Compromis entre sécurité et efficacité : Garanties formelles pour les MDP orientés but

Matisse Roche[1], Yoko Watanabe[1], Caroline P.C. Chanel[2]

Fédération ENAC ISAE-SUPAERO ONERA,
Université de Toulouse, Toulouse, France
Email: {prénom}.{nom}@onera.fr [1], caroline.chanel@isae-supaero.fr [2]

Résumé

Dans les processus de décision markoviens orientés but avec états catastrophiques, nous travaillons dans un cadre de maximisation d'utilité, une fonction qui prend en compte à la fois le coût cumulé et l'atteinte ou non du but. Ce cadre intègre un paramètre K_g représentant le bonus que l'agent reçoit lorsqu'il atteint un état but. Notre contribution principale est une formule analytique qui, pour un seuil de probabilité d'attendre le but P_{th} spécifié par l'utilisateur, détermine la valeur minimale de K_g garantissant que la politique optimale respecte ce seuil de sécurité. Cette transformation du MDP permet ainsi d'utiliser toutes les méthodes classiques et efficaces de résolution tout en bénéficiant automatiquement de la garantie de sécurité, sans recourir à la programmation par contraintes. Les évaluations expérimentales conduites sur le "river problem" démontrent l'avantage de notre approche par rapport aux méthodes traditionnelles qui privilégient exclusivement la sécurité maximale, souvent au détriment d'une efficacité raisonnable.

Mots-clés

MDP orientés but ; Garantie de sécurité ; Plus court chemin stochastique

Abstract

In goal-oriented Markov decision processes with dead-ends, we work within a utility maximization framework. Here, utility is a function that takes into account both the cumulative cost and whether the goal is reached. This framework uses a parameter K_g representing the bonus that the agent receives when it reaches a goal. Our main contribution is to propose an analytical formula that, for a user-specified threshold for a probability of reaching the goal P_{th}, determines the minimum value of K_g guaranteeing that the solution policy respects this safety threshold. This MDP model transformation thus allows the use of classical and efficient resolution algorithms for finding a policy with the safety guarantee, without the need of constraint programming. Experimental evaluations conducted on the "river problem" demonstrate the significant advantage of our approach compared to traditional methods that always prioritize maximum safety, often at the expense of reasonable efficiency.

Keywords

Goal-oriented MDPs ; Safety Guarantee ; Stochastic Shortest Path

1 Introduction

Les problèmes de plus court chemin stochastique (SSP, *Stochastic Shortest Path*) ont été formellement introduits par Bertsekas et Tsitsiklis [1] comme une classe spécifique des Processus de Décision Markoviens (MDP) orientés but. Les SSP-MDP constituent un formalisme puissant pour modéliser et résoudre des problèmes de planification en environnement incertain [6].

Dans de nombreuses applications réelles comme la robotique autonome, la planification de mission ou la gestion de systèmes critiques, ces modèles doivent prendre en compte l'existence d'états catastrophiques – des états à partir desquels il devient impossible d'atteindre le but quelle que soit la politique suivie. Face à ces états catastrophiques, les approches traditionnelles de résolution des problèmes SSP-MDP se révèlent inadaptées [7]. En effet, ces approches reposent sur deux hypothèses fondamentales : l'existence d'une politique propre, politique atteignant le but avec probabilité 1, et l'accumulation d'un coût infini pour toute politique impropre. Ces hypothèses sont rarement satisfaites dans des environnements complexes où les états catastrophiques sont inévitables.

Diverses extensions ont été proposées pour contourner ces limitations. Le cadre SSPUDE (*SSP with unavoidable dead-ends*) de Kolobov (voir [7]) traite les états catastrophiques en leur attribuant soit une pénalité finie (fSSPUDE), soit une pénalité infinie (iSSPUDE). Le cadre S3P de Teichteil-Königsbuch [9] adopte quant à lui une approche lexicographique qui maximise d'abord la probabilité d'atteindre le but, puis minimise le coût espéré parmi les politiques plus sûres. Ces approches partagent cependant un inconvénient majeur : en priorisant systématiquement la sécurité maximale, elles peuvent conduire à des comportements excessivement conservateurs et inefficaces. Dans de nombreux contextes industriels et applications réelles, il est plus pertinent de rechercher un compromis entre sécurité et efficacité. Une autre approche est la recherche de politique assurant un minimum de sécurité, notamment [8], qui introduit le problème *AtLeastProb*. Ce problème consiste à trouver une po-

litique garantissant un seuil de probabilité d'atteinte du but défini par l'utilisateur. Dans cette approche, aucun critère lié au coût ou à l'efficacité n'est pris en compte dans la recherche de la politique optimale.
Alternativement, un nouveau model, basé sur la maximisation de l'utilité a été proposé par [5]. Ce nouveau cadre, appelé le modèle GUBS, permet un contrôle du compromis entre sécurité et efficacité : on peut préférer une politique moins sûre mais la perte de sécurité est limitée. à l'aide d'un paramètre K_g représentant un bonus en cas d'atteinte du but.
Dans cet article on aborde le problème de maximisation de l'efficacité sous la contrainte de sécurité minimale. Inspirés par le modèle GUBS, nous proposons une formule analytique pour déterminer la valeur de K_g qui garantisse que la politique optimale respecte ce seuil de sécurité. Notre approche à l'avantage de ne pas considérer la contrainte lors de l'optimisation. On peut ainsi utiliser les algorithmes classiques de résolution de MDP. Les évaluations expérimentales conduites sur un problème de référence démontrent que notre méthode respecte systématiquement les contraintes de sécurité imposées. De plus, notre méthode est capable de fournir différentes politiques en fonction de la contrainte de sécurité choisie.
Nous présentons d'abord les fondements théoriques des processus de décision markoviens orientés but et les approches existantes pour traiter les états catastrophiques (section 2). Nous développons ensuite notre contribution principale sur le compromis efficacité-sécurité, en établissant une formule analytique qui garantit un seuil minimal de sécurité (section 3). Nous détaillons notre méthode de résolution numérique (section 4) avant de présenter les résultats de nos expérimentations sur le problème de la rivière, qui confirment la validité de notre approche (section 5).

