# Problèmes de Plus Sûr et Plus Court Chemin Stochastique

Florent Teichteil-Königsbuch

florent.teichteil@onera.fr
Onera — The French Aerospace Lab
F-31055, Toulouse, France

**Résumé** : Résoudre de manière optimale des problèmes de Plus Court Chemin Stochastique (SSP en anglais) nécessite en général qu'il existe au moins une politique qui atteint le but avec probabilité 1 depuis l'état initial. Cette condition est très forte et empêche de résoudre de nombreux problèmes intéressants, par exemple où toutes les politiques possibles atteignent des états "culs-de-sac" avec une probabilité strictement positive. Nous introduisons un critère d'optimisation duale plus général et plus riche, qui minimise le coût moyen (non pondéré) des seuls chemins menant au but parmi toutes les politiques qui maximisent la probabilité d'atteindre le but. Nous présentons des équations de mise à jour de la politique sous la forme de la programmation dynamique pour ce nouveau critère dual. Ces équations sont différentes des équations standards de Bellman, mais elles produisent la même solution s'il existe une politique menant au but avec probabilité 1 depuis l'état initial. Nous démontrons que nos équations convergent en horizon infini sans aucune condition sur la structure du problème ni sur ses politiques, ce qui étend de fait la classe des problèmes de Plus Court Chemin Stochastique qui peuvent être résolus. Nous montrons expérimentalement que notre critère dual fournit des solutions bien fondées à des SSP qui n'ont pas de solution avec le critère standard, et que le fait d'utiliser un facteur d'actualisation avec ce dernier fournit certes des politiques solutions, mais qui ne sont pas optimales au regard du critère dual.

## 1 Introduction

Les recherches sur la planification probabiliste se sont essentiellement intéressées à soit maximiser la probabilité d'atteindre le but depuis l'état initial (Kolobov *et al.* (2011); Teichteil-Königsbuch *et al.* (2010); Puterman (1994)), ou à minimiser les coûts moyens cumulés s'il existe une politique qui atteint le but avec probabilité 1, appelée "politique correcte" (Bertsekas & Tsitsiklis (1996); Bonet & Geffner (2005, 2003); Kolobov *et al.* (2010, 2011); Yoon *et al.* (2010)). Cependant, du mieux que nous sachions, aucune approche existante optimise à la fois la probabilité d'atteindre le but, et minimise les coûts moyens cumulés, au sein d'un même cadre théorique cohérent. De plus, si la probabilité maximum d'atteindre le but est strictement inférieure à 1 pour un problème donné, c'est-à-dire s'il n'existe aucune politique correcte, il est toujours possible de minimiser les coûts moyens cumulés *actualisés* (Teichteil-Königsbuch *et al.* (2011)), mais cette approche présente l'inconvénient de prendre inutilement (et malheureusement) en compte les coûts des chemins qui n'atteignent pas le but lors de l'optimisation des coûts moyens cumulés.

Dans ce papier, nous proposons d'abord un nouveau critère dual d'optimisation en horizon infini, qui sélectionne les politiques minimisant les coûts moyens cumulés (non pondérés) des seuls chemins qui atteignent le but, parmi toutes les politiques maximisant la probabilité d'atteindre le but. Ce critère dual est souvent considéré comme une métrique d'évaluation intéressante (Younes *et al.* (2005)), mais, du mieux que nous sachions, aucun moyen théorique ni pratique n'existe pour optimiser ces métriques main dans la main. Nous illustrons les bénéfices de notre critère dual par rapport aux approches existantes, à l'aide d'exemples contenant à la fois des états but et des états "cul-de-sac". Ensuite, nous proposons des équations de mise à jour pour évaluer ce critère dual pour une politique stationnaire quelconque, et nous prouvons que ces équations convergent toujours vers des solutions de valeurs finies à mesure que l'horizon de raisonnement tend vers $+\infty$, sans aucune hypothèse sur la structure du problème considéré ni de ses politiques (contrairement aux approches précédentes).

Néanmoins, en pratique, il apparaît particulièrement difficile de construire des politiques stationnaires optimales pour ce critère dual dans le cas général, c'est-à-dire en présence de coûts positifs ou négatifs.

Ainsi, nous proposons des équations d'optimalité pour notre critère dual seulement dans le cas où tous les coûts sont positifs. Ces équations sont différentes des équations standards de Bellman, mais : (i) la complexité temporelle du calcul de leurs solutions est également polynomiale en le nombre d'états et d'actions du problème, et (ii) elles fournissent les même politiques optimales que les SSPs pour les problèmes où les hypothèses des SSPs sont vérifiées. Finalement, nous démontrons expérimentalement sur la base de plusieurs problèmes de référence, que les approches existantes – qui optimisent soit la probabilité d'atteindre le but, soit les coûts moyens cumulés le long de toutes les trajectoires (pas uniquement celles qui atteignent le but) – ne fournissent pas nécessairement des politiques optimales au sens de notre critère dual.

## 1.1 Processus Décisionnel Markovien orienté but

Nous considérons des problèmes de planification probabiliste définis sous la forme de Processus Décisionnels Markoviens (MDP en anglais) orientés but, qui sont des tuples $\langle S, A, T, c, G\rangle$ tels que (Bertsekas & Tsitsiklis (1996)) :

- $S$ est l'ensemble fini des états ;
- $G \subset S$ est l'ensemble fini des états but, supposés absorbants ($T(g, a, g) = 1, \forall a \in A$) ;
- $A$ est l'ensemble fini des actions ;
- $T : S \times A \times S \to [0; 1]$ est une fonction de transition telle que, pour tout $(s, a, s') \in S \times A \times S$ et étape de décision $t \in \mathbb{N}$, $T(s, a, s') = Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ ;
- $c : S \times A \times S \to \mathbb{R}$ est une fonction de coût telle que, pour tout $(s, a, s') \in S \times A \times S$, $c(s, a, s')$ est le coût payé par l'agent en exécutant l'action $a$ et en allant en une étape de l'état $s$ vers l'état $s'$ ; nous ne supposons pas des coûts positifs dans le cas général ; les buts sont supposés ne rien coûter à l'agent ($c(g, a, g) = 0, \forall a \in A$).

Nous notons $app : S \to 2^A$ la fonction qui indique l'ensemble des actions applicables dans un état donné.

La solution d'un MDP orienté but est une politique $\pi : S \to A$ qui optimise un critère donné, en général la probabilité d'atteindre le but depuis n'importe quel état initial, ou les coûts moyens cumulés depuis n'importe quel état initial.

## 1.2 Problèmes de Plus Court Chemin Stochastique

Des méthodes efficaces sont disponibles pour résoudre des MDP orientés but, à condition que deux hypothèses fondamentales sont vérifiées (Bertsekas & Tsitsiklis (1996)) : (i) il existe au moins une politique $\pi$ qui atteint le but avec probabilité 1, appelée *politique correcte*, et (ii) toutes les politiques incorrectes cumulent un coût moyen infini. L'hypothèse (ii) signifie que tous les cycles dans le graphe de transition, qui ne mènent pas à des états but, ont des coûts strictement positifs. Les problèmes qui vérifient ces deux hypothèses sont appelés *Plus Court Chemin Stochastique* (SSP en anglais). Les méthodes pour résoudre les SSP calculent le point fixe $C^*$ de l'équation de Bellman suivante, appelé *critère total*, qui est le coût moyen cumulé optimal le long de toutes les trajectoires commençant dans n'importe quel état initial $s$ :

$$C^*(s) = \min_{a \in app(s)} \sum_{s' \in S} T(s, a, s') \left(c(s, a, s') + C^*(s')\right) \tag{1}$$

Si les hypothèses des SSP ne sont pas vérifiées, par exemple en présence d'états "cul-de-sac" [1] atteignables avec un probabilité strictement positive en exécutant toutes les politiques possibles, l'équation précédente pourrait ne pas avoir de solution. Toutefois, cette équation peut être légèrement modifiée en multipliant $C^*(s')$ par un coefficient d'actualisation $0 < \gamma < 1$, donnant lieu au *critère $\gamma$-pondéré* $C^*_\gamma$, pour lequel il est prouvé qu'il existe toujours une solution (Puterman (1994)). Certains travaux (par exemple ceux de Teichteil-Königsbuch *et al.* (2011)) ont proposé des méthodes efficaces pour optimiser des MDP orientés but en présence d'états "cul-de-sac", basées sur le critère pondéré, mais nous montrerons dans la suite que de telles approches pourraient être inappropriées dans certains cas présentant des structures de coût complexes.

1. Un état *cul-de-sac* est un état depuis lequel aucun chemin ne mène au but avec une probabilité strictement positive, quelle que soit la politique exécutée.

