

Calcul d'une politique déterministe dans un MDP avec récompenses imprécises

Pegah Alizadeh **, Aomar Osmani *, Emiliano Traversi*

*Léonard de Vinci Pôle Universitaire, Centre Recherche

** LIPN-UMR CNRS 7030, PRES Sorbonne Paris Cité

Résumé. Pour beaucoup d'applications réelles nécessitant une prise de décision séquentielle dans un cadre incertain, on utilise un processus de décision Markovien avec récompenses inconnues (IRMDP) en calculant naturellement des politiques stochastiques. Une politique stochastique n'est pas facilement interprétable pour l'utilisateur final. Celui-ci a souvent besoin d'une politique déterministe et compréhensible. Pour mieux motiver l'utilisation d'une procédure exacte pour trouver une politique déterministe, nous montrons quelques cas où l'idée intuitive d'utiliser une politique déterministe obtenue après une «détermination» (arrondi) de la politique stochastique optimale donne une politique déterministe différente de la politique optimale.

1 Introduction

Les processus de décision de Markov (MDP) se sont avérés être des modèles efficaces pour représenter et résoudre des problèmes de décision séquentiels avec incertitude. Les applications dans divers domaines (robotique, navigation, composition de services, etc.) nécessitant une prise de décision dans un environnement dynamique utilisent naturellement des modèles MDP. Dans le cas de la conduite autonome, par exemple, à chaque étape, les effets probabilistes de l'exécution d'une action (déplacement, rotation) conduit aux étapes suivantes, chacune avec une récompense ou pénalité différente. Ceci motive l'utilisation des MDP qui rendent compte de ces incertitudes dans les paramètres du modèle. Dans ce cas, l'environnement (ici le trafic routier) sera modélisé par un ensemble d'actions probabilistes (par exemple, les probabilités d'accidents et leur coûts) et un ensemble d'états. L'objectif peut être défini comme un problème de maximisation de la somme des récompenses attendue.

En dépit de connaître l'objectif final, spécifier des récompenses ou des punitions pour choisir des actions à partir des états n'est jamais évident. De plus, comme l'a montré Mannor et al. (2007), les stratégies trouvées via un processus d'optimisation sous MDPs avec des paramètres numériques, peuvent parfois être bien pire que la politique anticipée. Parmi les raisons qui peuvent expliquer cette situation, on peut citer : (1) l'insuffisance des données pour estimer les récompenses, (2) la complexité de la construction des modèles¹ et (3) la présence d'élicitations contradictoires entre les utilisateurs. Dans le cas du véhicule autonome, si le modèle est

1. Dans le cas du véhicule autonome, par exemple, la définition des récompenses exactes pour toutes les actions prend du temps et est compliquée à faire. De plus, ces récompenses varient au cours du processus de conduite.

Calcul d'une politique déterministe dans un IRMDP

conçu pour différents conducteurs ayant différentes préférences, le MDP ne sera pas précis. Dans le cas des MDPs à récompenses inconnues (IRMDP), le système disposera de toutes les informations concernant la dynamique (route et trafic) ainsi que l'objectif final à atteindre (destination sans accident), mais ne nécessite pas la connaissance des préférences des utilisateurs à l'intérieur du système. Plusieurs approches ont été proposées dans la littérature pour trouver la meilleure *politique* dans un environnement avec des récompenses imprécises. Ces travaux sont axés essentiellement sur le critère de *regret minmax*. L'idée de base est de trouver la politique avec une perte minimale en comparaison avec les autres politiques possibles. Minimiser le regret maximal est plus optimiste que minimiser le pire cas ; il est largement utilisé dans la littérature (Regan et Boutilier (2009); Xu et Mannor (2009)).

Les méthodes de résolution exactes et approchées d'un MDP ont des politiques réalisables et sont de nature stochastique. Une politique est dite stochastique si, pour un état donné, l'action à prendre est choisie avec une probabilité associée à chaque état destination. L'utilisation de politiques stochastiques présente deux avantages principaux par rapport à une approche déterministe. D'un point de vue algorithmique (comme le montre le travail présenté dans ce papier), trouver la politique stochastique optimale est généralement plus rapide que de trouver la politique déterministe. En outre, le choix d'une politique stochastique implique l'exploration d'un espace de recherche plus grand par rapport à celui d'une politique déterministe, permettant ainsi d'avoir la possibilité d'atteindre une meilleure politique optimale. Malgré ces avantages évidents des politiques stochastiques, leur utilisation est souvent déconseillée voire éthiquement problématique à mettre en œuvre. Dans le cas des véhicules autonomes, par exemple, on peut se retrouver devant le dilemme du tramway posé par Philippe Foot ; cette expérience de pensée pose le problème de décision de l'action à prendre (le conducteur choisit une voix) qui favoriserait (laisserait en vie) un groupe au détriment d'un individu (en le tuant). La politique optimale devrait être déterministe sans obliger l'utilisateur à devoir prendre une décision avec une probabilité donnée p de rester sur la même voix ou de prendre une autre avec une probabilité de $1 - p$. Plus généralement, une politique déterministe est souvent nécessaire de par la nature discrète/combinatoire du problème étudié et du fait que l'algorithme ne peut être exécuté qu'une seule fois rendant ainsi l'aspect stochastique moins pertinent. De plus, une politique déterministe est souvent plus simple à comprendre du point de vue de l'utilisateur humain et donc plus susceptible d'être utilisée dans la pratique.