2 Définitions et travaux connexes

2.1 Définitions préliminaires

Les processus de décision markoviens (MDP) [2, 6] constituent un cadre fondamental pour la modélisation et la résolution de problèmes de décision séquentielle en environnement stochastique. Parmi leurs nombreuses variantes, les MDP orientés but représentent une classe particulièrement pertinente pour diverses applications comme la planification de chemin dans le cadre de la robotique autonome.
Formellement, un MDP orienté but est défini comme un tuple $\langle S, A, T, c, \mathcal{G}\rangle$ où :

- S est l'ensemble fini des états ;
- $\mathcal{G} \subseteq S$ est l'ensemble fini des états but ;
- A est l'ensemble fini des actions ;
- $T : S \times A \times S \to [0, 1]$ est la fonction de transition ;
- $c : S \times A \times S \to \mathbb{R}$ est la fonction de coût.

Dans ce cadre, les états but $g \in \mathcal{G}$ possèdent deux propriétés fondamentales :

- Ils sont absorbants : $T(g, a, g) = 1$ pour toute action $a \in A$;
- Ils n'engendrent aucun coût supplémentaire : $c(g, a, g) = 0$ pour toute action $a \in A$.

Une *politique* pour un MDP est une règle de décision qui spécifie quelle action choisir en fonction de l'historique du processus. Formellement, une politique générale peut être définie comme une fonction $\pi = (\pi_0, \pi_1, \ldots)$ où $\pi_t : H_t \to \Delta(A)$ associe à chaque historique $h_t \in H_t$ jusqu'à l'instant t une distribution de probabilité sur les actions. Ici, H_t représente l'ensemble des historiques possibles jusqu'à l'instant t, c'est-à-dire les séquences $h_t = (s_0, a_0, s_1, a_1, \ldots, s_t)$.
Une *politique stationnaire* est une politique qui ne dépend que de l'état courant et non de l'historique complet ou du temps. Elle est définie comme une fonction $\pi : S \to \Delta(A)$ où $\pi(s)$ représente une distribution de probabilité sur les actions pour l'état s. Dans le cas déterministe, une politique stationnaire est simplement une fonction $\pi : S \to A$ qui associe à chaque état une unique action.
Une *trajectoire* θ correspond à un historique $h = \{s_0, a_0, s_1, a_1, \ldots\}$. Nous notons $\Theta_{\mathcal{G}}^{\pi,s,t}$ l'ensemble des trajectoires jusqu'à l'instant t qui atteignent un état but $g \in \mathcal{G}$ lorsque la politique π est appliquée à partir de l'état initial s. t est omis lorsque l'horizon est infini : $\Theta_{\mathcal{G}}^{\pi,s} = \Theta_{\mathcal{G}}^{\pi,s,\infty}$.

2.2 Problèmes de plus court chemin stochastique (SSP)

Les problèmes de plus court chemin stochastique (SSP, *Stochastic Shortest Path*) ont été formellement introduits par Bertsekas et Tsitsiklis [1] comme une classe spécifique de MDP orientés but. Les SSP reposent sur deux hypothèses essentielles :

1. Il existe au moins une politique stationnaire (dite *propre*) qui atteint un état but avec probabilité 1 depuis tout état initial.
2. Toute politique impropre (n'atteignant pas un but avec probabilité 1) accumule un coût infini.

La seconde hypothèse implique généralement que tous les coûts des transitions entre états non-buts sont strictement positifs, ce qui garantit que les cycles infinis sont pénalisés par un coût infini.
Sous ces conditions, résoudre un SSP consiste à trouver une politique π qui minimise l'espérance du coût cumulé jusqu'à l'atteinte d'un état but. Ce critère d'optimisation définit la fonction valeur telle que :

$$V^*(s) = \min_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} c(s_t, \pi(s_t), s_{t+1}) | s_0 = s\right]$$

.
Bertsekas et Tsitsiklis [2] ont montré que, sous les hypothèses SSP, cette fonction valeur est l'unique solution de l'équation de Bellman :

$$V^*(s) = \min_{a \in A} \sum_{s' \in S} T(s, a, s') \left[c(s, a, s') + V^*(s')\right]$$

avec $V^*(g) = 0$ pour tout $g \in \mathcal{G}$.

Des algorithmes classiques de résolution de MDP, tels que *Value Iteration* (VI) ou *Policy Iteration* (PI), peuvent être utilisés pour résoudre des SSP. D'autres algorithmes, plus efficaces que VI ou PI, ont été développés pour résoudre les SSP. Citons notamment *Labelled Real-Time Dynamic Programming* (LRTDP) [3], une variante de VI qui explore l'espace d'états de manière plus efficace en se concentrant sur les trajectoires les plus pertinentes.

2.3 Extensions pour les problèmes avec états catastrophiques

Dans de nombreux contextes pratiques, l'hypothèse selon laquelle un état but est toujours atteignable avec probabilité 1 n'est pas réaliste. Certains environnements comportent des états catastrophiques (ou "dead-ends") – des états à partir desquels il est impossible d'atteindre un état but, quelle que soit la politique suivie.

2.3.1 Approches par facteur d'actualisation

Une première approche pour traiter les états catastrophiques consiste à introduire un facteur d'actualisation $0 < \gamma < 1$ dans la fonction valeur :

$$V_\gamma^*(s) = \min_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t, s_{t+1}) | s_0 = s \right]$$

.

Ce facteur garantit la convergence de la somme, même en présence de trajectoires infinies. Cependant, cette approche présente plusieurs inconvénients :

- Elle modifie la structure du problème original en dévalorisant les coûts futurs.
- Elle ne distingue pas les trajectoires qui atteignent le but de celles qui aboutissent à des états catastrophiques.
- Le choix de γ influence la politique obtenue sans critère clair pour sa détermination.