# 2 Plus *Sûr* et Plus Court Chemin Stochastique

Le critère d'optimisation traditionnellement utilisé dans les SSP n'est pas nécessairement mathématiquement défini quand il n'existe pas de politique correcte, c'est-à-dire de politique qui atteint le but avec probabilité 1. En effet, dans ce cas, ce critère peut diverger en sommant un nombre infini de coûts positifs le long des chemins qui ne mènent pas au but. Si actualisé (critère $\gamma$-pondéré), il converge à coup sûr, mais : 1) il peut donner lieu à des politiques qui ne maximisent pas la probabilité d'atteindre le but (car les coûts des chemins qui n'atteignent pas le but pourraient attirer la politique vers ces chemins), et 2) il n'est de toute façon pas optimale au regard des coûts moyennés seulement le long des chemins qui atteignent le but. Nous pensons que la seule façon mathématiquement correcte d'optimiser d'une part la probabilité d'atteindre le but, et d'autre part le coût moyen cumulé uniquement le long des chemins qui mènent au but, est de séparer ces deux critères concurrents dans deux schémas d'évaluation et d'optimisation différents, mais parallèles.

## 2.1 Fonctions *probabilité de but* et *coût de but*

Pour un état $s \in S$, politique $\pi \in A^S$, et $n \in \mathbb{N}$ donnés, nous notons $P_n^{G,\pi}(s)$ la probabilité d'atteindre le but $G$ en au plus $n$ étapes en exécutant $\pi$ depuis $s$. Cette fonction est appelée *fonction de probabilité de but* en au plus $n$ étapes (restantes). En horizon fini, $\pi$ est une série de politiques $(\pi_0, \cdots, \pi_{H-1}), H \in \mathbb{N}$, où $\pi_k$ est la politique exécutée à l'étape $k$. Nous notons également $C_n^{G,\pi}(s)$ le coût moyen cumulé en exécutant $\pi$ depuis $s$, moyenné uniquement sur les chemins qui atteignent le but $G$ avec une probabilité positive. Nous l'appelons *fonction de coût de but* en au plus $n$ étapes (restantes). Il est important de remarquer que cette fonction est différente de la fonction de valeur utilisée traditionnellement dans les MDP, car cette dernière est moyennée sur tous les chemins partant de $s$ (et non uniquement ceux qui atteignent le but).

## 2.2 Critère d'optimisation dual en horizon infini

Nous prouverons plus loin dans cet article que les fonctions $P_n^{G,\pi}$ et $C_n^{G,\pi}$ convergent toutes les deux vers des valeurs finies lorsque l'horizon $H$ (ou les étapes restantes $n$) tendent vers $+\infty$, pour tout MDP orienté but, politique stationnaire $\pi$, et *sans aucune condition* sur la structure du MDP. Le lecteur notera que cette propriété assez forte est spécifique à nos fonctions de probabilité de but et de coût de but ; elle n'est en particulier pas valide pour les critères d'optimisation standards des MDP, dont la garantie de convergence est généralement conditionnée à l'existence de caractéristiques particulières pour la chaîne de Markov contrôlée sous-jacente, ou à celle d'un facteur d'actualisation $\gamma$ comme discuté précédemment.

Sur la base des métriques de probabilité de but et de coût de but, nous définissons les problèmes de *Plus Sûr et Plus Court Chemin Stochastique* ($S^3P$ en anglais), qui sont des MDP orientés but où, pour tout $s \in S$, nous cherchons une politique $\pi^*(s)$ qui *minimise le coût moyen accumulé moyenné uniquement sur les chemins qui atteignent le but depuis $s$, parmi toutes les politiques qui maximisent la probabilité d'atteindre le but* :

$$\pi^*(s) \in \underset{\pi:\forall s' \in S, \pi(s') \in \operatorname{argmax}_{\pi' \in A^S} P_\infty^{G,\pi'}(s')}{\operatorname{argmin}} C_\infty^{G,\pi}(s) \tag{2}$$

Les $S^3P$ incluent les problèmes de *Plus Court Chemin Stochastique* traditionnels (SSP ou $S^2P$ en anglais) : $S^2P \subset S^3P$. Il est intéressant de mentionner que les $S^3P$ incluent aussi le critère d'optimisation récemment proposé par Kolobov *et al.* (2011), qui constituait jusqu'alors la plus grande classe connue de MDP orientés but, appelée GSSP, pour optimiser soit le coût moyen cumulé parmi les seules politiques correctes (c'est-à-dire qui atteignent le but avec probabilité 1), les coûts immédiats pouvant être positifs ou négatifs, soit la fonction de probabilité de but. Néanmoins, les GSSP ne permettent pas d'optimiser à la fois ces deux critères, contrairement à l'approche présentée dans le présent article, et d'autant plus que nous prenons également en compte des coûts immédiats positifs ou négatifs. Ainsi, la classe de MDP orientés but, pour lesquels des solutions optimales et mathématiquement bien fondées existent, est dès lors étendue : $S^2P \subset$ GSSP $\subset S^3P$.

## 2.3 Exemples illustratifs

L'équation 2 montre clairement, pour un MDP orienté but donné, que les politiques optimales pour le critère $S^3P$ sont les mêmes que pour le critère SSP s'il existe une politique qui atteint le but avec probabilité

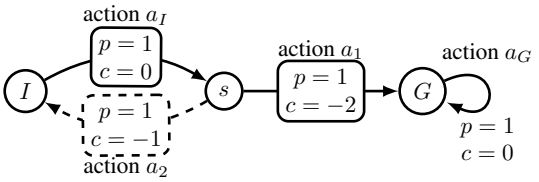


FIGURE 1 – S$^3$P avec existence de politique correcte. $I$ est l'état initial, $G$ l'état but, et $s$ un état intermédiaire.

1 (hypothèse (i) des SSP) et si de telles politiques optimales pour le critère SSP sont mathématiquement bien fondées (équivalent à l'hypothèse (ii)). Cette dernière hypothèse peut être violée pour deux raisons différentes : soit la fonction de valeur du SSP a une valeur infinie dans certains états initiaux, soit elle a une valeur finie mais ces valeurs ne peuvent malheureusement pas être obtenus par l'équation de Bellman (équation 1).

### 2.3.1 Exemple de MDP orientié but avec existence de politique correcte

La figure 1 présente un problème S$^3$P pour lequel l'hypothèse (i) des SSP est vérifiée, mais pas l'hypothèse (ii). Il existe donc des politiques stationnaires (une seule en fait) qui atteignent le but avec probabilité 1, mais aussi des politiques stationnaires qui n'atteignent jamais le but et qui cumulent des coûts négatifs, au point que l'équation de Bellman 1 ne peut pas fournir la politique optimale du SSP. Il y a dans cet exemple seulement deux politiques possibles, selon que les actions $a_1$ ou $a_2$ sont choisies dans l'état intermédiaire $s$. Nommons ces politiques respectivement $\pi_1$ et $\pi_2$. La politique $\pi_1$ est correcte, c'est-a-dire que la probabilité d'atteindre $G$ en partant de $I$ et en exécutant $\pi_1$ est 1. De plus, exécuter $\pi_1$ depuis $I$ permet à l'agent de cumuler a coût fini égal à $-2$. D'un autre côté, exécuter $\pi_2$ depuis $I$ ne lui permet jamais d'atteindre $G$, et il cumule de surcroît un coût de $-\infty$. Ainsi, le critère total utilisé classiquement dans les SSP n'est mathématiquement pas bien défini, au point qu'aucune politique SSP optimale n'existe. Même en supposant que $-\infty$ est une valeur minimum acceptable, elle ne peut être obtenue par programmation dynamique, car le critère d'arrêt ne sera alors jamais atteint. Cependant, $\pi_1$ est une politique S$^3$P mathématiquement bien définie : elle maximise la fonction de probabilité de but ($P_\infty^{G,\pi_1}(I) = 1 > 0 = P_\infty^{G,\pi_2}(I)$) et sa fonction de coût de but est finie ($C_\infty^{G,\pi_1}(I) = -2$). C'est donc l'unique politique S$^3$P optimale dans cet exemple.

Maintenant, supposons que le coût immédiat de l'action $a_2$ est 0 (au lieu de $-1$). Clairement, $\pi_1$ est encore une politique S$^3$P optimale avec la même fonction de coût de but. toutefois, le critère total utilisé dans les SSP est dès lors de valeur finie pour la politique $\pi_2$, et égal à 0 dans les états $I$ et $s$. Ainsi, $\pi_1$, dont le coût total est $-2 < 0$, est une politique SSP optimale. Malheureusement, cette politique optimale ne peut pas être obtenue en utilisant l'équation de programmation dynamique 1, car l'opérateur de cette équation n'est pas une contraction : selon la fonction de valeur initiale choisie pour les états $I$ et $s$, l'équation 1 pourrait malencontreusement attribuer un coût inférieur à la politique $\pi_2$. L'hypothèse (ii) des SSP, qui en fait n'est pas vérifiée dans cet exemple, garantit si vérifiée que l'équation 1 produit toujours des politiques optimales. En fait, les équations que nous présenterons dans la suite de cet article pour résoudre des S$^3$P, souffrent d'un problème similaire : malgré la garantie théorique d'existence de politiques optimales stationnaires pour tout S$^3$P, nous proposerons des moyens algorithmiques sous la forme de programmation dynamique uniquement pour des MDP dont tous les coûts sont strictement positifs (sauf ceux de l'état but).

