Athena Scientific.

FEARNLEY J. (2010). Exponential lower bounds for policy iteration. In *Proceedings of the 37th international colloquium conference on Automata, languages and programming : Part II*, ICALP'10, p. 551–562, Berlin, Heidelberg : Springer-Verlag.

HANSEN T., MILTERSEN P. & ZWICK U. (2013). Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J. ACM*, **60**(1), 1 :1–1 :16.

HANSEN T. & ZWICK U. (2010). Lower bounds for howard's algorithm for finding minimum mean-cost cycles. In *ISAAC (1)*, p. 415–426.

HOLLANDERS R., DELVENNE J. & JUNGERS R. (2012). The complexity of policy iteration is exponential for discounted markov decision processes. In *51st IEEE conference on Decision and control (CDC'12)*.

MANSOUR Y. & SINGH S. (1999). On the complexity of policy iteration. In *UAI*, p. 401–408.

MELEKOPOGLOU M. & CONDON A. (1994). On the complexity of the policy improvement algorithm for markov decision processes. *INFORMS Journal on Computing*, **6**(2), 188–192.

POST I. & YE Y. (2012). *The simplex method is strongly polynomial for deterministic Markov decision processes.* Rapport interne, arXiv :1208.5083v2.

PUTERMAN M. (1994). *Markov Decision Processes*. Wiley, New York.

SCHERRER B. (2013). *Improved and Generalized Upper Bounds on the Complexity of Policy Iteration*. Rapport interne, hal-00829532.

SCHMITZ N. (1985). How good is howard's policy improvement algorithm ? *Zeitschrift für Operations Research*, **29**(7), 315–316.

YE Y. (2011). The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Math. Oper. Res.*, **36**(4), 593–603.