2.3.2 Cadre SSPUDE

Kolobov et al. [7] ont proposé une extension plus formelle appelée SSPUDE (*Stochastic Shortest Path with Unavoidable Dead-Ends*). Ce cadre distingue deux variantes :

- **fSSPUDE** (Finite-penalty SSPUDE) : Attribue une pénalité finie D à chaque état catastrophique, ce qui permet d'utiliser des algorithmes standards comme VI ou LRTDP.
- **iSSPUDE** (Infinite-penalty SSPUDE) : Attribue une pénalité infinie aux états catastrophiques, formalisant l'idée qu'ils doivent être évités à tout prix. Cette variante est plus complexe algorithmiquement car elle nécessite une approche en deux phases : d'abord identifier les états à éviter, puis optimiser les coûts sur les états restants.

L'avantage principal de fSSPUDE est sa facilité d'implémentation, tandis qu'iSSPUDE offre des garanties théoriques plus fortes sur l'évitement des états catastrophiques. fSSPUDE est computationnellement beaucoup plus efficace qu'iSSPUDE. Pour une valeur suffisamment grande de la pénalité D, fSSPUDE produit des politiques équivalentes à celles d'iSSPUDE. Malheureusement, il n'existe pas de méthode analytique pour déterminer cette valeur critique de pénalité, ce qui rend le paramétrage de fSSPUDE délicat en pratique.

2.3.3 Le cadre S3P

Teichteil-Königsbuch [9] a introduit le cadre S3P (*Stochastic Safest and Shortest Path*) qui adopte explicitement une approche bicritère pour les problèmes avec états catastrophiques. Ce cadre définit un ordre lexicographique sur les politiques :

1. Maximiser d'abord la probabilité d'atteindre un état but : $P^\pi(s) = Pr(\theta \in \Theta_\mathcal{G}^{\pi,s})$
2. Parmi les politiques maximisant cette probabilité, minimiser l'espérance du coût conditionnel des trajectoires réussies :

$$C^\pi(s) = \mathbb{E}[C(\theta) | \theta \in \Theta_\mathcal{G}^{\pi,s}]$$

.

Formellement, une politique optimale π_{S3P}^* pour S3P est définie par :

$$\pi_{S3P}^*(s) \in \arg \min_{\pi \in \arg\max_{\pi'} P^{\pi'}(s)} C^\pi(s)$$

.

Cette approche présente plusieurs avantages :

- Elle est bien définie même en présence d'états catastrophiques inévitables ;
- Elle se concentre uniquement sur les coûts des trajectoires qui atteignent effectivement le but ;
- Elle généralise le cadre SSP standard (lorsque des politiques propres existent, les solutions S3P sont identiques aux solutions SSP).

Dans tous ces cadres, on fait généralement l'hypothèse que les coûts des transitions entre états non-buts sont strictement positifs. Cette hypothèse permet d'éviter les problèmes de cycles de coût nul ou négatif, qui compliqueraient considérablement l'analyse théorique et les algorithmes de résolution.

2.3.4 Le cadre GUBS

Le cadre GUBS, proposé par Freire et Delgado [5], évalue une trajectoire θ en fonction à la fois de son coût cumulé et de l'atteinte ou non d'un état but. Plus précisément, l'utilité d'une trajectoire est définie par :

$$U(\theta) = u(C(\theta)) + K_g \cdot \mathbf{1}_{\theta \in \Theta_\mathcal{G}} \quad (1)$$

où :

- u est une fonction d'utilité sur les coûts satisfaisant les deux propriétés suivants :
  1. $u : \mathbb{R} \cup \{\infty\} \to [U_{min}, U_{max}]$, et
  2. u est strictement décroissante ;
- $C(\theta)$ est le coût cumulé de la trajectoire θ ;

- K_g est un paramètre représentant le bonus d'utilité accordé pour l'atteinte d'un état but ;
- $\mathbf{1}_{\theta \in \Theta_{\mathcal{G}}}$ est l'indicateur d'atteinte du but.

Dans le cadre GUBS, l'évaluation d'une politique π se fait par la fonction de valeur définie comme l'espérance de l'utilité sur toutes les trajectoires possibles :

$$V^{\pi}_{GUBS}(s) = \mathbb{E}\left[U\left(C(\theta), \mathbf{1}_{\theta \in \Theta_{\mathcal{G}}}\right) \mid \theta \in \Theta^{\pi,s}\right] \quad (2)$$

Cette formulation peut être développée comme suit :

$$V^{\pi}(s) = \left(U^{\pi}(s) + K_g\right) P^{\pi}(s) + \left(1 - P^{\pi}(s)\right) U_{\min} \quad (3)$$

où

$$U^{\pi}(s) = \mathbb{E}\left[u(C(\theta)) \mid \theta \in \Theta^{\pi,s}_{\mathcal{G}}\right] \quad (4)$$

La politique optimale π^*_{GUBS} pour GUBS est celle qui maximise la fonction de valeur (2.3.4). Pour garantir que le cadre GUBS priorise l'atteinte du but, il faut que $K_g > U_{max} - U_{min}$. Cette condition assure qu'une trajectoire atteignant le but aura toujours une utilité supérieure à une trajectoire qui n'y parvient pas, indépendamment des coûts.

2.3.5 Propriétés fondamentales du cadre GUBS

Pour calculer l'utilité d'un chemin dans le cadre GUBS, il est nécessaire de savoir si ce chemin atteint ou non un état but. Cette particularité entraîne une conséquence importante : la politique optimale pour le critère GUBS n'est généralement pas stationnaire. En effet, l'action optimale à exécuter dans un état dépend non seulement de l'état courant, mais aussi du coût cumulé jusqu'à ce point : $a^*_t = \pi^*_{GUBS}(s_t, C(\theta_t))$. Cela présente des défis computationnels significatifs, car cette dépendance au coût cumulé revient à créer un nouveau MDP en augmentant l'espace d'états avec la variable C, ce qui peut engendrer un nombre potentiellement infini d'états, tant en termes de calcul que de mémoire nécessaire pour stocker la politique. La résolution n'est d'ailleurs possible que si les coûts sont des rationnels. Néanmoins, Crispino et al. [4] ont démontré un résultat important : il existe une valeur C_{max} calculable à partir des données du problème, telle que pour tout état s et tout coût cumulé $C \geq C_{max}$, la politique optimale sous GUBS devient stationnaire. De plus, cette politique stationnaire correspond exactement à la politique optimale selon le critère lexicographique risque-sensible, comme la politique optimale S3P. Formellement, pour deux politiques π et π', on a $\pi \succ \pi'$ (c'est-à-dire que π est préférée à π') si :

- $P^{\pi}(s) > P^{\pi'}(s)$, ou
- $P^{\pi}(s) = P^{\pi'}(s)$ et $U^{\pi}(s) > U^{\pi'}(s)$

Lorsque la fonction d'utilité choisie est de type exponentielle, on parle alors de eGUBS (exponential GUBS). Pour résoudre ce problème particulier, Crispino et al. [4] ont proposé l'algorithme eGUBS-VI une approche qui adapte les méthodes de programmation dynamique classiques pour résoudre un MDP avec leurs critères.