Dans cet article, nous présentons une étude originale sur la recherche de la politique déterministe qui minimise le regret maximal dans un IRMDP. Typiquement, dans les MDPs, les fonctions de récompenses sont estimées à partir d'observations ou de sources externes. Les travaux de Mannor et al. (2007), qui ont montré que la politique trouvée via l'optimisation d'un MDP avec des paramètres numériques n'est pas garantie, ont motivé une modélisation mathématique des IRMDPs incluant des modélisations avec récompenses numériques (Regan et Boutilier (2009); Xu et Mannor (2009)). On se focalisera, dans notre travail, sur les approches prenant le point de vue de la théorie de la décision et on considérera un ensemble de MDP avec diverses fonctions de récompenses imprécises qu'on nomme *IRMDP*.

Une approche courante pour le calcul d'une solution robuste est la méthode *minmax* qui est une politique qui maximise la valeur par rapport au scénario le plus défavorable (Nilim et Ghaoui (2005)). La robustesse du minmax peut être vue comme un jeu à deux joueurs ; le premier choisit la politique qui maximise la récompense tandis que le deuxième propose une instantiation qui minimise la valeur attendue. Certains travaux récents (Mannor et al. (2012); Wie-

semann et al. (2013)) traitent de problème des incertitudes interdépendantes dans les MDPs. Dans ce papier, nous ne traiterons que les cas des fonctions de récompenses indépendantes les unes des autres. Les politiques *maxmin* sont naturellement conservatrices, c'est pour cette raison que les approches de *regret minmax* (Regan et Boutilier (2009); Xu et Mannor (2009)) ont été introduites. L'objectif de l'approche de *regret minmax* est de trouver la politique qui a le moins de regret possible sur toutes les instances des récompenses. Plusieurs méthodes se sont concentrées sur le calcul de la politique stochastique optimale pour les IRMDPs (Ahmed et al. (2017); Regan et Boutilier (2009); Xu et Mannor (2009)). A notre connaissance, il n'y a aucun travail qui gère le calcul de politiques déterministes des MDPs sous incertitudes comme nous le présentons dans ce papier.

Nous montrons également, dans ce travail, que l'utilisation d'une technique de détermination intuitive pour obtenir une politique déterministe réalisable basée sur la solution stochastique optimale peut mener à une politique significativement différente de l'optimale. Ce qui, de ce fait, exclut l'utilisation d'approches stochastiques de ce type pour le calcul d'actions déterministes. Le résultat de notre travail est aussi validé expérimentalement sur des MDP aléatoires et en diamants. Nous donnerons, dans cet article, les résultats des analyses de performances de nos algorithmes. Le reste de l'article est organisé comme suit : la Section 2 définit le MDP, la formulation du problème sous forme d'un critère de *regret minmax* avec récompenses imprécises et pose le problème comme un problème de jeu à deux adversaires. La Section 4 présente les principaux résultats théoriques portant sur la politique déterministe optimale. Les résultats expérimentaux sur des instances de la littératures sont présentés à la Section 5 et la Section 6 donne quelques conclusions et perspectives.

2 Préliminaires

Un MDP *Markov Decision Process* (Puterman (1994)) est défini par un tuple $M(S, A, P, r, \gamma, \beta)$, où : S est un ensemble fini d'états ; A un ensemble fini d'actions, $P : S \times A \times S \rightarrow [0, 1]$ est une fonction de transition tel que $\mathbb{P}(s'|s, a)$ encode la probabilité de passer dans l'état s' en étant dans l'état s en choisissant l'action a ; $r : S \times A \rightarrow \mathbb{R}$ est une *fonction de gain* (ou de pénalité) obtenue en choisissant l'action a en étant dans l'état s ; $\gamma \in [0, 1[$ est le facteur d'actualisation et $\beta : S \rightarrow [0, 1]$ est une *distribution initiale des états* indiquant par $\beta(s)$ la probabilité de commencer par l'état s . Une *politique déterministe stationnaire* est une fonction $\pi : S \rightarrow A$, qui préconise de prendre l'action $\pi(s)$ quand on est dans l'état s . Une *politique stochastique stationnaire* est une fonction $\tilde{\pi} : S \times A \rightarrow [0, 1]$ qui indique avec une probabilité $\tilde{\pi}(s, a)$, que l'action a est choisie dans l'état s selon la politique $\tilde{\pi}$.

Une politique $\tilde{\pi}$ induit une *fonction de la fréquence de visite des états* $f^{\tilde{\pi}}$ où $f^{\tilde{\pi}}(s, a)$ est la probabilité conjointe totale d'être dans l'état s et de choisir l'action a (voir la Section 6.9 dans Puterman (1994)) où la somme est prise sur des trajectoires définies par $S_0 \sim \beta, A_t \sim \tilde{\pi}(S_t)$ et $S_{t+1} \sim P(\cdot | S_t, A_t)$. La politique est calculable à partir de $f^{\tilde{\pi}}$ à travers $\tilde{\pi}(s, a)$:

$$f^{\tilde{\pi}}(s, a) = \sum_{s' \in S} \beta(s') \sum_{t=0}^{\infty} \gamma^{t-1} (S_t = s', A_t = a | S_1 = s), \quad \tilde{\pi}(s, a) = \frac{f^{\tilde{\pi}}(s, a)}{\sum_{a'} f^{\tilde{\pi}}(s, a')}.$$

Pour une politique déterministe π , nous avons $f^{\pi}(s, a) = 0, \forall a \neq \pi(s)$. Les politiques sont évaluées avec la *fonction valeur* $V : S \rightarrow \mathbb{R}$:

Calcul d'une politique déterministe dans un IRMDP

$$V^\pi(s) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))\right).$$

Une autre manière de définir la qualité de la politique est la *fonction Q-value* $Q : S \times A \rightarrow \mathbb{R}$:

$$Q^\pi(s, a) = r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V^\pi(s'). \quad (1)$$

Pour un état initial β , la valeur de la politique optimale est $\beta \cdot V^{\bar{\pi}}$, cette quantité peut être exprimée en fonction de la fréquence de visite des états (voir Puterman (1994)) : $\beta \cdot V^{\bar{\pi}} = r \cdot f^{\bar{\pi}}$. Un MDP a toujours une politique optimale π^* tel que : $\pi^* = \operatorname{argmax}_\pi \beta \cdot V^\pi$ ou $f^* = \operatorname{argmax}_f r \cdot f$. Un IRMDP (Regan et Boutilier (2009)) est un tuple $M(S, A, P, r, \gamma, \beta)$ où S, A, P, γ et β sont définis comme dans le cas précédent, r est un ensemble de fonctions de récompense possible sur $S \times A$ et r modélise l'incertitude sur les valeurs des récompenses réelles. Comme c'est le cas des travaux de l'état de l'art (Alizadeh et al. (2015); Benavent et Zanuttini (2018); Regan et Boutilier (2009); Weng et Zanuttini (2013)), nous supposons que l'ensemble des récompenses possibles est représenté par un polytope $\mathcal{R} = \{r : Cr \leq d\}$.

Critère de regret minmax. Afin de résoudre l'IRMDP, nous utilisons le critère de *regret minmax* (voir Regan et Boutilier (2009); Xu et Mannor (2009)). Le *regret* d'une politique f^π basée sur une fonction de récompense $r \in \mathcal{R}$ est la perte ou la différence de valeur entre f et la politique optimale sous r , c'est à dire : $R(f^\pi, r) = \max_g r \cdot g - r \cdot f$. Le *regret maximal* pour une politique f^π est le regret maximal de cette politique par rapport à l'ensemble des récompenses \mathcal{R} défini comme suit : $MR(f^\pi, \mathcal{R}) = \max_{r \in \mathcal{R}} R(f^\pi, r)$.

En d'autres termes, quand nous devons sélectionner la politique f ; quelle est la pire des pertes sur toutes les récompenses possibles \mathcal{R} ? Considérant cela comme un jeu, l'adversaire essaie de trouver une valeur de récompense afin de maximiser notre perte. Enfin, Nous définissons le *regret minmax* de l'ensemble des récompenses réalisables possibles \mathcal{R} comme suit :

$$MM(\mathcal{R}) = \min_{f^\pi} MR(f^\pi, \mathcal{R}).$$

Toute politique f^* qui minimise le maximum de regret est une *politique de regret minmax optimale* Alizadeh et al. (2015); Benavent et Zanuttini (2018); Regan et Boutilier (2009); da Silva et Costa (2011); Xu et Mannor (2009). Dans cet article, nous utilisons l'approche présentée par Regan et Boutilier (2009) basée sur la décomposition de Benders (1962). L'idée est de formuler le problème sous la forme de séries de programmes linéaires (LP) et de programmes linéaires à nombres entiers mixtes (MILP) :

Programme Maître

$$\min_{\delta, f} \delta \quad (2)$$

$$\text{s.t. : } r \cdot g - r \cdot f \leq \delta \quad \forall (g, r) \in \text{GEN} \quad (3)$$

$$\gamma E^\top f + \beta = 0 \quad (4)$$

Le programme maître est un programme linéaire calculant le minimum de regret en ce qui concerne toutes les combinaisons possibles de récompenses et de politiques adverses. Nous appelons GEN l'ensemble contenant toutes les combinaisons de récompenses et de politiques adverses. Le deuxième ensemble de contraintes du problème principal, $\gamma E^\top f + \beta = 0$ garantit que f est une fonction de fréquence de visite valide. Par souci d'abréviation, la matrice E est générée en fonction de la fonction de transition P ; E est une matrice $|S||A| \times |S|$ avec une ligne pour chaque action, et une colonne pour chaque état :

$$E_{sa,s'} = \begin{cases} P(s'|s, a) & \text{if } s' \neq s \\ P(s'|s, a) - \frac{1}{\gamma} & \text{if } s' = s \end{cases}.$$

L'intuition derrière cette contrainte est liée au programme linéaire dual de l'équation de Bellman (voir par exemple Sutton et Barto (1998), Chapitre 4 ou Puterman (1994), Section 6.9).

Programme esclave

$$\max_{Q,V,I,r} \beta \cdot V - r \cdot f \quad (5)$$

$$\text{s.t. : } Q_a = r_a + \gamma P_a V \quad \forall a \in A \quad (6)$$

$$V \geq Q_a \quad \forall a \in A \quad (7)$$

$$V \leq (1 - I_a)M_a + Q_a \quad \forall a \in A \quad (8)$$

$$Cr \leq d \quad (9)$$

$$\sum_{a \in A} I_a = 1 \quad (10)$$

$$I_a(s) \in \{0, 1\} \quad \forall s \in S, a \in A \quad (11)$$

$$M_a = M^\top - M_a^\perp \quad \forall a \in A \quad (12)$$

Le programme esclave reçoit une politique possible f^* et recherche une politique et une valeur de récompense qui maximise le regret de la politique donnée. Si ce n'est pas le cas, la procédure s'arrête et f^* est la politique (stochastique) qui minimise le regret maximum. L'interaction entre les programmes maître et esclave peut être vue comme un jeu à deux joueurs. Le programme maître trouve une politique optimale qui minimise le regret des adversaires donnés jusqu'ici par le programme esclave, tandis que le programme esclave recherche un adversaire avec le gain maximum par rapport à la politique maître.