### 2.3.2 Exemple de MDP orienté but sans politique correcte

La figure 2 illustre un MDP orienté but pour lequel ni les hypothèses (i) et (ii) des SSP sont vérifiées, mais pour lequel la fonction de valeur des SSP est toutefois finie – cela signifie qu'elle ne peut pas être obtenue à partir de l'équation 1. Il y a 4 politiques possibles, selon que l'action $a_1$ ou $a_2$ ou $a_3$ ou $a_I$ est choisie dans l'état initial $I$. Notons les respectivement $\pi_1$, $\pi_2$, $\pi_3$ et $\pi_4$. En partant de $I$, les politiques $\pi_1$ et $\pi_2$ mènent à $G$ (resp. $d$) avec la même probabilité $0.9 + 0.1 \times 0.5 = 0.95$ (resp. $0.05$). La politique $\pi_3$ mène à $G$ (resp. $d$) avec la probabilité $0.1 \times 0.5 = 0.05$ (resp. $0.95$). L'action $a_I$ est absorbante et ne mène pas du tout à $G$. Ainsi, les politiques qui maximisent la fonction de probabilité de but sont $\pi_1$ et $\pi_2$. Cependant, minimiser la fonction de valeur des SSP revient à choisir la politique $\pi_3$. En effet, en utilisant l'équation 1, la fonction de valeur de $s$ est 1, si bien que la fonction de valeur de $\pi_1$ dans $I$ est

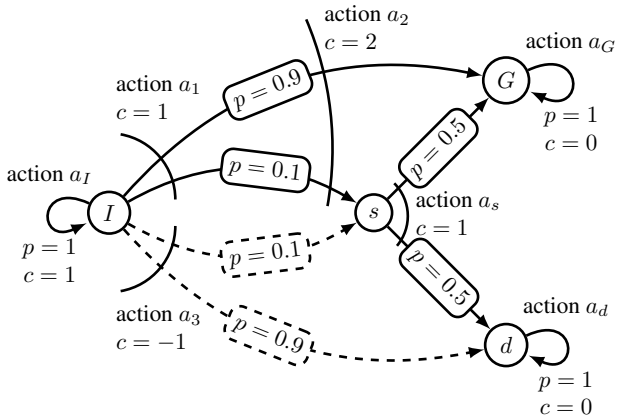


FIGURE 2 – S$^3$P sans politique correcte. $I$ est l'état initial, $G$ l'état but, $d$ un état "cul-de-sac", et $s$ un état intermédiaire.

$C^{\pi_1}(I) = 0.9 \times (1+0) + 0.1 \times (1+1) = 1.1$, celle de $\pi_2$ est $C^{\pi_2}(I) = 0.9 \times (2+0) + 0.1 \times (2+1) = 2.1$, et celle de $\pi_3$ est $C^{\pi_3}(I) = 0.1 \times (-1+1) + 0.9 \times (-1+0) = -0.9$. La politique $\pi_4$ a une fonction de valeur infinie et sera donc rejetée. Cela signifie que le critère d'optimisation standard des SSP, dans ce cas précis, ne sélectionne pas la politique qui maximise la probabilité d'atteindre le but depuis $I$.

Néanmoins, si nous optimisons d'abord la fonction de probabilité de but ainsi que définie précédemment, nous sélectionnerons les politiques $\pi_1$ et $\pi_2$. En exécutant la politique $\pi_1$ (resp. $\pi_2$) depuis $I$, on ne trouve que deux chemins menant au but : un premier chemin de probabilité $0.9$ et de coût total $1$ (resp. 2), un second chemin de probabilité $0.05$ et de coût total 2 (resp. 3). La probabilité cumulée de ces chemins est $0.95$, et peut être considérée comme une constante de normalisation du coût cumulé et moyenné sur ces deux chemins. Ainsi, la fonction de coût de but de $\pi_1$ ainsi que définie dans cet article, est : $1/0.95 \times (0.9 \times 1 + 0.05 \times 2) = 1/0.95 \simeq 1.05$. Celle de $\pi_2$ est : $1/0.95 \times (0.9 \times 2 + 0.05 \times 3) = 1.95/0.95 \simeq 1.85$. Par conséquent, $\pi_1$ est la politique S$^3$P optimale, dont la fonction de coût de but est égale à $1.05$. La fonction de valeur standard utilisée dans les SSP a une valeur supérieure pour $\pi_1$, car elle moyenne les coûts cumulés aussi sur le chemin menant à l'état "cul-de-sac" $d$ : $0.9 \times 1 + 0.05 \times 2 + 0.05 \times 2 = 1.1$.

Finalement, imaginons que le coût immédiat de l'action $a_d$ est 1 (au lieu de 0). Dans ce cas, la fonction de valeur utilisée dans les SSP diverge vers $+\infty$ depuis n'importe quel état, c'est-à-dire qu'elle n'est mathématiquement pas bien fondée. Toutefois, le lecteur pourra vérifier que les calculs du paragraphe précédents pour le critère S$^3$P sont encore valides et conduisent aux mêmes fonctions de probabilité de but et coût de but, et donc à la même politique optimale $\pi_1$. Il peut être pertinent de se demander si le fait de pondérer la fonction de valeur des SSPS avec un facteur d'actualisation $0 < \gamma < 1$, de manière à ce qu'elle ait une valeur finie, produira la même politique optimale qu'avec le critère S$^3$P. Nous prouverons plus loin dans la partie sur les expérimentations algorithmiques que ce n'est pas le cas pour des problèmes présentant des structures de coût complexes.

# 3 Évaluation des politiques S$^3$P en horizon fini

Dans cette section, nous présentons un théorème pour évaluer des politiques S$^3$P en horizon fini, ce qui est fondamental pour étudier les propriétés mathématiques des fonctions de probabilité de but et de coût de but. Nous pourrions prouver que $P_n^{G,\pi}$ peut être calculé via une traduction du MDP original vers un MDP où tous les revenus sont égaux à 0 exceptés ceux des transitions directes vers le but, qui sont égales à 1 (Kolobov *et al.* (2011)). Cependant, l'équation de mise à jour de la fonction de coût de but, présentée ci-dessous, n'est pas équivalente aux équations d'évaluation de Bellman pour les MDP, car les coûts cumulés sont moyennés uniquement sur les chemins menant au but, et pas sur tous les chemins comme dans les SSP.

**Théorème 1.** ***(Équations d'évaluation des politiques S$^3$P en horizon fini)*** *Soit $H \in \mathbb{N}$ l'horizon fini du problème. Pour toute étape de décision restante $1 \leqslant n < H$, toute politique histoire-dépendante $\pi =$*

$(\pi_0, \cdots, \pi_{H-1})$ *et tout état* $s \in S$ *:*

$$P_n^{G,\pi}(s) = \sum_{s' \in S} T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s'), \text{ avec :}$$

$$P_0^{G,\pi}(s) = 0, \forall s \in S \setminus G, \text{et } P_0^{G,\pi}(g) = 1, \forall g \in G \quad (3)$$

*Si* $P_n^{G,\pi}(s) > 0$*,* $C_n^{G,\pi}(s)$ *est défini, et satisfait :*

$$C_n^{G,\pi}(s) = \frac{1}{P_n^{G,\pi}(s)} \sum_{s' \in S} T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s') \left[ c(s, \pi_{H-n}(s), s') + C_{n-1}^{G,\pi}(s') \right], \text{ avec :}$$

$$C_o^{G,\pi}(s) = 0, \forall s \in S \quad (4)$$

*Démonstration.* L'équation 3 peut être facilement obtenue à l'aide de raisonnement similaire, mais plus simple, à celui nous permettant de démontrer l'équation 4. Par conséquent, nous présentons seulement la démonstration de cette dernière. Soit $\Phi_n^{G,\pi}(s)$ l'ensemble des chemins qui mènent au but $G$ en au plus $n$ étapes de décision en exécutant la politique $\pi$ depuis l'état $s$. Pour tout chemin $\phi \in \Phi_n^{G,\pi}(s)$, nous notons $|\phi|$ la longueur de $\phi$ jusqu'à ce qu'il atteigne le but, $\phi(i)$ le $i^{\text{ème}}$ état visité dans le chemin pour $0 \leqslant i \leqslant |\phi|$, et $\phi_i$ le sous-chemin de $\phi$ commençant en $\phi(i)$.