2.4 Discussion sur ces approches

L'introduction d'états catastrophiques inévitables dans le modèle soulève une question fondamentale : comment évaluer et comparer les différentes politiques ? En effet, dans ce contexte, deux critères entrent en compétition :

- La sécurité : maximiser la probabilité d'atteindre un état but
- L'efficacité : minimiser le coût moyen des trajectoires atteignant un état but

Plusieurs approches ont été proposées dans la littérature pour traiter ce dilemme. La plus courante, utilisée dans S3P [9] ou iSSPUDE [7], adopte une préférence lexicographique qui priorise la sécurité maximale, puis minimise le coût parmi les politiques maximalement sûres. Cette approche présente cependant un inconvénient majeur : elle peut conduire à des politiques extrêmement inefficaces en termes de coût pour obtenir un gain marginal en probabilité. Par exemple, pour augmenter la probabilité d'atteindre sa destination de 95% à 96%, un voyageur pourrait devoir tripler son temps de trajet, ce qui est rarement acceptable en pratique.
De plus, dans de nombreux contextes industriels et applications réelles, il n'est pas toujours nécessaire ou même souhaitable de maximiser absolument la probabilité de succès. On recherche plutôt des politiques qui garantissent un niveau de sécurité suffisant (au-dessus d'un certain seuil) tout en optimisant l'efficacité. D'autres travaux ont exploré cette direction, notamment [8], qui introduit le problème AtLeast-Prob. Ce problème consiste à trouver une politique garantissant un seuil de probabilité d'atteinte du but défini par l'utilisateur $P_{th} \in [0, 1]$, soit formellement une politique π telle que $P^{\pi}(s) \geq P_{th}$ (ou prouver qu'une telle politique n'existe pas). Bien que cette formulation permette de fixer un niveau de sécurité minimal acceptable, elle ne traite pas de l'aspect efficacité. En effet, aucun critère lié au coût ou à l'efficacité n'est pris en compte dans la recherche de la politique optimale.

3 Le compromis Efficacité-Sécurité

3.1 MDP orienté but sous la contrainte de sécurité

Dans ce contexte, nous proposons une approche alternative qui permet de spécifier un niveau minimal de sécurité souhaité tout en optimisant l'efficacité. Formellement, nous cherchons à résoudre le problème d'optimisation suivant :

$$\min_{\pi \in \Pi} \mathbb{E}[C(\theta) | \theta \in \Theta^{\pi,s}_{\mathcal{G}}]$$
$$\text{t.q.} \quad P^{\pi}(s) \geq P_{th}$$

où P_{th} est un seuil de probabilité spécifié par l'utilisateur. Nous appelons ce seuil, le seuil de sécurité.

3.2 Garantie de sécurité pour GUBS

Pour obtenir une politique efficace, qui respecte le seuil de sécurité, nous nous appuyons sur le cadre théorique GUBS [5], présenté dans la section 2.3.4.

Nous proposons une méthode permettant de garantir qu'une politique optimale sous GUBS, π^*_{GUBS}, atteigne une probabilité de succès minimale P_{th} dans un MDP orienté but, en ajustant le bonus K_g. Pour cela, nous considérons les deux politiques extrêmes suivantes :

- La politique optimale pour l'ordre lexicographique π^*_L, qui atteint la probabilité maximale $P^*(s) = \max_\pi P^\pi(s)$ d'attendre le but ;
- La politique la plus efficace π^*_E, qui atteint l'espérance conditionnelle maximale $U^{\pi^*_E}(s) = \max_\pi U^\pi(s)$.

Théorème 3.1 (Garantie de sécurité pour GUBS)**.** *Soit un MDP orienté but $M = \langle S, A, T, c, \mathcal{G}\rangle$ et soit $P^*(s_0)$ la probabilité maximale d'atteindre $\mathcal{G}$ depuis l'état initial s_0. Pour tout seuil de sécurité $P_{th} \in]0, P^*(s_0)[$, définissons*

$$K^*_{P_{th}} = \frac{P_{th}(U^{\pi^*_E}(s_0) - U_{min}) - P^*(s_0)(U^{\pi^*_L}(s_0) - U_{min})}{P^*(s_0) - P_{th}}, \tag{5}$$

avec π^*_L la politique optimale pour l'ordre lexicographique et π^*_E pour la politique la plus efficace. $U^\pi(s_0)$ est l'espérance conditionnelle de l'utilité du coût cumulé pour les trajectoires atteignant un état but, défini par (2.3.4).
Alors, pour tout $K_g > K^*_{P_{th}}$, toute politique optimale selon le critère GUBS avec paramètre K_g possède une probabilité d'atteindre un état but supérieure ou égale à P_{th}.