Le problème esclave est une reformulation du $\text{MR}(f, \mathcal{R})$ pour la politique reçue f du programme maître. La fonction objective $r \cdot g - r \cdot f$ est réécrite comme $\beta \cdot V - r \cdot f$. La contrainte (8) assure que l'équation (1) est satisfaite et les contraintes (9) et (10) assurent que $Q(s, a) = V(s), \forall s$.² I est une matrice $|S| \times |A|$ définissant la politique liée à V . Les contraintes (10) et (11) imposent d'avoir une politique déterministe, c'est-à-dire avec une et une seule action sélectionnée a par état s . Notez que le programme esclave propose un adversaire déterministe au programme maître, alors que le programme maître approxime toujours une politique stochastique.

2. Pour chaque a , nous savons que la constante M_a est égale à $M^\top - M^\perp$, où M^\top est la valeur de la politique optimale pour les valeurs de récompenses maximales et M^\perp est la Q-valeur pour la politique optimale avec les récompenses minimales sur \mathcal{R} .

3 Le schéma branch-and-bound pour trouver la politique déterministe optimale

Nous nous intéressons, dans ce qui suit, à la manière d'obtenir une politique déterministe optimale pour un IRMDP. L'algorithme utilisé pour atteindre cet objectif est le branch-and-bound (voir Bertsimas et Weismantel (2005), Section 11 pour une explication détaillée). Dans notre application, la racine de l'arborescence de l'algorithme est associée à l'ensemble des politiques déterministes, tandis qu'une branche est obtenue en sélectionnant un couple (s, a) d'état et d'action et en imposant ensuite la disjonction suivante sur les deux nœuds enfants : (1) $f_{s,a'} = 0, \forall a' \neq a$ pour le nœud enfant "gauche"; (2) $f_{s,a} = 0$ pour le nœud enfant "droit". Les disjonctions imposent au nœud enfant gauche de ne représenter que les politiques déterministes avec $f_{s,a} \neq 0$ (c'est-à-dire $\pi(s, a) = 1$). D'un autre côté, le nœud enfant droit représente les politiques déterministes avec $f_{s,a} = 0$ (c'est-à-dire $\pi(s, a) = 0$)³. Pour éviter d'explorer l'arbre entier, nous avons besoin d'un algorithme de calcul de borne inférieure pour élaguer les branches qui ne contiennent pas de politiques optimales. Dans notre cas, nous utilisons la politique stochastique optimale comme sous-estimateur de la politique déterministe optimale pour une branche donnée de l'arbre (nous sommes face à un problème de minimisation, de ce fait le sous-estimateur peut être vu comme une estimation optimiste de la politique). De cette façon, si un nœud a une politique stochastique supérieure à la meilleure politique déterministe trouvée jusqu'à présent, il n'est pas nécessaire de continuer à explorer cette branche et le nœud peut être élagué. L'ingrédient final du branch-and-bound est une procédure pour trouver des politiques déterministes réalisables. Dans notre implémentation, chaque politique stochastique calculée dans les procédures de délimitation est également déterministe; sa valeur peut être utilisée pour mettre à jour la meilleure valeur connue de la politique déterministe.

La figure 1 présente le pseudocode de notre implémentation de l'algorithme branch-and-bound. L'algorithme commence par initialiser la valeur de la politique déterministe la plus connue à $+\infty$ et la liste des nœuds inexplorés au nœud racine. La boucle *while* extrait un nœud inexploré de la liste, corrige le f correspondant à sa sous-région de politiques déterministes réalisables et calcule une borne inférieure avec la décomposition de Benders. Si la politique stochastique optimale résultante a un regret maximum δ^* supérieur ou égal au regret maximum inférieur trouvé jusqu'ici pour une politique déterministe, aucun nœud supplémentaire n'est créé et la boucle extrait un autre nœud de la liste. Si le nœud n'est pas élagué mais la solution stochastique est déterministe, la valeur de la meilleure solution déterministe est mise à jour à δ^* . Comme dernière option, si la solution stochastique n'est pas déterministe, un état s avec plus d'un f différent de zéro est trouvé et le $f_{s,a}^*$ avec la valeur la plus élevée est utilisée pour créer les deux prochains nœuds enfants.

4 Analyse théorique de la politique déterministe optimale

Dans cette section, nous introduisons d'abord le concept intuitif d'heuristique déterministe, moyen d'obtenir une politique déterministe réalisable à partir d'une politique optimale stochas-

3. Le nombre total de choix (c'est-à-dire le nombre de paires d'états-actions) est fini, donc la taille de l'arbre du branch-and-bound est également finie.