Le calcul de $C_n^{G,\pi}$ est moyenné sur la base d'une distribution de probabilité conditionnelle, conditionnée sur les seules trajectoires qui atteignent le but. Ainsi, ce calcul nécessite de calculer d'abord l'équation de mise à jour de la probabilité conditionnelle suivante ($n \geqslant 1$), où nous notons $\omega_{s,n}^{G,\pi}$ l'événement "l'exécution de $\pi$ depuis l'état $s$ produira un chemin qui atteindra le but en au plus $n$ étapes restantes", qui n'a de sens que si $P_n^{G,\pi}(s) > 0$ (nous avons $Pr(\omega_{s,n}^{G,\pi}) = P_n^{G,\pi}(s)$, et les conditionnements des probabilité vis-à-vis de $\pi$ sont implicites) :

$$\begin{aligned}
p_{\phi,n}^{G,\pi} &= Pr(\text{“exécuter } \phi\text{”} \mid \omega_{\phi(0),n}^{G,\pi}) \\
&= \frac{Pr(\omega_{\phi(0),n}^{G,\pi} \mid \text{“exécuter } \phi\text{”}) Pr(\text{“exécuter } \phi\text{”})}{Pr(\omega_{\phi(0),n}^{G,\pi})} \\
&= \frac{Pr(\omega_{\phi(1),n-1}^{G,\pi} \mid \text{“ex. } \phi_1\text{”}) Pr(\text{“}\phi(0) \text{ to } \phi(1)\text{”}) Pr(\text{“ex. } \phi_1\text{”})}{P_n^{G,\pi}(\phi(0))} \\
&= T(\phi(0), \pi_{H-n}(\phi(0)), \phi(1)) \frac{P_{n-1}^{G,\pi}(\phi(1))}{P_n^{G,\pi}(\phi(0))} \underbrace{\frac{Pr(\omega_{\phi(1),n-1}^{G,\pi} \mid \phi_1) Pr(\text{“ex. } \phi_1\text{”})}{P_{n-1}^{G,\pi}(\phi(1))}}_{Pr\left(\text{“ex. } \phi_1\text{”} \middle| \omega_{\phi(1),n-1}^{G,\pi}\right) = p_{\phi_1,n-1}^{G,\pi}}
\end{aligned}$$

Maintenant, en notant $c(\phi)$ le coût cumulé le long d'un chemin $\phi$, nous obtenons pour tout état $s$ tel que $P_n^{G,\pi}(s) > 0$ :

$$\begin{aligned}
C_n^{G,\pi}(s) &= \sum_{\phi \in \Phi_n^{G,\pi}(s)} Pr(\text{“executing } \phi\text{”} \mid \omega_{s,n}^{G,\pi})\, c(\phi) \\
&= \sum_{\phi \in \Phi_n^{G,\pi}(s)} p_{\phi,n}^{G,\pi} \times (c(s, \pi_{H-n}(s), \phi(1)) + c(\phi_1)) \\
&= \sum_{\phi \in \Phi_n^{G,\pi}(s)} T(s, \pi_{H-n}(s), \phi(1)) \frac{P_{n-1}^{G,\pi}(\phi(1))}{P_n^{G,\pi}(s)} p_{\phi_1,n-1}^{G,\pi} \times (c(s, \pi_{H-n}(s), \phi(1)) + c(\phi_1)) \\
&= \frac{1}{P_n^{G,\pi}(s)} \sum_{s' \in S} T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s') \sum_{\phi_1 \in \Phi_{n-1}^{G,\pi}(s')} p_{\phi_1,n-1}^{G,\pi} \times (c(s, \pi_{H-n}(s), s') + c(\phi_1)) \\
&= \frac{1}{P_n^{G,\pi}(s)} \sum_{s' \in S} T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s') \Big[ c(s, \pi_{H-n}(s), s') + \underbrace{\sum_{\phi_1 \in \Phi_{n-1}^{G,\pi}(s')} p_{\phi_1,n-1}^{G,\pi}\, c(\phi_1)}_{C_{n-1}^{G,\pi}(s')} \Big]
\end{aligned}$$

La dernière étape du calcul est due au fait que $\sum_{\phi_1 \in \Phi_{n-1}^{G,\pi}(s')} p_{\phi_1,n-1}^{G,\pi} = 1$. □

La division par $P_n^{G,\pi}(s)$ dans l'équation 4 pourrait paraître surprenante, mais il s'agit en fait d'une constante de normalisation de l'espérance qui définit la fonction de coût de but. En effet, la somme de $T(s, \pi_{H-n}(s), s')P_{n-1}^{G,\pi}(s')$ sur les états successeurs $s'$ dans l'équation 4 est en fait égale à $P_n^{G,\pi}(s)$.

Les équations 3 et 4 d'évaluation de politiques S$^3$P pourraient nous permettre de proposer des équations de mise à jour sous forme de programmation dynamique pour obtenir des politiques S$^3$P histoire-dépendantes optimales en horizon fini. Pour cela, il suffirait de remplacer l'équation 3 par une opération de maximisation sur les actions applicables, et l'équation 4 par une opération de minimisation sur les actions de probabilité maximum. Toutefois, nous nous concentrons dans cet article sur l'optimisation de MDP orientés but en horizon infini, qui est souvent plus utile en pratique, mais qui est aussi théoriquement bien plus compliquée.

# 4 Optimisation de problèmes S$^3$P en horizon infini

Le cas de l'horizon infini est primordial dans de nombreuses applications, par exemple si l'on ne connaît pas à l'avance le nombre minimum d'étapes de décision nécessaires pour atteindre le but depuis l'état initial, ou pour à la fois maximiser la fonction de probabilité de but et minimiser la fonction de coût de but. Comme nous l'avons mentionné précédemment, ce travail a été essentiellement motivé par des problèmes d'horizon infini, où le critère SSP ne fournit pas de solutions s'il n'existe pas de politique atteignent le but avec probabilité 1 – à moins de pondérer tous les coûts immédiats par un facteur d'actualisation $0 < \gamma < 1$ (même ceux qui appartiennent à des trajectoires ne menant pas au but !), ce qui est de fait rarement une métrique d'optimisation souhaitée. Il y aurait en effet une probabilité positive, pour toutes les politiques y compris celles qui sont optimales, de prendre en compte les coûts liés aux chemins indésirables.

Cette remarque est en fait la clef pour comprendre l'intuition liée à notre critère d'optimisation dual, et pourquoi il converge toujours vers un unique point fixe en horizon infini. La fonction de probabilité de but converge parce que : 1) les états depuis lesquels aucun chemin ne mène au but sont systématiquement affectés d'une valeur constante 0 après chaque mise à jour ; 2) les autres états mènent nécessairement à $G$ (avec une probabilité croissante) ou dans un des états précédemment mentionnés, ce qui assure la convergence de la fonction de probabilité de but. La fonction de coût de but converge car : 1) elle n'est pas définie pour les états depuis lesquels aucun chemin ne mène au but (ainsi que souhaité) ; 2) les coûts des autres états sont cumulés (et moyennés) uniquement le long des chemins qui mènent au but, dont les probabilités de présence convergent vers 0 au fur et à mesure que la longueur considérée de ces chemins tend vers $+\infty$, et qui ne payent plus de coûts une fois le but atteint. En comparaison, le critère d'optimisation utilisé dans les SSP cumule les coûts aussi le long des chemins qui ne mènent pas au but (si tant est qu'il en existe), dont les coûts peuvent éventuellement diverger vers $\pm\infty$ dans le cas général, au point de supprimer toute garantie de convergence de la fonction de valeur.

Les fondements mathématiques de cette intuition sont en fait relativement complexes, et reposent sur le lemme suivant, qui prouve que *l'opérateur de transition du MDP restreint à un sous-ensemble stable des états de $S \setminus G$ menant au but avec une probabilité (strictement) positive pour une politique stationnaire donnée, est une contraction.* Ce lemme peut être vu comme une généralisation non triviale de la propriété de contraction de l'opérateur de transition des politiques correctes, utilisé dans les SSP (voir Bertsekas & Tsitsiklis (1996)). Cet opérateur, pour les SSP et contrairement à notre approche, doit être une contraction sur l'espace d'état tout entier.