Démonstration. Nous voulons démontrer que, pour $K_g \geq K^*_{P_{th}}$, aucune politique de proba de succès inférieure à P_{th} ne peut avoir une utilité moyenne supérieure à celle de la politique optimale pour l'ordre lexicographique.
La formule suivante définie $U^{\pi^*_E}(s)$:

$$U^{\pi^*_E}(s) = \max_\pi \lim_{T\to\infty} \mathbb{E}[u(C(\theta)) \mid \theta \in \Theta^{\pi,s,T}_{\mathcal{G}}]$$

Soit une politique π telle que $P^\pi(s_0) < P_{th}$, on majore l'utilité moyenne d'une telle politique :

$$\begin{aligned} V^\pi(s_0) &= \mathbb{E}\left[u(C(\theta)) + K_g \mathbf{1}_{\theta\in\Theta^{\pi,s_0}_{\mathcal{G}}} \mid \theta \in \Theta^{\pi,s_0}\right] \\ &= P^\pi(s_0)\left(\mathbb{E}[u(C(\theta)) \mid \theta \in \Theta^{\pi,s_0}_{\mathcal{G}}] + K_g\right) \\ &\quad + (1 - P^\pi(s_0))U_{min} \\ &= P^\pi(s_0)\left(\mathbb{E}[u(C(\theta)) \mid \theta \in \Theta^{\pi,s_0}_{\mathcal{G}}]\right. \\ &\quad \left. + K_g - U_{min}\right) + U_{min} \\ &< P_{th}(U^{\pi^*_E}(s_0) + K_g - U_{min}) + U_{min} \end{aligned}$$

D'autre part, pour π_L la politique optimale pour l'ordre lexicographique, on a :

$$V^{\pi^*_L}(s_0) = P^*(s_0)(U^{\pi^*_L}(s_0) + K_g - U_{min}) + U_{min}$$

Si l'on suppose que $V^{\pi^*_L}(s_0) \leq V^\pi(s_0)$, alors :

$$\begin{aligned} &P^*(s_0)(U^{\pi^*_L}(s_0) + K_g - U_{min}) + U_{min} \\ &\quad \leq P_{th}(U^{\pi^*_E}(s_0) + K_g - U_{min}) + U_{min} \end{aligned}$$

Ce qui implique :

$$\begin{aligned} K_g(P^*(s_0) - P_{th}) \leq\ & P_{th}(U^{\pi^*_E}(s_0) - U_{min}) \\ & - P^*(s_0)(U^{\pi^*_L}(s_0) - U_{min}) \end{aligned}$$

Finalement :

$$K_g \leq \frac{P_{th}(U^{\pi^*_E}(s_0) - U_{min}) - P^*_G(s_0)(U^{\pi^*_L} - U_{min})}{P^*_G(s_0) - P_{th}}$$

Par contraposée, si $K_g > K^*_{P_{th}}$, avec $K^*_{P_{th}}$ définit par l'équation (5), alors la politique π^*_L domine toute politique de probabilité de succès strictement inférieure à P_{th}. La politique optimale a donc une probabilité de succès supérieure à P_{th}, ce qui constitue la garantie de sécurité recherchée. □

3.3 Garantie de sécurité pour GUBS sans la politique la plus efficace

Dans l'expression précédente, le terme $U^{\pi^*_E}(s_0)$ n'est pas accessible sans résoudre le MDP en maximisant $U^\pi(s_0)$. Si on a accès à un majorant de $U^{\pi^*_E}(s_0)$ on peut l'utiliser à la place du terme $U^{\pi^*_E}(s_0)$ dans la formule précédente et la contrainte de sécurité sera toujours garantie. On peut obtenir un tel majorant en considérant le graphe de transition du MDP.

Théorème 3.2 (Garantie de sécurité pour GUBS sans π^*_E)**.** *Soit un MDP orienté but $M = \langle S, A, T, c, \mathcal{G}\rangle$ et soit $P^*(s_0)$ la probabilité maximale d'atteindre $\mathcal{G}$ depuis l'état initial s_0. Pour tout seuil de sécurité $P_{th} \in]0, P^*(s_0)[$, définissons*

$$K^+_{P_{th}} = \frac{P_{th}(u(d_G(s_0,\mathcal{G}) - U_{min}) - P^*(s_0)(U^{\pi^*_L}(s_0) - U_{min})}{P^*(s_0) - P_{th}}, \tag{6}$$

avec π^*_L la politique optimale pour l'ordre lexicographique et $d_G(s_0,\mathcal{G}) = \min_{\theta\in\Theta_{\mathcal{G}}} C(\theta)$.
Alors, pour tout $K > K^+_{P_{th}}$, toute politique optimale selon le critère GUBS avec paramètre K possède une probabilité d'atteindre un état but supérieure ou égale à P_{th}.

Démonstration. D'après le théorème 3.1, il suffit de montrer que $K^*_{P_{th}} \leq K^+_{P_{th}}$.
Par hypothèse u est décroissante, donc pour tout π et tout $\theta \in \Theta^\pi_{\mathcal{G}}$, $u(\theta) \leq u(d_G(s_0,\mathcal{G}))$, donc pour tout π, $U^\pi(s) \leq u(d_G(s_0,\mathcal{G}))$, d'où $U^{\pi^*_E}(s_0) \leq u(d_G(s_0,\mathcal{G}))$. On a donc bien $K^*_{P_{th}} \leq K^+_{P_{th}}$. □

4 Résolution numérique

Dans cette section, nous présentons l'algorithme eGUBS-VI, proposé par Freire et Delgado [4] qui calcule de manière efficace la fonction de valeur optimale et la politique optimale associée pour un paramètre K_g donné. Notre approche consiste à choisir une valeur du paramètre K_g de sorte à ce que la contrainte de sécurité soit garantie. Nous proposons également une méthode pour calculer $u(d_G(s_0,\mathcal{G}))$. On peut donc résoudre le MDP, avec la valeur de K_g fournit au théorème 3.2, et obtenir une politique qui respecte la contrainte de sécurité.

Comme Crispino et al [4], nous faisons l'hypothèse que la fonction d'utilité est une exponentielle décroissante :

$$u(C) = e^{\lambda C} \tag{7}$$

avec $\lambda < 0$ un paramètre. Cette hypothèse nous permet d'utiliser l'algorithme de résolution proposé par Crispino et al. Toutefois, notre méthode et notre garantie théorique restent valides quel que soit le choix de la fonction d'utilité.