Algorithme 1 : branch-and-bound pour la recherche d'une politique déterministe optimale

```

1  $BestVal := +\infty$  /* borne inf initialisée à l'infini */
2  $\mathcal{N} = \{\{\emptyset\}\}$  /* la liste des nœuds ouverts est initialisée à l'ensemble vide */
3 while  $\mathcal{N} \neq \emptyset$  do
4   extraire le nœud  $N$  de  $\mathcal{N}$ 
5   for each  $f$  in  $N$  do :
6     fixer  $f = 0$  dans le problème maître
7     résoudre le problème maître avec la décomposition de Benders
8     soit  $(\delta^*, f^*) :=$  la solution optimale du problème maître
9     if  $\delta^* < BestVal$  then : /*comparer la politique stochastique à la meilleure politique
    déterministe*/
10      if  $f^*$  est déterministe then :
11         $BestVal = \delta^*$  /* mettre à jour la meilleure politique déterministe */
12      else : /* créer les deux nœuds enfants */
13        trouver  $f_{s,a}^*$  qui correspond à l'état non déterministe
14         $N_L := N \cup_{s' \neq s} f_{s',a}$ 
15         $N_R := N \cup f_{s,a}$ 
16         $\mathcal{N} := \mathcal{N} \cup N_L \cup N_R$ 

```

tique. Par la suite, nous analysons les situations dans lesquelles une telle heuristique pourrait fournir un maximum de regrets, loin de celui donné par la politique déterministe optimale.

Une heuristique de déterminisation. Pour obtenir une politique déterministe réalisable à partir d'une politique optimale stochastique, nous introduisons le concept intuitif d'heuristique de déterminisation. Cette heuristique sera comparée, dans les expérimentations à l'algorithme optimal proposé dans la section précédente. Soit f une valeur de fréquence de visite donnée pour une politique stochastique optimale. La politique déterminisée⁴ $\hat{\pi}$ peut être calculée comme suit : pour chaque $s' \in S$: trouver une action $a' = \operatorname{argmax}_{a \in A} f_{s',a}$ et fixer le reste de l'action à zéro : $\hat{f}_{s',a} = 0, \forall a \neq a'$. Et au final, récupérer la politique déterministe $\hat{\pi}$ obtenue.

Exemple : IRMDP trident. On définit un *IRMDP trident* avec trois états s_0, s_1, s_2 , trois actions a_0, a_1, a_2 , un facteur de réduction $\gamma = 1$, une fonction de transition $P(s_0|s_2, a_0) = 1$, $P(s_1|s_2, a_1) = 1$, $P(s_0|s_2, a_2) = T_0$ et $P(s_1|s_0, a_2) = T_1$ ⁵ et deux récompenses inconnues associées à s_0 et s_1 : $r(a_0) = r_0 \in [-A, +A]$ et $r(a_1) = r_1 \in [-A + B, +A + B]$ avec $A, B > 0$ et $A \gg B$. Ainsi, $\mathcal{R} = [-A, +A] \times [-A + B, +A + B]$. La distribution initiale sur les états est $\beta(s_0) = \beta(s_1) = 0$ et $\beta(s_2) = 1$.

Les propositions suivantes donnent une caractérisation complète des politiques optimales stochastiques et déterministes pour le MDP Trident. Par facilité de notation, nous utilisons

4. Nous utiliserons, dans la suite de ce papier, "politique déterminisée" toute politique déterministe obtenue à partir d'une politique stochastique en utilisant notre algorithme.

5. Dans cette formulation, les récompenses dépendent des états. Ils peuvent être facilement modifiés pour la notation de fonction de récompense donnée dans cet article $r(s, a)$

Calcul d'une politique déterministe dans un IRMDP

π_a à la place de $\pi(s_2, s_a)$, r_0 à la place de $r(s_0)$ et r_1 à la place de $r(s_1)$. Chaque politique stochastique sur le MDP trident peut être démontrée sous la forme d'un tuple $\pi = (\pi_0, \pi_1, \pi_2)$. Les fréquences sont également simplifiées comme suit : $f = (f_0, f_1, f_2)$.

Proposition 1. *La stratégie stochastique optimale minimisant le maximum de regret (voir la section 2) pour le MDP Trident est la stratégie $\tilde{\pi} = (\pi_0, \pi_1, \pi_2)$ définie comme suit :*

$$\pi_0 = \frac{2A - B}{4A}, \quad \pi_1 = \frac{2A + B}{4A}, \quad \pi_2 = 0.$$

Démonstration. Nous constatons d'abord que pour chaque politique $\pi' = (\pi'_0, \pi'_1, \pi'_2)$ avec $\pi'_2 > 0$, il est possible de construire une politique $\pi'' = (\pi''_0, \pi''_1, \pi''_2)$ avec $\pi''_2 = 0$ et de la même manière : $\pi''_0 = \pi'_0 + \pi'_2 T_0$, $\pi''_1 = \pi'_1 + \pi'_2 T_1$. Et si nous calculons la valeur de la première politique, on obtient : $\beta \cdot V^{\pi'} = V^{\pi'}(s_2) = r_0 \pi'_0 + r_1 \pi'_1 + r_0 T_0 \pi'_2 + r_1 T_1 \pi'_2 = r_0(\pi'_0 + T_0 \pi'_2) + r_1(\pi'_1 + T_1 \pi'_2) = r_0 \pi''_0 + r_1 \pi''_1 = V^{\pi''}(s_2) = \beta \cdot V^{\pi''}$. Ce qui montre que les deux politiques ont les mêmes valeurs. De plus, $\beta \cdot V^{\pi'} = \beta \cdot V^{\pi''}$, $\forall r \in \mathcal{R}$. On déduit que π' et π'' ont un maximum de regret équivalent à cause de l'égalité suivante : $MR(\pi', \mathcal{R}) = \max_r \max_g r \cdot g - \beta \cdot V^{\pi'} = \max_r \max_g r \cdot g - \beta \cdot V^{\pi''} = MR(\pi'', \mathcal{R})$. On peut supposer qu'il existe une politique stochastique optimale avec $\pi_2 = 0$ comme solution du regret minmax. La deuxième partie de la preuve consiste à calculer la valeur de la politique optimale en considérant $\tilde{\pi} = (\pi_0, \pi_1, 0)$ où $\pi_0, \pi_1 \geq 0$ de même $\tilde{f} = (f_0, f_1, 0)$. On constate que la politique de l'adversaire dont la fréquence de visite est donnée par g est aussi déterministe (voir 2). Il n'existe, de ce fait, que deux règles d'adversaires : $g = (g_0, g_1, g_2)$ où $g_0 = g_2 = 0$ et $g_1 > 0$ ou à l'inverse : $g' = (g_0, g_1, g_2)$ où $g_0 > 0$ et $g_1 = g_2 = 0$. Nous remarquons qu'avec des arguments analogues à ceux de la première partie de la preuve, nous pouvons exclure le cas où $g_2 \geq 0$.