**Lemme 1.** *Soit $\mathcal{M}$ un MDP orienté but quelconque, $\pi$ une politique stationnaire, $T^\pi$ la matrice de transition pour la politique $\pi$, et pour tout $n \in \mathbb{N}$, $\mathcal{X}_n^\pi = \{s \in S \setminus G : P_n^{G,\pi}(s) > 0\}$. Alors : (i) pour tout $s \in S$, $P_n^{G,\pi}(s)$ converge vers une valeur finie lorsque $n$ tend vers $+\infty$ ; (ii) il existe $\mathcal{X}^\pi \subset S$ tel que $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ pour tout $n \in \mathbb{N}$ et $T^\pi$ est une contraction sur $\mathcal{X}^\pi$.*

*Démonstration.* (i) Un raisonnement par récurrence simple, en utilisant l'équation 3 montre que, pour tout $s \in S$, $P_n^{G,\pi}(s)$ est croissante avec $n$. Comme toutes les valeurs $P_n^{G,\pi}(s)$ sont bornées par 1, elles convergent vers une valeur finie $P_\infty^{G,\pi}(s)$ pour tout $s \in S$. (ii) Ce raisonnement par récurrence montre aussi que : $\forall n \in \mathbb{N}, \mathcal{X}_n^\pi \subset \mathcal{X}_{n+1}^\pi \subset S$. Comme $S$ est fini, il existe $\mathcal{X}^\pi \in S$ and $n_0 \in \mathbb{N}$ tel que pour tout $n \in \mathbb{N}$, $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ et pour tout $n \geqslant n_0$, $\mathcal{X}_n^\pi = \mathcal{X}^\pi$. Ainsi, pour tout $s \in \mathcal{X}^\pi$ et $n \geqslant n_0$, $P_n^{G,\pi}(s) > 0$. De plus, ainsi que précédemment mentionné, $P_n^{G,\pi}(s)$ est croissant avec $n$, de sorte que : pour tout $s \in \mathcal{X}^\pi$,

$P_\infty^{G,\pi}(s) = \lim_{n\to+\infty} P_n^{G,\pi}(s) > 0$. Par conséquent, la probabilité que n'importe quel état de $\mathcal{X}^\pi$ soit absorbé par le but est strictement positive, ce qui signifie que $\mathcal{X}^\pi$ est un sous-ensemble des états transitoires de la chaîne de Markov induite par la politique $\pi$. Soit $W^\pi$ la sous-matrice de $T^\pi$ qui associe les états transitoires à eux-mêmes (transitions entre les états transitoires). Il a été démontré (cf. Proposition A.3 dans Puterman (1994)) que $W^\pi$ est une contraction, c'est-à-dire que $\rho(W^\pi) < 1$, où $\rho(W^\pi)$ est la plus grande valeur propre de $W^\pi$ en valeur absolue. Aussi, en réordonnant les états transitoires de telle manière que les états de $\mathcal{X}^\pi$ apparaissent en premier, nous pouvons re-écrire $W^\pi$ sous la forme : $W^\pi = \begin{pmatrix} T^\pi_{|\mathcal{X}^\pi} & A^\pi \\ 0 & B^\pi \end{pmatrix}$. En effet, si la sous-matrice en bas à gauche n'était pas nulle, nous pourrions aller d'un état $\tilde{s}$ n'appartenant pas $\mathcal{X}^\pi \cup G$ vers un état dans $\mathcal{X}^\pi$ avec une probabilité non nulle, depuis lequel nous pourrions ensuite aller vers le but (par définition de $\mathcal{X}^\pi$) : cela signifie qu'il existerait $n_1 \in \mathbb{N}$ tel que $P_{n_1}^{G,\pi}(\tilde{s}) > 0$, ce qui contredit le fait que $P_n^{G,\pi}(s) = 0$ pour tout état $s \notin \mathcal{X}^\pi \cup G$ et $n \in \mathbb{N}$. Finalement, grâce à la forme précédente de $W^\pi$, nous avons : $\rho(T^\pi_{|\mathcal{X}^\pi}) \leqslant \rho(W^\pi) < 1$, c'est-à-dire $T^\pi$ est une contraction sur $\mathcal{X}^\pi$. □

## 4.1 Évaluation de politiques S$^3$P en horizon infini

Grâce au précédent lemme, nous pouvons dès lors démontrer la convergence des équations d'évaluation et d'optimisation des politiques S$^3$P. Comme pour les MDP standards, nous introduisons un opérateur de mise à jour qui s'avère utile pour démontrer cette convergence. Pour un $n \in \mathbb{N}^*$ et une politique stationnaire $\pi$ donnés, nous notons $\mathcal{L}_n^\pi$ l'opérateur suivant, défini sur l'ensemble des fonctions $J : S \to \mathbb{R}$ :

$$(\mathcal{L}_n^\pi J)(s) = \sum_{s' \in S} T(s, \pi(s), s')[P_{n-1}^{G,\pi}(s')c(s, \pi(s), s') + J(s')]$$

où $P_{n-1}^{G,\pi}$ est défini de manière récursive dans le théorème 1.

**Théorème 2.** *Soit $\mathcal{M}$ un MDP orienté but quelconque, et $\pi$ une politique stationnaire quelconque pour $\mathcal{M}$. Les équations d'évaluation du théorème 1 convergent vers des valeurs finies $P_\infty^{G,\pi}(s)$ et $C_\infty^{G,\pi}(s)$ pour tout $s \in S$ (par convention, $C_n^{G,\pi}(s) = 0$ if $P_n^{G,\pi}(s) = 0, n \in \mathbb{N}$).*

*Démonstration.* Comme le montre le lemme 1, la convergence de la série des fonctions de probabilité de but est indépendante des fonctions de coût de but. En notant $\mathcal{X}_n^\pi = \{s \in S \setminus G : P_n^{G,\pi}(s) > 0\}$, ce lemme montre également qu'il existe $\mathcal{X}^\pi \subset S$ tel que $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ pour tout $n \in \mathbb{N}$ et $T^\pi$ est une contraction sur $\mathcal{X}^\pi$.

Nous pouvons remarquer en étudiant l'équation 4, que pour tout $n \in \mathbb{N}$, $C_n^{G,\pi}(s) = 0$ pour tout $s \in G$. Pour les états $s \in S \setminus (G \cup \mathcal{X}^\pi)$, $P_n^{G,\pi}(s) = 0$ pour tout $n \in \mathbb{N}$ (car $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ pour tout $n \in \mathbb{N}$), si bien que $P_n^{G,\pi}(s)$ n'est pas défini mais systématiquement égal à 0 par convention. Par conséquent, l'opérateur $\mathcal{L}_n^\pi$, restreint pour opérer sur le sous-espace des fonctions $\Gamma = \{J : S \to \mathbb{R}\,; J(s) = 0, s \in S \setminus \mathcal{X}^\pi\}$, est équivalent à l'équation de mise à jour 4, ce qui signifie que cette dernière converge si et seulement si $\mathcal{L}_n^\pi$ converge sur l'ensemble de fonctions $\Gamma$. Nous allons de fait démontrer que $\mathcal{L}_n^\pi$ est une contraction sur $\Gamma$ ; pour tout $J_1$ et $J_2$ dans $\Gamma$, et $s \in \mathcal{X}^\pi$, nous avons :

$$\begin{aligned}
|(\mathcal{L}_n^\pi J_1)(s) - (\mathcal{L}_n^\pi J_2)(s)| &\leqslant \max_{s \in \mathcal{X}^\pi} |(\mathcal{L}_n^\pi J_1)(s) - (\mathcal{L}_n^\pi J_2)(s)| \\
&= \max_{s \in \mathcal{X}^\pi} \left| \sum_{s' \in S} T(s, \pi(s), s')(J_1(s') - J_2(s')) \right| \\
&= \max_{s \in \mathcal{X}^\pi} \left| \sum_{s' \in \mathcal{X}^\pi} T^\pi_{|\mathcal{X}^\pi}(s, s')(J_1(s') - J_2(s')) \right| \\
&= \| T^\pi_{|\mathcal{X}^\pi}(J_1 - J_2) \|_{\mathcal{X}^\pi} \\
&\leqslant \rho\left(T^\pi_{|\mathcal{X}^\pi}\right) \cdot \|J_1 - J_2\|_{\mathcal{X}^\pi}
\end{aligned}$$

Ainsi : $\|(\mathcal{L}_n^\pi J_1) - (\mathcal{L}_n^\pi J_2)\|_{\mathcal{X}^\pi} \leqslant \rho\left(T^\pi_{|\mathcal{X}^\pi}\right) \cdot \|J_1 - J_2\|_{\mathcal{X}^\pi}$. De plus, par définition de $\Gamma$ : $\|(\mathcal{L}_n^\pi J_1) - (\mathcal{L}_n^\pi J_2)\|_{S \setminus \mathcal{X}^\pi} = 0$. Comme $T^\pi_{|\mathcal{X}^\pi}$ est une contraction, $\mathcal{L}_n^\pi$ est lui-même une contraction sur $\Gamma$, pour tout

$n \in \mathbb{N}^*$, et sa constante de contraction ainsi que $\Gamma$ ne dépendent pas de $n$. Par conséquent, selon le théorème (généralisé) du point fixe de Banach, toute suite de fonctions $J_n$ de $\Gamma$, telle que $J_{n+1} = \mathcal{L}_n^\pi J_n$, converge vers un unique point fixe $J_\infty = P_\infty^{G,\pi} C_\infty^{G,\pi} \in \Gamma$. □

## 4.2 Optimisation de politiques S$^3$P en horizon infini

Nous venons de prouver que, pour toute politique stationnaire $\pi \in A^S$, les fonctions de probabilité de but et de coût de but sont mathématiquement définies (c'est-à-dire qu'elles ont des valeurs finies) en horizon infinit, et qu'elles peuvent être calculées de manière itérative avec les équations 3 et 4. Ainsi, puisque le nombre d'états et d'actions est fini, et donc aussi le nombre de politiques stationnaires, nous pouvons immédiatement établir la proposition suivante, qui prouve que *tout* problème S$^3$P en horizon infini a une solution associée à des fonctions de probabilité de but et de coût de but finies.