4.1 Algorithme eGUBS-VI

L'algorithme eGUBS-VI [4] suit les étapes suivantes :

1. **Calcul de la politique optimale lexicographique sensible au risque :** D'abord, calculer la politique optimale pour le critère lexicographique sensible au risque en utilisant l'algorithme RiskSensitiveLexicographicVI, une variante de VI qui procède en deux phases à chaque itération. Pour chaque état s, il calcule d'abord la probabilité maximale d'atteindre l'objectif :

$$P_{\mathcal{G}}(s) \leftarrow \max_{a \in A} \sum_{s' \in S} T(s,a,s') P'_{\mathcal{G}}(s')$$

Ensuite, il définit l'ensemble des actions qui maximisent cette probabilité :

$$A_s^{P_{\mathcal{G}}} = \text{Argmax}_{a \in A} \sum_{s' \in S} T(s,a,s') P_{\mathcal{G}}(s')$$

Enfin, il met à jour la fonction d'utilité en utilisant uniquement ces actions maximisant la probabilité :

$$U_\lambda(s) = \max_{a \in A_s^{P_{\mathcal{G}}}} \left\{ \sum_{s' \in S} T(s,a,s') e^{\lambda c(s,a,s')} U_\lambda(s') \right\}$$

La fonction $U_\lambda(s)$ converge vers $U^{\pi^*_L}(s)$, qui représente l'utilité conditionnelle espérée pour la politique lexicographique sensible au risque π^*_L. De même, $P_{\mathcal{G}}(s)$ converge vers $P^*(s)$, la probabilité maximale d'atteindre un état objectif à partir de s.

2. **Calcul de C_{max} :** Ensuite, déterminer la valeur C_{max} à partir de laquelle la politique optimale devient stationnaire. L'algorithme pour calculer C_{max} est présenté dans [4] page 26.
3. **Calcul de la politique optimale GUBS :** L'algorithme calcule itérativement en remontant de C_{max} à 0 :
   i. **La politique optimale** pour chaque état s et coût C :

$$\pi^*(s,C) = \arg\max_{a \in A} \{Q(s,a,C) + K_g \cdot P'_{\mathcal{G}}(s,a,C)\}$$

   ii. **Les composantes pour évaluer les actions :**

$$Q(s,a,C) = \sum_{s'} T(s,a,s') \cdot e^{\lambda C'} \cdot V^*(s',C')$$

$$P'_{\mathcal{G}}(s,a,C) = \sum_{s'} T(s,a,s') \cdot P_{\mathcal{G}}(s',C')$$

   où $C' = C + c(s,a,s')$ et $s' \in S$.

   iii. **Mise à jour des valeurs optimales :** Après avoir déterminé $a^* = \pi^*(s,C)$, on met à jour :

$$V^*(s,C) = Q(s,a^*,C)$$
$$P_{\mathcal{G}}(s,C) = P'_{\mathcal{G}}(s,a^*,C)$$

   Ces valeurs sont utilisées pour calculer la politique aux étapes précédentes.

4.2 Calcul de $d_G(s_0, \mathcal{G})$

La méthode proposé dans cet article nécessite de calculer la valeur de $K^+_{P_{th}}$ en utilisant l'équation (6) dans le théorème 3.2. L'algorithme e-GUBS VI fournit déjà les termes $P^*(s_0)$ et $U^{\pi^*_L}(s_0)$. Nous expliquons ici comment calculer rapidement le dernier terme manquant, $d_G(s_0, \mathcal{G})$.

Définition 4.1 (Graphe déterministe d'un MDP)**.** Soit $M = \langle S, A, T, c \rangle$ un processus décisionnel de Markov. Le graphe déterministe associé à M est un graphe orienté pondéré $G = (V, E, w)$ où :

- $V = S$ est l'ensemble des sommets correspondant aux états du MDP ;
- $E \subseteq S \times S$ est l'ensemble des arêtes tel que $(s,s') \in E$ si et seulement s'il existe une action $a \in A$ telle que $T(s,a,s') > 0$;
- $w : E \to \mathbb{R}^+$ est la fonction de pondération des arêtes définie par :

$$w(s,s') = \min_{a \in A : T(s,a,s') > 0} c(s,a,s')$$

Dans ce graphe, le poids d'une arête représente le coût minimal pour transiter d'un état à un autre, à condition qu'une telle transition soit possible avec une probabilité strictement positive dans le MDP orienté but original.
On note que $d_G(s_0, \mathcal{G})$ est égal au coût du plus court chemin de s_0 vers un état but dans le graphe déterministe G.
Pour calculer $d_G(s_0, \mathcal{G})$, nous utilisons l'algorithme de Dijkstra, qui détermine efficacement le plus court chemin dans un graphe à pondérations positives.

5 Évaluations expérimentales

Pour évaluer notre approche, nous avons utilisé l'environnement "River Problem" introduit par Freire et Delgado [5] dans le cadre de leurs travaux sur GUBS. Ce banc d'essai constitue un cas d'étude permettant de tester notre méthode et de visualiser les différents compromis entre sécurité et efficacité.

5.1 L'environnement *River Problem*

Le problème de la rivière (*River Problem*), illustré dans la figure 1), considère une carte sous forme de grille de dimensions $N_x \times N_y$, où les extrémités en coordonnée x (soit $x = 1$ et $x = N_x$) représentent les berges de la rivière. L'agent doit traverser la rivière, ce qui peut être réalisé de deux manières :

- en nageant à partir de n'importe quel point de la berge, ou

- en longeant la berge jusqu'à atteindre un pont situé à $y = N_y$.

Cependant, la rivière s'écoule vers une chute d'eau (située à $y = 1$), où l'agent peut se retrouver piégé. Ces états correspondent à des états catastrophiques dans notre modèle.
L'état initial se trouve d'un côté de la rivière et loin du pont : $x_0 = 1$ et $y_0 = 2$. L'état but se situe de l'autre côté de la rivière, loin du pont : $x_g = N_x$ et $y_g = 1$.
Les actions peuvent être prises dans chacune des directions cardinales : Nord (N), Sud (S), Est (E) et Ouest (W). Si les actions sont effectuées sur la berge ou sur le pont, les transitions sont déterministes vers les directions cardinales correspondantes. Chaque action engendre un coût de 1. En revanche, si les actions sont effectuées dans la rivière, les transitions deviennent probabilistes : l'agent suit la direction cardinale choisie avec une probabilité $(1 - P)$, ou est emporté vers le bas (Sud) de la rivière avec une probabilité P. La chute d'eau est modélisée comme un ensemble d'états catastrophiques.