Sachant que le regret maximum est le maximum entre deux choix pour les politiques de l'adversaire, le regret maximum associé à la politique $g = (0, g_1, 0)$ (obtenue en fixant $r_0 = -A$ et $r_1 = A + B$) est : $r \cdot g - r \cdot \tilde{f} = A + B + A\pi_0 - (A + B)\pi_1$. Et le regret maximum associé à la politique $g' = (g_0, 0, 0)$ où $g_0 > 0$ est obtenue en fixant $r_0 = A$ et $r_1 = -A + B$, donnant : $r \cdot g - r \cdot \tilde{f} = A - A\pi_0 - (B - A)\pi_1$.

Minimiser le regret maximum, consiste dans notre cas, à trouver les valeurs de π_0 et π_1 qui minimisent le maximum des deux quantités précédemment citées. La politique stochastique optimale peut donc être obtenue par la résolution du système à deux équations suivant : $A + B + A\pi_0 - (A + B)\pi_1 = A - A\pi_0 - (B - A)\pi_1$ et $\pi_0 + \pi_1 = 1$. Cela a comme solution optimale les valeurs $\pi_0 = \frac{2A - B}{4A}$ et $\pi_1 = \frac{2A + B}{4A}$, concluant la preuve. \square

De cette proposition 1, on déduit le lemme suivant :

Lemme 1. *La politique de déterminisation pour le MDP Trident est $\hat{\pi} = (0, 1, 0)$ et son regret maximum est $MR(\hat{f}, \mathcal{R}) = 2A - B$.*

Démonstration. C'est une conséquence directe du fait que dans la politique stochastique optimale nous avons toujours $\pi_1 > \pi_2$ et $\pi_0 = 0$. \square

Proposition 2. *Si $T_1 > T_0$ alors la politique déterministe optimale prendra la valeur de $\pi^* = (0, 0, 1)$ et son regret maximum sera de $MR(f^*, \mathcal{R}) = A - AT_0 + (A - B)T_1$.*

Démonstration. Cette preuve est apportée en calculant explicitement le maximum de regret des trois politiques déterministes possibles suivantes : $\pi = (1, 0, 0)$, $\pi' = (0, 1, 0)$, $\pi'' = (0, 0, 1)$

Le regret maximum (RM) pour $\pi = (1, 0, 0)$. Nous voulons trouver la politique de l'adversaire qui maximise le regret de la politique π . Nous le faisons en calculant toutes les combinaisons possibles de politiques et de récompenses de l'adversaire :

- Si la fréquence de visite pour la politique de l'adversaire est $g = (0, g_1, 0)$ où $g_1 > 0$, la récompense maximisant le regret sera $r_0 = -A$ et $r_1 = A + B$, conduisant à un regret maximum de (a) $A + B - (-A) = 2A + B$
- Si la politique de l'adversaire est $g' = (0, 0, g_2)$ où $g_2 > 0$, nous devons vérifier les quatre combinaisons de récompenses extrêmes :
 - $r_0 = -A$ et $r_1 = A + B$. RM de (b) $-AT_0 + (A + B)T_1 + A = (1 - T_0 + T_1)A + T_1B$
 - $r_0 = A$ et $r_1 = A + B$. RM de (c) $AT_0 + (A + B)T_1 - A = (-1 + T_0 + T_1)A + T_1B$
 - $r_0 = A$ et $r_1 = -A + B$. RM de (d) $AT_0 + (-A + B)T_1 - A = (-1 + T_0 - T_1)A + T_1B$
 - $r_0 = -A$ et $r_1 = -A + B$. MR de (e) $-AT_0 + (-A + B)T_1 + A = (1 - T_0 - T_1)A - T_1B$

Par construction, nous avons $A \gg B$ et $T_0 + T_1 = 1$, cela implique que l'équation (a) $\geq \max \{(b), (c), (d), (e)\}$. Par conséquent, le regret maximum si $g_2 > 0$ est de $MR(f^\pi, \mathcal{R}) = 2A + B$.

Regret maximum pour $\pi' = (0, 1, 0)$. Il est trivial de vérifier, avec des calculs analogues à celui utilisé ci-dessus pour calculer le regret de π , que le regret maximum est dans ce cas égal à $MR(f^{\pi'}, \mathcal{R}) = 2A - B$.