**Proposition 1.** *Soit $\mathcal{M}$ un MDP orienté but quelconque. (I) Il existe une politique stationnaire optimale $\pi^*$ qui minimise la fonction de coût de but en horizon infini, parmi toute les politiques qui maximisent la fonction de probabilité de but en horizon infini, c'est-à-dire $\pi^*$ est solution de l'équation 2. (II) Les fonctions limites de probabilité de but $P_\infty^{G,\pi^*}$ et de coût de but $C_\infty^{G,\pi^*}$ ont des valeurs finies (c'est-à-dire sont mathématiquement bien définies).*

Cette proposition est très générale ; en particulier, elle permet aux concepteurs de systèmes autonomes de s'intéresser à certains problèmes de MDP orientés but embêtants, où l'hypothèse (i) des SSP est certes vérifiée, mais pas l'hypothèse (ii). Souvenons-nous que l'hypothèse (ii) des SSP impose que le graphe de transition du MDP ne doit pas contenir de cycles avec des coûts non (strictement) positifs, composés d'états non connectés au but, ce qui garantit que le critère total standard utilisé dans les SSP est mathématiquement bien fondé (absence de tels cycles avec coûts strictement négatifs) *et* qu'il peut être optimisé par programmation dynamique (pas de tels cycles avec coûts nuls). Aucune des deux conditions n'est nécessaire dans les S$^3$P, car la fonction de coût de but n'est optimisée que sur l'ensemble des états transitions atteignant le but avec une probabilité non nulle (c'est-à-dire pas sur les cycles composés d'états ne menant pas au but). En pratique, cela signifie que le critère S$^3$P permet dès lors aux concepteurs de systèmes autonomes de résoudre des problèmes de plus court chemin contenant des cycles d'état avec coûts négatifs ou nuls, ou également en l'absence de politiques propres.

Néanmoins, même si la proposition 1 garantit l'existence de politiques stationnaires optimales pour tout problème S$^3$P, il n'est pas pour autant acquis que le calcul de telles politiques soit aisé en pratique. En d'autres termes, il n'existe pas nécessairement de moyens pratiques algorithmiques pour optimiser les fonctions $P_\infty^{G,\pi^*}$ et $C_\infty^{G,\pi^*}$ dans le cas général. La raison en est relativement compliquée et identique à une problématique similaire qui se pose lors de l'optimisation du critère total pour les MDP dans le cas général (cf. les travaux de Dawen (1986) et les chapitres 7 et 10 de Puterman (1994)), ou plus spécifiquement lors de l'optimisation des SSP contenant des cycles de coûts nuls composés d'états ne menant pas au but. Dans notre contexte, le problème se pose dès que la fonction de probabilité de but a convergé : sa valeur limite peut être différente de celle de la politique retournée par l'équation de mise à jour. Par exemple, considérons l'exemple représenté dans la figure 2 : dès que la fonction de probabilité de but optimale a été obtenue ($0.95$), l'application de l'action $a_I$ depuis l'état $I$ fournit en une étape (additionnelle) la même probabilité de but optimale que les actions $a_1$ et $a_2$ en deux étapes ($1 \times 0.95 = 0.95$), mais la fonction de probabilité de but de la politique stationnaire $\pi_4 = (a_I, a_I, \cdots)$ vaut en réalité $0$. Ainsi, puisque l'optimisation de la fonction de probabilité de but ne produit pas nécessairement des politiques dont la fonction de probabilité de but est égale à la fonction de probabilité de but optimisée, la fonction de coût de but – qui dépend de la fonction de probabilité de but d'après l'équation 4 – ne converge pas nécessairement.

Heureusement, nous avons pu obtenir des équations de mise à jour légèrement différentes, présentées ci-dessous, dont nous avons pu prouver qu'elles convergent vers des politiques stationnaires optimales, *à condition que toutes les transitions depuis des états non buts ont des coûts strictement positifs*. Notre intuition est la suivante : dès que la fonction de probabilité de but optimale a convergé, l'optimisation itérative des fonctions de coût de but sélectionne *indirectement* des politiques stationnaires dont la fonction de probabilité de but est égale à la fonction de probabilité de but optimisée, en rejetant les autres politiques qui ont *nécessairement* des fonctions de coût de but plus élevées. En effet, si l'on analyse le schéma d'optimisation suivant, nous réalisons que toutes les politiques dont la probabilité de but est inférieure à la probabilité de but optimisée ont un coût de but infini, si bien qu'elles sont automatiquement écartées en minimisant la

fonction de coût de but. La preuve mathématique est relativement complexe, et de fait non présentée dans cet article.

**Théorème 3.** *Soit $\mathcal{M}$ un MDP orienté but dont toutes les transitions depuis les états non buts ont des coûts strictement positifs. Soit $P_n^* : S \to [0;1]$ la série de fonctions définie comme :*

$$P_n^*(s) = \max_{a \in app(s)} \sum_{s' \in S} T(s,a,s') P_{n-1}^*(s'), \text{ avec :}$$

$$P_0^*(s) = 0, \forall s \in S \setminus G; P_0^*(g) = 1, \forall g \in G \quad (5)$$

*Les fonctions $P_n^*$ convergent vers des fonctions $P_\infty^*$ de valeurs finies.*
*Soit $C_n^* : S \to \mathbb{R}_+$ la série de fonctions définie comme : $C_n^*(s) = 0$ si $P_\infty^*(s) = 0$, sinon si $P_\infty^*(s) > 0$ :*

$$C_n^*(s) = \min_{a \in app(s): \sum_{s' \in S} T(s,a,s') P_\infty^*(s') = P_\infty^*(s)} \frac{1}{P_\infty^*(s)} \sum_{s' \in S} T(s,a,s') P_\infty^*(s') \left[ c(s,a,s') + C_{n-1}^*(s') \right],$$

$$\text{avec : } C_o^*(s) = 0, \forall s \in S \quad (6)$$

*Les fonctions $C_n^*$ convergent vers des fonctions $C_\infty^*$ de valeurs finies et toute politique stationnaire $\pi^*$ obtenue à partir de l'équation précédente à convergence est optimale pour le critère S$^3$P.*

La preuve de ce théorème établit également que la vitesse de convergence des fonctions de probabilité de but et de coût de but optimales dépend de la constante de contraction, qui est égale au rayon spectral de $T^{\pi}_{|\mathcal{X}^{\pi^*}}$ (voir le lemme 1). Pour une précision de convergence $\epsilon > 0$ donnée, les équations 5 et 6 convergent en temps fini, et la complexité temporelle pire-cas de ce schéma itératif est polynomial en le nombre d'états et d'actions, comme les équations de Bellman pour les SSP standards.

Nous avons implémenté le schéma d'optimisation du théorème 3 au sein d'un algorithme nommé `GPCI` (Goal-Probability and -Cost Iteration), dont le pseudo-code est présenté dans l'algorithme 1. Cet algorithme est partagé en trois phases : calcul de la fonction de probabilité de but optimale (lignes 1 à 11) ; calcul de la fonction de coût optimale (lignes 12 à 21) ; extraction de la politique optimale (lignes 22 à 25). Notons que la précision du calcul des fonctions de probabilité de but et de coût de but n'est pas réellement contrôlée. En effet, contrairement au critère $\gamma$-pondéré, et comme le critère total standard, la distance entre deux fonctions successives (lignes 11 ou 21) ne suffit pas pour contrôler la distance à la fonction optimale. Il faudrait pour cela calculer la politique courante $\pi$ et le rayon spectral associé de $T^{\pi}_{|\mathcal{X}^{\pi}}$ (ce qui n'est pas nécessairement très coûteux), ou analyser finement la structure du MDP, comme dans Hansen (2011) où ce problème de distance à l'optimum se pose de toute façon de fait dans les SSP standards. Notons finalement que la politique optimale $\pi$ renvoyée par l'algorithme `GPCI` n'est calculée que sur l'ensemble des états $s \in S$ tels que $P_\infty^*(s) > 0$. En effet, si $\epsilon$ est choisie suffisamment petit, l'exécution de $\pi$ depuis l'état initial ne devrait pas atteindre d'états où $P_\infty^*$ est nul. Comme la version présentée de `GPCI` ne contrôle pas réellement $\epsilon$, il est sans doute plus prudent d'initialiser $\pi$ à une valeur choisie au hasard parmi les actions applicables dans chaque état.