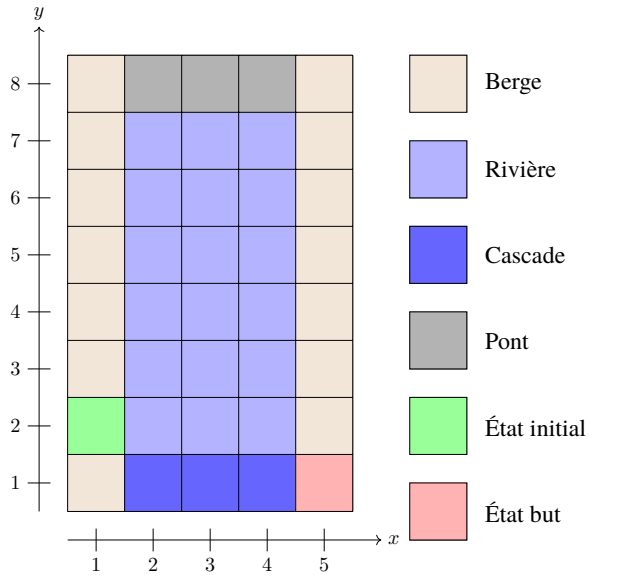

FIGURE 1 – Représentation de l'environnement "river problem" pour les paramètres $N_x = 5$ et $N_y = 8$.

5.2 Configuration expérimentale

Pour nos expériences, nous avons fixé les paramètres suivants : $N_x = 5$, $N_y = 15$ et $P = 0.4$. Le coût de chaque action est constant et égal à 1. On choisit $\lambda = -0.1$ comme constante pour l'utilité exponentielle 4.
Afin de simplifier l'analyse et la visualisation des politiques, nous avons légèrement modifié l'environnement original en n'autorisant que l'action d'aller vers l'Est (E) lorsque l'agent se trouve dans la rivière. Cette modification présente plusieurs avantages :

- Une politique peut être résumée essentiellement par le moment où l'agent décide d'entrer dans l'eau (la coordonnée y d'entrée dans la rivière).
- Cette configuration permet de dériver une formule analytique exacte pour calculer la probabilité d'atteindre le but en fonction de P et de la coordonnée y d'entrée dans la rivière, ce qui nous permet de comparer notre méthode numérique avec un oracle. Ici l'oracle est donc, pour un niveau de sécurité P_{th} donné, la politique la plus efficace, de proba de succès supérieure à P_{th}.

Cette simplification permet de représenter visuellement les différentes politiques par un seul paramètre : la hauteur à laquelle l'agent décide de traverser. Elle met également en évidence le compromis fondamental de ce problème : plus l'agent monte au Nord avant de traverser, plus il augmente sa sécurité (probabilité d'atteindre le but), mais en contrepartie, le coût du trajet augmente également.

5.3 Protocole expérimental

Notre protocole expérimental a consisté à :

1. Choisir plusieurs valeurs de seuil de sécurité P_{th}.
2. Pour chaque valeur de P_{th}, calculer la valeur minimale de $K^+_{P_{th}}$ en utilisant notre formule analytique présentée dans le théorème 3.2.
3. Résoudre le problème GUBS avec cette valeur de $K_g = K^+_{P_{th}}$ à l'aide de l'algorithme eGUBS-VI [4].
4. Calculer la probabilité exacte d'atteindre le but pour chaque politique, en exploitant les caractéristiques de cet environnement qui permettent de déterminer cette probabilité en fonction de la coordonnée d'entrée dans la rivière.
5. Déterminer la politique oracle pour comparer et évaluer les performances de notre approche par rapport à cette référence optimale.

5.4 Résultats et analyse

Les résultats de notre méthode sont présentés dans plusieurs graphiques qui illustrent différents aspects du compromis entre sécurité et efficacité.
La figure 2 illustre l'évolution de la politique optimale en fonction du seuil de sécurité P_{th}. L'axe des ordonnées représente la coordonnée y à laquelle l'agent décide d'entrer dans la rivière selon la politique calculée.
Comme prévu, plus le seuil de sécurité P_{th} augmente, plus la politique optimale choisit de traverser la rivière à une altitude élevée, renforçant ainsi la sécurité du parcours.
La figure 3 compare la probabilité de succès de notre méthode avec celle de la politique oracle. Notez que ces probabilités sont des valeurs exactes calculées analytiquement, et non des résultats empiriques d'exécutions répétées. Nous constatons que notre méthode respecte systématiquement la contrainte de sécurité, tous les points se situant au-dessus de la droite d'équation $y = x$. Cela confirme que la probabilité d'atteindre le but pour un P_{th} donné est toujours supérieure ou égale à P_{th} avec notre approche.
Néanmoins, comme le montre la Figure 4, qui représente l'utilité du coût cumulé en fonction de P_{th}, notre méthode produit des politiques plus conservatrices que l'oracle. La

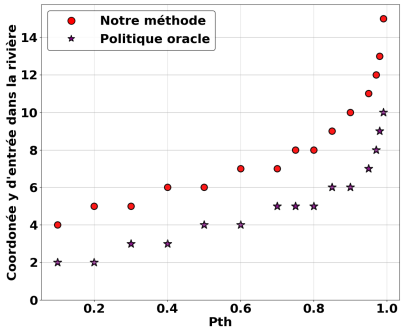

FIGURE 2 – Coordonnée d'entrée dans la rivière en fonction du seuil de sécurité P_{th}

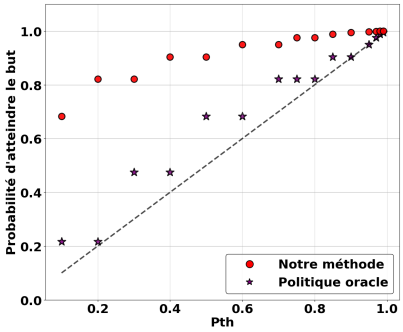

FIGURE 3 – Probabilité de succès observée en fonction du seuil de sécurité P_{th}.