Regret maximum pour $\pi'' = (0, 0, 1)$. Dans ce cas également, nous devons examiner les deux cas suivants : $g = (g_0, 0, 0)$ où $g_0 > 0$ et $g' = (0, g_1, 0)$ avec $g_1 > 0$. Pour g , nous fixons $r_0 = A$ et $r_1 = -A + B$, obtenant ainsi un regret égal à (f) $A - AT_0 + (A - B)T_1$. Et pour g' , nous fixons $r_0 = -A$ et $r_1 = A + B$, obtenant un regret égal à (g) $A + B + AT_0 - (A + B)T_1$. Le maximum entre (f) et (g) dépend des valeurs de T_0 et T_1 . En fixant $A - AT_0 + (A - B)T_1 \geq A + B + AT_0 - (A + B)T_1$ nous obtenons : $2AT_1 \geq 2AT_0 + B$.

Nous rappelons que par construction nous avons $A \gg B$, cela implique que si $T_1 > T_0$ (resp. $T_1 \leq T_0$) le maximum de regret est alors égale à (f) (resp. (g)). Le minimum des regrets maximum trouvés jusqu'à présent est celui obtenu pour $\pi = \pi' = (0, 1, 0)$, qui est égal à $MR(f^{\pi'}, \mathcal{R}) = 2A - B$. Il reste donc à vérifier pour quelles valeurs de $T_0 > T_1$ nous avons $2A - B \geq (f) : A - AT_0 + (A - B)T_1 \leq 2A - B \Leftrightarrow A - A(1 - T_1) + (A - B)T_1 \leq 2A - B \Leftrightarrow (2A - B)T_1 \leq 2A - B \Leftrightarrow T_1 \leq 1$. Puisque nous avons par construction $T_1 \leq 1$, nous pouvons conclure que pour tout $T_1 > T_0$ la politique déterministe optimale est $\pi^* = \pi'' = (0, 0, 1)$ et son regret maximum est égal à $MR(f^{\pi''}, \mathcal{R}) = A - AT_0 + (A - B)T_1$. \square

La proposition 2 et le lemme 1 montrent que pour tout MDP Trident, la politique déterministe optimale et la politique déterministe sont toujours différentes. Le lemme suivant montre que la politique de déterminisation peut être bien pire que la politique déterministe optimale :

Lemme 2. *Le rapport entre le regret maximum de la politique déterminisée et la politique déterministe optimale passe à 2 avec l'augmentation de la valeur de A par rapport à B et l'augmentation de T_1 . En d'autres termes : $\lim_{A/B \rightarrow \infty, T_1 \rightarrow \frac{1}{2}^+} \frac{2A - B}{A - AT_0 + (A - B)T_1} = 2$*

Démonstration. La preuve découle du calcul de la limite. \square

Calcul d'une politique déterministe dans un IRMDP

$ S $	$ A $	VR	TR	% diff	Comp. Time
5	2	1.07	1.83	50%	2.59
	3	1.03	2.44	20%	5.05
	4	1.09	2.17	50%	5.28
	5	1.07	2.85	50%	8.03
	10	1.02	2.5	30%	13.76
10	2	1.11	4.11	90%	21.78
	3	1.15	7.63	80%	81.67
	4	1.04	9.19	60%	312.05
	5	1.06	8.42	90%	570.15
	10	1.01	18.79	90%	1886.05
15	2	1.04	6.91	60%	94.95
	3	1.05	18.75	80%	2240.4
	4	1.01	20.04	80%	5366.92
	5	1.03	32.1	100%	7677.25
Avg.		1.06	7.77	70%	1306.14

TAB. 1 – Ratio temporel et de valeurs pour les MDPs aléatoires.

D'un point de vue théorique, on ignore toujours si certains MDP peuvent avoir un ratio supérieur à 2 (ou même un ratio allant à l'infini). D'un point de vue pratique, ce petit exemple montre comment l'utilisation de la politique de déterminisation peut conduire à un maximum de regret de 100% loin de l'optimum.

5 Résultats expérimentaux

Nous allons proposer une évaluation expérimentale de nos algorithmes basée sur deux classes d'instances d'IRMDP : (1) MDPs (aléatoires) et (2) MDPs en (Diamants). Pour un MDP donné, soit $MR(f^{\hat{\pi}}, \mathcal{R})$ le regret maximum de la politique déterminisée et $MR(f^{\pi^*}, \mathcal{R})$ le regret maximum de la politique déterministe optimale. Nous définissons le Ratio de valeur de tels MDPs comme suit : $VR = MR(f^{\hat{\pi}}, \mathcal{R})/MR(f^{\pi^*}, \mathcal{R})$. De plus, soit \hat{T} (respectivement T^*) le temps de calcul nécessaire pour calculer la politique déterminisée (respectivement optimale), nous définissons le ratio temporel comme suit : $TR = T^*/\hat{T}$.

5.1 MDPs aléatoires

Analyse des résultats Dans le tableau 1, nous présentons les résultats concernant les performances de notre algorithme sur des MDP aléatoires (voir Alizadeh et al. (2015)) avec $|S| \in \{5, 10, 15\}$, $|A| \in \{2, \dots, 5, 10\}$. Pour chaque combinaison d'états et d'actions, nous avons présenté des résultats moyens supérieurs à 10 pour différentes simulations. Les deux premières colonnes indiquent le rapport de valeur et le rapport de temps (VR et TR). La colonne % diff montre le pourcentage de cas où la politique optimale est différente de la politique déterminisée. La dernière colonne montre le temps de calcul de l'algorithme de branch-and-bound (Base) présenté dans la Section 3. Nous remarquons qu'en moyenne, dans 70% du

temps, la politique déterministe optimale diffère de la politique déterminisée, tandis que le regret maximal de la politique déterminisée est de 6% plus mauvais que la politique déterministe optimale. Cet écart modéré est probablement dû au fait que les MDP aléatoires ne présentent pas de structure particulière. Le calcul de la politique déterministe optimale est à un ordre de grandeur plus lent que le calcul de la politique déterminisée. Cette écart est acceptable, si l'on considère que dans chaque nœud du branch-and-bound la procédure de délimitation se réduit à un problème aussi difficile que le calcul de la politique stochastique optimale.