# 5 Évaluation expérimentale

Le but de cette section est de prouver expérimentalement que des politiques qui minimisent le critère total classiquement utilisé dans les SSP, ne sont pas nécessairement des politiques S$^3$P optimales pour des problèmes où il n'existe pas de politiques atteignant le but avec probabilité 1 (politiques correctes). Afin de vérifier cette hypothèse, nous évaluons les fonctions de probabilité de but et de coût de but pour des politiques optimisées sur la base du critère total standard ; cette évaluation est réalisée en utilisant les équations du théorème 1 jusqu'à convergence (prouvée dans le théorème 2). Nous comparons alors les fonctions de probabilité de but et de coût de but pour un état initial donné avec celles optimisées par notre algorithme `GPCI` (qui implémente le théorème 3).

Nous avons testé deux algorithmes optimaux pour le critère total utilisé dans les SSP : `VI` (voir Puterman (1994)), qui a la même complexité temporelle que `GPCI`, et `LRTDP` de Bonet & Geffner (2003), qui est un algorithme de recherche heuristique populaire. Nous nous comparons également à un algorithme non optimal mais efficace : `RFF` de Teichteil-Königsbuch *et al.* (2010), qui tente à la fois de maximiser la fonction de probabilité de but et de minimiser la fonction de coût de but sans garanties théoriques.

**Algorithme 1:** Goal-Probability and -Cost Iteration (GPCI)

$n \leftarrow 0$;
**for** $s \in S$ **do**
  **if** $s \in G$ **then**
    $P_0(s) \leftarrow 1$;
  **else**
    $P_0(s) \leftarrow 0$;
**repeat**
  **for** $s \in S$ **do**
    $P_{n+1}(s) \leftarrow \max_{a \in A} \sum_{s' \in S} T(s,a,s') P_n(s')$;
  $n \leftarrow n+1$;
**until** $\|P_n - P_{n-1}\| < \epsilon$;
$m \leftarrow 0$;
**for** $s \in S$ **do**
  $C_0(s) \leftarrow 0$;
**repeat**
  **for** $s \in S$ **do**
    **if** $P_n(s) > 0$ **then**
      $Aselect \leftarrow \{a \in app(s) : |P_n(s) - \sum_{s' \in S} T(s,a,s') P_n(s')| < \epsilon\}$;
      $C_{m+1}(s) \leftarrow \min_{a \in Aselect} \frac{1}{P_n(s)} \sum_{s' \in S} T(s,a,s') P_n(s') \left[c(s,a,s') + C_m(s')\right]$;
  $m \leftarrow m+1$;
**until** $\|C_m - C_{m-1}\| < \epsilon$;
**for** $s \in S$ **do**
  **if** $P_n(s) > 0$ **then**
    $Aselect \leftarrow \{a \in app(s) : |P_n(s) - \sum_{s' \in S} T(s,a,s') P_n(s')| < \epsilon\}$;
    $\pi(s) \leftarrow \text{argmin}_{a \in Aselect} \frac{1}{P_n(s)} \sum_{s' \in S} T(s,a,s') P_n(s') \left[c(s,a,s') + C_m(s')\right]$;
**return** $\pi$, $P_n$, $C_m$

Il est intéressant de mentionner que certains problèmes difficiles, notamment à cause de leur structure probabiliste, permettent à GPCI d'être plus efficace que tous les autres algorithmes – y compris les algorithmes heuristiques –, en plus d'être le seul dont il est prouvé qu'il est $S^3P$ optimal. Notons que les compétitions de planification internationales (IPC) passées ont en partie classée les différents planificateurs sur la base de leurs fonctions de probabilité de but et de coût de but (Younes *et al.* (2005)), dont nous sommes les premiers à notre connaissance à proposer des moyens de calcul théoriques et pratiques. Nous pensons également que l'optimisation duale des fonctions de probabilité de but et de coût de but est particulièrement intéressante dans de nombreuses applications réalistes, où il n'existe a priori pas de politique atteignant le but avec probabilité 1.

Nos résultats expérimentaux sont résumés dans la figure 3 pour des domaines SSP variés, détaillés ci-dessous. Pour chaque domaine, nous présentons des résultats obtenus pour un seul problème particulier, car (i) nous avons obtenu exactement les mêmes résultats relatifs pour tous les problèmes de chaque domaine qui ont pu être résolu par tous les algorithmes à la fois, et (ii) le but de ces tests n'est bien entendu pas de comparer combien de problèmes peuvent être résolus par chaque algorithme. Le problème de plus grande taille a $2^{47}$ états. Pour VI et GPCI, nous utilisons la connaissance de l'état initial pour supprimer au préalable tous les états qui ne sont pas atteignables depuis l'état initial, quelque soit la politique exécutée.

## 5.1 Domaines blocksworld et rectangle-tireworld domains

Ces domaines sont issus de la compétition internationale de planification. Pour ces deux domaines, il existe en fait des politiques correctes, si bien que les politiques optimisées en utilisant le critère SSP ont des valeurs de coût moyen cumulé finies et sont donc mathématiquement bien définies. En d'autres termes, nous pouvons lancer sans risque les algorithmes VI et LRTDP sans facteur d'actualisation, en utilisant simple-

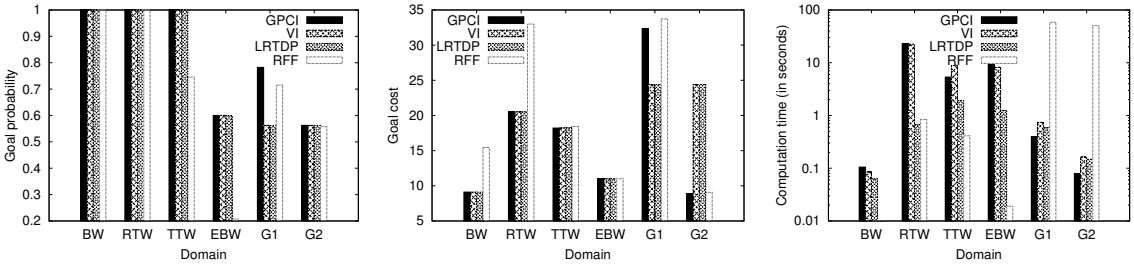


FIGURE 3 – Comparaison des probabilités de but (graphique de gauche), des coûts de but (graphique au centre), et des temps de calcul (graphique de droite), d'algorithmes différents pour des domaines SSP variés : `blocksworld` (BW), `rectangle-tireworld` (RTW), `triangle-tireworld` (TTW), `exploding-blocksworld` (EBW), `grid-I` (G1), `grid-II` (G2).

ment le critère total utilisé avec les SSP (Bertsekas & Tsitsiklis (1996)). Comme attendu, nous constatons que `GPCI`, `VI` et `LRTDP`, trouvent tous trois une politique qui atteint le but avec probabilité 1, et avec la même fonction de coût de but (cf. figure 3). `RFF` est efficace en termes de temps de calcul, mais sa fonction de coût de but est loin d'être optimale.

## 5.2 Domaines triangle-tireworld and exploding-blocksworld domains

Ces domaines sont également issus de la compétition internationale de planification, mais contrairement aux deux domaines précédents, il n'existe pas de politique correcte (pour triangle-tireworld, la probabilité de but maximum est très légèrement inférieure à 1). Ainsi, le critère total utilisé dans les SSP n'est mathématiquement pas bien défini, de telle sorte que `VI` et `LRTDP` ne convergent jamais (testé lors de nos expérimentations mais non présenté dans cet article). Le seul autre critère existant en horizon infini qui puisse être utilisé avec ces algorithmes de sorte qu'ils convergent *sans changer la structure de coût des problèmes*, est le critère $\gamma$-pondéré, qui pondère tous les coûts cumulés futurs par un facteur d'actualisation $0 < \gamma < 1$ (cf. Puterman (1994); Teichteil-Königsbuch *et al.* (2011)). Dans nos expérimentations, pour triangle-tireworld, nous avons pu atteindre des valeurs respectivement maximum et minimum des fonctions de probabilité de but et de coût de but à l'aide de `VI` et `LRTDP` avec le critère $\gamma$-pondéré, avec tout $\gamma \geqslant 0.95$. Pour exploding-blocksworld, nous avons dû prendre $\gamma \geqslant 0.3$. Comme le montre la figure 3, `VI` et `LRTDP` sont capables de trouver les mêmes politiques $S^3P$ optimales que `GPCI` : en effet, dans ces domaines, il existe une constante $\alpha > 0$ telle que tous les états n'atteignant pas le but ont le même coût moyen cumulé $\alpha/(1-\gamma)$, au point qu'il existe une valeur minimum de $\gamma$ garantissant que la minimisation de la fonction de coût (sur tous les chemins, y compris ceux ne menant pas au but) favorisera implicitement les politiques qui maximisent aussi la fonction de probabilité de but. La fonction de coût de but sera alors nécessairement (et implicitement) optimale, car tous les états qui ne mènent pas au but (c'est-à-dire ceux appartenant à des chemins non connectés au but) ont le même coût moyen cumulé pondéré, qui est donc neutre vis-à-vis de l'optimisation de la fonction de coût de but. Toutefois, notons que *la valeur minimum du facteur d'actualisation $\gamma$ garantissant l'optimalité implicite du critère $S^3P$ n'est pas connue à l'avance*. Concernant `RFF`, il n'est pas $S^3P$ optimal puisqu'il ne trouve même pas des politiques de probabilité de but optimale.