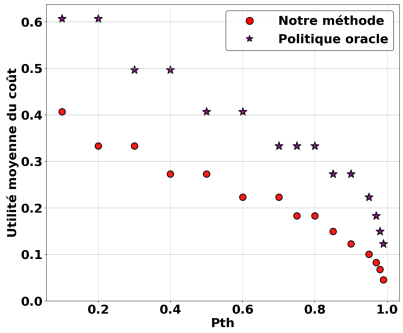

FIGURE 4 – Utilité du coût cumulé total en fonction du seuil de sécurité P_{th}.

politique oracle présente systématiquement une meilleure utilité pour un même seuil de sécurité. Cette observation suggère que notre approche, bien que garantissant formellement le respect du seuil de sécurité, tend à être un peu trop prudente dans certains cas. Des raffinements de notre formule analytique pourraient permettre de réduire cet écart dans de futurs travaux.

La Figure 2 révèle une discontinuité significative dans la politique optimale : l'agent ne sélectionne jamais l'altitude $y = 14$ comme point d'entrée dans la rivière. Cette discontinuité dans l'espace des politiques optimales s'explique avec la théorie de GUBS.

Pour comprendre ce phénomène, la Figure 5 présente l'évolution de C_{max} en fonction de P_{th}. Dans la théorie GUBS,

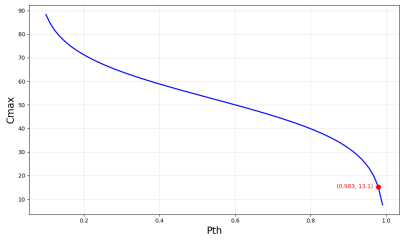

FIGURE 5 – Évolution de C_{max} en fonction du seuil de sécurité P_{th}.

pour un coût cumulé supérieur à C_{max}, la politique optimale correspond à celle de l'ordre lexicographique. On observe que pour $P_{th} \approx 0.98$, C_{max} chute en dessous de 12. Cette valeur est cruciale car elle représente exactement le coût accumulé par l'agent lorsqu'il se trouve encore sur la berge à la position $y = 14$. Ainsi, lorsque l'agent atteint cette position avec un P_{th} supérieur à 0.98, il bascule automatiquement vers la politique la plus sûre (la politique lexicographique) et continue donc son chemin jusqu'au pont situé à $y = 15$, sans jamais traverser la rivière à $y = 14$. Ce phénomène explique l'absence de politique d'entrée à $y = 14$ dans la Figure 2 et illustre comment la théorie GUBS influence directement le comportement de l'agent en fonction du seuil de sécurité imposé.

6 Conclusion et perspectives

Dans cet article, nous avons abordé la problématique du compromis entre sécurité et efficacité dans les processus de décision markoviens orientés but avec états catastrophiques. Nous avons constaté qu'il n'existe pas de critère d'optimalité canonique dans ce cadre, car nous sommes confrontés à une optimisation multi-objectif où sécurité et efficacité sont souvent en opposition.

Notre contribution principale réside dans une formule analytique qui détermine une valeur du bonus d'utilité K_g garantissant qu'une politique optimale dans le cadre GUBS respecte un seuil de probabilité d'atteindre le but spécifié par l'utilisateur. Cette approche, à notre connaissance, constitue la première tentative formelle d'obtenir des garanties de sécurité tout en optimisant l'efficacité pour les MDP orientés but avec états catastrophiques.

Les évaluations expérimentales conduites sur le problème de la rivière démontrent que notre méthode respecte systématiquement les contraintes de sécurité imposées. Cependant, comme l'illustrent nos résultats, notre approche actuelle tend à être conservatrice, produisant des politiques dont la probabilité de succès dépasse souvent le seuil requis, parfois au détriment de l'efficacité optimale.

Dans nos travaux futurs, nous envisageons principalement d'affiner notre méthode en développant des majorations plus précises pour l'estimation du paramètre K_g. Cette perspective d'amélioration est essentielle car elle permettrait de réduire le caractère conservateur de notre approche actuelle tout en maintenant les garanties formelles de sécurité qui constituent l'avantage principal de notre méthode. En paral-

lèle, nous prévoyons d'étendre nos expérimentations à des MDP orientés but de plus grande dimension afin d'évaluer la scalabilité de notre approche. Nous envisageons également de réaliser une analyse comparative entre notre méthode et les approches existantes de la littérature.
En conclusion, bien que notre approche ne fournisse pas encore la politique optimale exacte pour le critère de sécurité minimale que nous avons défini, elle offre néanmoins un cadre formel solide et une méthode pratique pour obtenir des politiques respectant des garanties de sécurité tout en maintenant une efficacité raisonnable, représentant ainsi une avancée significative par rapport aux approches existantes.

Références

[1] Dimitri P. Bertsekas and John N. Tsitsiklis. *Analysis of Stochastic Shortest Path Problems*, volume 16. INFORMS, 1991.

[2] Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3) :580–595, 1991.

[3] Blai Bonet and Hector Geffner. Labeled RTDP : Improving the convergence of real-time dynamic programming. In *Proceedings of the 13th International Conference on Automated Planning and Scheduling (ICAPS)*, pages 12–21. AAAI Press, 2003.

[4] G. N. Crispino, V. Freire, and K. V. Delgado. Gubs criterion : Arbitrary trade-offs between cost and probability-to-goal in stochastic planning based on expected utility theory. *Artificial Intelligence*, 316 :103848, 2023.

[5] V. Freire and K. V. Delgado. GUBS : A utility-based semantic for goal-directed Markov decision processes. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1260–1268. IFAAMAS, 2017.

[6] Andrey Kolobov et al. *Planning with Markov decision processes : An AI perspective*, volume 17. Morgan & Claypool Publishers, 2012.

[7] Andrey Kolobov, Mausam, and Daniel S. Weld. A theory of goal-oriented mdps with dead ends. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 438–447, 2012.

[8] Marcel Steinmetz, Jörg Hoffmann, and Olivier Buffet. Goal probability analysis in probabilistic planning : Exploring and enhancing the state of the art. *Journal of Artificial Intelligence Research*, 57 :229–271, 2016.

[9] Florent Teichteil-Königsbuch. Stochastic safest and shortest path problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1825–1831, 2012.