5.2 MDPs en diamant

p	5	10	15	20	25	30	35	40	45	Avg.
VR	1.66	1.24	1.16	1.13	1.15	1.15	1.15	1.14	1.16	1.22
TR	10.23	7.44	6.32	6.48	7.67	5.93	7.62	10.46	13.80	8.44

TAB. 2 – *Le Ratio temporel et Le Ratio de valeurs pour les MDP en Diamant.*

Analyse des résultats Cette classe de MDPs a été introduite dans Benavent et Zanuttini (2018). Nous proposons une généralisation de cette famille de MDP en testant une gamme de paramètres pour la probabilité $p \in \{0.05, 0.10, \dots, 0.45\}$ pour atteindre le nœud enfant gauche (resp. droit) et une probabilité de $1 - p$ (resp. p) pour atteindre son nœud parent sinon. Dans la table 2, nous montrons comment le Ratio de Valeur change avec l'augmentation de p . Dans les MDPs en diamant, la situation est différente : le regret maximal de la politique déterminisée est de 20 % plus mauvaise que celle de la politique déterministe optimale. De plus, le temps de calcul de la politique déterministe optimale est inférieur d'un ordre de grandeur à celui requis par la politique déterminisée. Ces résultats montrent comment, en présence d'une structure spécifique, la différence entre $MR(f^{\hat{\pi}}, \mathcal{R})$ et $MR(f^{\pi^*}, \mathcal{R})$ augmente significativement.

6 Conclusions

Nous avons présenté un algorithme pour trouver une politique déterministe optimale qui minimise le regret maximal d'un processus de décision de Markov avec des récompenses imprécises et inconnues. L'algorithme proposé consiste en un branch-and-bound qui utilise la décomposition de Benders comme procédure de délimitation. Nous motivons (au niveaux théorique et expérimental) l'utilisation de politiques déterministes par rapport à des politiques stochastiques en montrant que les procédures de déterminisation de base peuvent trouver des politiques déterministes loin d'être optimales.

Références

Ahmed, A., P. Varakantham, M. Lowalekar, Y. Adulyasak, et P. Jaillet (2017). Sampling based approaches for minimizing regret in uncertain markov decision processes (mdps). *J. Artif. Intell. Res.* 59, 229–264.

- Alizadeh, P., Y. Chevaleyre, et J. Zucker (2015). Approximate regret based elicitation in markov decision process. In *RIVF*, pp. 47–52. IEEE.
- Benavent, F. et B. Zanuttini (2018). An Experimental Study of Advice in Sequential Decision-Making under Uncertainty. In *AAAI*.
- Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numer. Math.*, 238–252.
- Bertsimas, D. et R. Weismantel (2005). *Optimization Over Integers*.
- da Silva, V. F. et A. H. R. Costa (2011). A geometric approach to find nondominated policies to imprecise reward mdps. ECML PKDD'11, pp. 439–454.
- Mannor, S., O. Mebel, et H. Xu (2012). Lightning does not strike twice : Robust mdps with coupled uncertainty. *CoRR abs/1206.4643*.
- Mannor, S., D. Simester, P. Sun, et J. N. Tsitsiklis (2007). Bias and variance approximation in value function estimates. *Management Science*, 308–322.
- Nilim, A. et L. E. Ghaoui (2005). Robust control of markov decision processes with uncertain transition matrices. *Operations Research* 53(5), 780–798.
- Puterman, M. L. (1994). *Markov Decision Processes : Discrete Stochastic Dynamic Programming* (1st ed.). New York, NY, USA : John Wiley & Sons, Inc.
- Regan, K. et C. Boutilier (2009). Regret-based reward elicitation for markov decision processes. In *UAI*, pp. 444–451. AUAI Press.
- Sutton, R. S. et A. G. Barto (1998). *Introduction to Reinforcement Learning* (1st ed.). Cambridge, MA, USA : MIT Press.
- Weng, P. et B. Zanuttini (2013). Interactive value iteration for markov decision processes with unknown rewards. In *IJCAI*, pp. 2415–2421.
- Wiesemann, W., D. Kuhn, et B. Rustem (2013). Robust markov decision processes. *Mathematics of Operations Research* 38(1), 153–183.
- Xu, H. et S. Mannor (2009). Parametric regret in uncertain markov decision processes. In *CDC*, pp. 3606–3613. IEEE.

Summary

In some real world applications of sequential decision making under uncertainty, a stochastic policy is not easily interpretable for the system users. This might be due to the nature of the problem or to the system requirements. In these contexts, it is more convenient (inevitable) to provide a deterministic policy to the user. We propose an approach for computing a deterministic policy for a Markov Decision Process with Imprecise Rewards. To motivate the use of an exact procedure for finding a deterministic policy, we show some cases where the intuitive idea of using a deterministic policy obtained after “determinising” (rounding) the optimal stochastic policy leads to a deterministic policy different from the optimal.