## 5.3 Domaines grid

Les expérimentations précédentes ont montré que la structure de coût des domaines IPC est trop simple pour exhiber des situations où (i) il n'existe pas de politique correcte au point que critère total utilisé dans les SSP serait inutilisable, et (ii) le critère $\gamma$-pondéré ne peut pas fournir de politiques $S^3P$ optimale pour toute valeur possible de $\gamma$. Aussi, nous proposons domaine "grid" présenté dans la figure 4, dont nous donnons deux variantes `grid-I` et `grid-II`. Un agent doit se déplacer depuis un état initial vers un état but en utilisant 5 actions disponibles qui coûtent 1 chacune : `up`, `down`, `right`, `left`, `stay`. Dans la variante `grid-I`, toutes les portes peuvent fermer avec probabilité $0.25$ et ensuite ne jamais ré-ouvrir ; quand les portes D1 et D2 (resp. D3 et D4) ferment, un coût additionnel de $1$ (resp. $3$) est ajouté à tous les déplacement unitaires futurs, si bien que les états "cul-de-sac" induisent des coûts différents en fonction de la provenance de l'agent. Clairement, il existe dans ce domaine un compromis difficile entre maximiser la

probabilité de but (qui nécessite de choisir le chemin moins risqué et plus direct à travers les portes D3 et D4), et minimiser les coûts cumulés (qui nécessiterait de traverser les portes D1 et D2 en prévision de la fermeture éventuelle des portes). Un tel compromis est un écueil pour les approches existantes qui reposent exclusivement sur le seul critère du coût moyen cumulé, qu'elles minimisent sur l'ensemble des chemins partant de l'état initial (y compris ceux qui n'atteignent jamais le but). Dans nos expériences, nous avons dû utiliser $\gamma \geqslant 0.99$ pour obtenir des valeurs de probabilité de but maximum et de coût de but minimum, mais la figure 3 montre que ni `VI` ni `LRTDP` ne sont capables de trouver des politiques de probabilité de but optimales, contrairement à `GPCI`. Même `RFF` (qui n'est a priori optimal pour aucun critère) trouve des politiques de probabilité de but meilleures que `VI` et `LRTDP`. Comme les fonctions de probabilité de but ne sont pas optimales, les fonctions de coût de but ne sont pas comparables, puisque moyennées sur des chemins dont la probabilité d'atteindre le but n'est pas optimale.

Dans la variante `grid-II`, les portes ne peuvent pas fermer, mais elles font disparaître le but (il n'est plus atteignable) avec une probabilité de $0.25$ lorsque chaque porte est traversée. Des coûts additionnels sont payés par l'agent pour chaque déplacement futur, comme dans la première variante du domaine. Dans la deuxième variante, les chemins à travers les portes D2 et D1, ou D3 et D4, ont la même probabilité d'atteindre le but. La figure 3 montre que `VI` et `LRTDP` (et aussi `RFF`) trouvent bien des politiques de fonctions de probabilité de but optimales, mais de fonctions de coût de but nettement plus élevées que celles des politiques obtenues avec `GPCI`. La raison en est que `VI` et `LRTDP` prennent en compte tous les chemins atteignables lors de la minimisation du coût, même ceux qui n'atteignent pas le but : ces chemins n'atteignant pas le but ont un coût inférieur si l'agent traverse les portes D1 et D2 en faisant disparaître le but. Toutefois, les chemins à travers les portes D3 et D4 ont bien la même probabilité d'atteindre le but mais un coût inférieur tant que le but n'a pas disparu. De plus, la figure 3 montre également que *`GPCI` présente le temps de calcul le plus faible pour les domaines "grid"*, parce qu'il ne perd pas de temps à optimiser les coûts des chemins qui n'atteignent pas le but (ce que même `LRTDP` et `RFF` perdent du temps à faire).

# 6 Conclusion

À notre connaissance, nous proposons le premier cadre mathématique et algorithmique pour résoudre les MDP orientés but, qui optimise à la fois la probabilité d'atteindre le but, qui ne nécessite pas d'être égale à 1, *et* les coûts cumulés et moyennés *uniquement* sur les chemins qui atteignent le but. Ces métriques sont souvent utilisées pour évaluer des planificateurs ou les performances de différentes politiques. Nous avons expérimentalement démontré que les critères total ou $\gamma$-pondéré, traditionnellement utilisés dans les MDP orientés but, ne trouvent pas nécessairement des politiques optimales pour ces deux métriques duales contrairement à notre approche, en particulier pour les problèmes avec une structure de coût complexe.

La prochaine étape consistera à concevoir des algorithmes de recherche heuristique efficaces pour ces métriques, sur la base des briques théoriques présentées dans cet article. Il nous faudra également chercher des heuristiques indépendantes du domaine de planification, qui sur-estiment la fonction de probabilité de but et sous-estiment celle de coût de but. Nous pensons qu'une telle approche est prometteuse, car l'algorithme optimal et relativement simple que nous avons proposé, `GPCI`, surclasse déjà en termes de

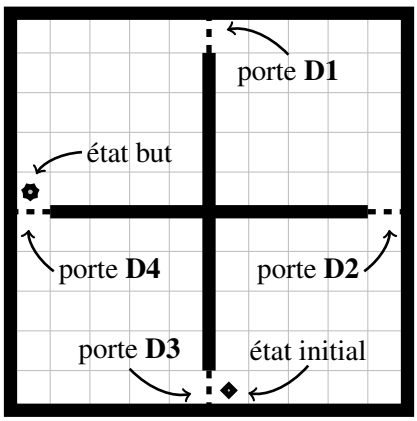


FIGURE 4 – Domaine "grid"

temps de calcul des algorithmes heuristiques pour les SSP comme `LRTDP`, pour des domaines SSP avec une structure de coût non triviale.

# Références


BERTSEKAS D. P. & TSITSIKLIS J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

BONET B. & GEFFNER H. (2003). Labeled RTDP : Improving the convergence of real-time dynamic programming. In E. GIUNCHIGLIA, N. MUSCETTOLA & D. NAU, Eds., *Proc. 13th International Conf. on Automated Planning and Scheduling*, p. 12–21, Trento, Italy : AAAI Press.

BONET B. & GEFFNER H. (2005). mGPT : A probabilistic planner based on heuristic search. *JAIR*, **24**, 933–944.

DAWEN R. (1986). Finite state dynamic programming with the total reward criterion. *Zeitschrift für Operations Research*, **30**(1), A1–A14.

HANSEN E. A. (2011). Suboptimality bounds for stochastic shortest path problems. In *UAI*, p. 301–310.

KOLOBOV A., MAUSAM & WELD D. S. (2010). SixthSense : Fast and Reliable Recognition of Dead Ends in MDPs. In *AAAI*.

KOLOBOV A., MAUSAM, WELD D. S. & GEFFNER H. (2011). Heuristic Search for Generalized Stochastic Shortest Path MDPs. In *ICAPS*.

PUTERMAN M. L. (1994). *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. New York, NY, USA : John Wiley & Sons, Inc., 1st edition.

TEICHTEIL-KÖNIGSBUCH F., KUTER U. & INFANTES G. (2010). Incremental plan aggregation for generating policies in MDPs. In *Proc. AAMAS*, p. 1231–1238.

TEICHTEIL-KÖNIGSBUCH F., VIDAL V. & INFANTES G. (2011). Extending classical planning heuristics to probabilistic planning with dead-ends. In *AAAI*.

YOON S. W., RUML W., BENTON J. & DO M. B. (2010). Improving determinization in hindsight for on-line probabilistic planning. In *ICAPS*, p. 209–217.

YOUNES H. L. S., LITTMAN M. L., WEISSMAN D. & ASMUTH J. (2005). The first probabilistic track of the International Planning Competition. *JAIR*, **24**, 851–887.