Sorbonne Université
Master ANDROIDE

Cavalier à marche aléatoire

Projet Madi

Julien Canitrot – Alan Adamiak

2023

Table des matières

Chapitre 1

Recherche d'une trajectoire prudente

1.1 Question A. Modélisation du problème

Modéliser le problème comme un processus décisionnel Markovien en donnant une récompense significative lorsqu'on atteint la case but, par exemple un bonus (réduction de coût) de 1000 points mais on pourra essayer d'autres valeurs en fonction de la dimension de la grille choisie. Écrire les équations de Bellman définissant la valeur vij d'être dans une case (i, j) pour une politique stationnaire optimale et un facteur d'actualisation γ.

Nous devons modéliser le problème sous la forme d'un processus décisionnel markovien. Un PDM est un quadruplet {S, A, T, R}, définissant :

— un ensemble d'états S,

— un ensemble d'actions A,

— une fonction de transition $T : S \times A \times S \to [0; 1]$,

— une fonction de récompense $R : S \times A \times S \times \mathbb{R} \to [0; 1]$

Dans notre cas, cette PDM correspond à :

— $S : \{Re \cup Bla \cup Ve \cup N \cup Ble \cup Ro\}$; où chaque acronyme correspond à un type d'état dans lequel le cavalier peut se retrouver. [1]

— $A : \{R, T, Y, U, J, H, G, F\}$; correspondant à l'ensemble des actions d'un cavalier dans le jeu d'échec.

— $T : (i, j) \times a \times (i', j') \to [0; 1]$, $\forall (i, j) \in S \forall a \in A \forall (i', j') \in S$; correspondant à l'ensemble des actions possibles dans l'échiquier crée.

— $R : \{\ R((i, j)) = 0, \quad \forall (i, j) \in Mu;$
$R((i, j)) = -1, \quad \forall (i, j) \in Bla;$
$R((i, j)) = -2, \quad \forall (i, j) \in Ve;$
$R((i, j)) = -4, \quad \forall (i, j) \in N;$
$R((i, j)) = -8, \quad \forall (i, j) \in Ro;$
$R((i, j)) = -16, \quad \forall (i, j) \in Ble;$
$R((i, j)) = 1000, \quad \forall (i, j) \in Re;\ \}$, où chaque type de case a une récompense spécifique associée.

1. état représenté sous forme de case de type Re : Récompense, Bla : Blanc, Ve : Vert, N : Noir, Ble : Bleu, Ro : Rouge

Équations de Bellman

$V_t(i,j) = \max_{a \in A} R((i,j),a) + \gamma \sum_{(i',j') \in S} T((i,j),a,(i',j'))V_{t-1}(i',j')$

Cette première équation est simplifiable, car la récompense de l'état ne dépend pas de notre action.

$V_t(i,j) = \max_{a \in A} R((i,j)) + \gamma \sum_{(i',j') \in S} T((i,j),a,(i',j'))V_{t-1}(i',j')$

1.2 Question B Le cas déterministe

Dans un premier temps, on se place dans le cas déterministe (la case cible est atteinte avec une probabilité 1 à chaque saut). Déterminer par itération de la valeur un chemin optimal de la case initiale vers la case but et tester sur une grille de taille puis 5×10, 10×15, 15×20 (on essaiera avec $\gamma = 1$ et $\gamma = 0.5$). On donnera les grilles en visualisant les trajectoires optimales et on donnera les temps de calcul.

Toutes ces expériences ont été réalisées avec epsilon = 1 (critère d'arret). De plus, les seeds sont identiques entre les expériences afin de pouvoir observer des variations dans le comportement.

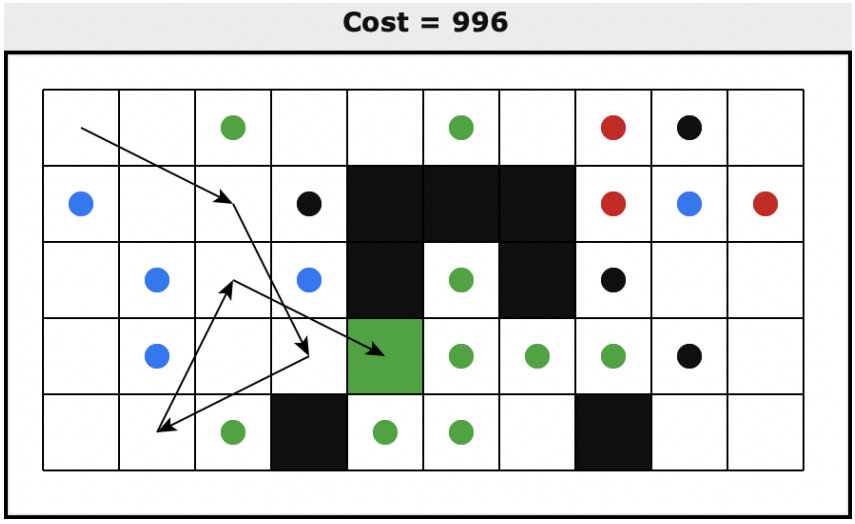

FIGURE 1.1 – nb_lines = 5, nb_columns = 10, gamma = 1, 5 itérations, Temps de calcul : 0.00097 secondes

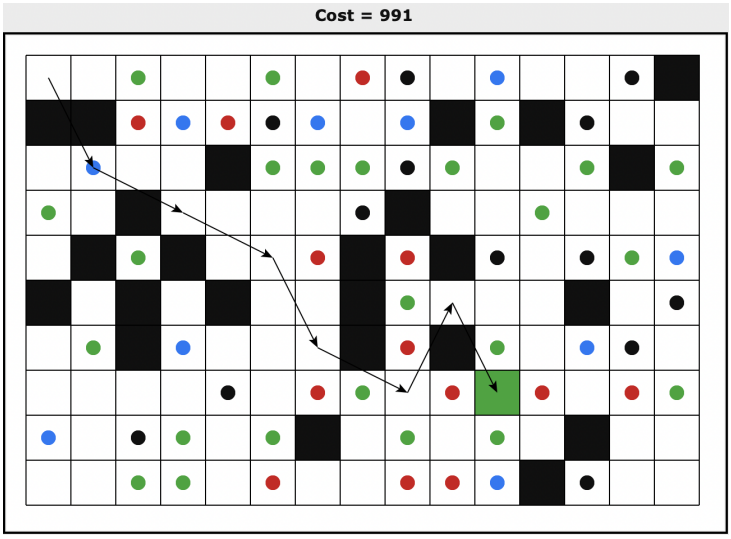

FIGURE 1.2 – nb_lines = 10, nb_columns = 15, gamma = 1, 7 itérations, Temps de calcul : 0.00525 secondes

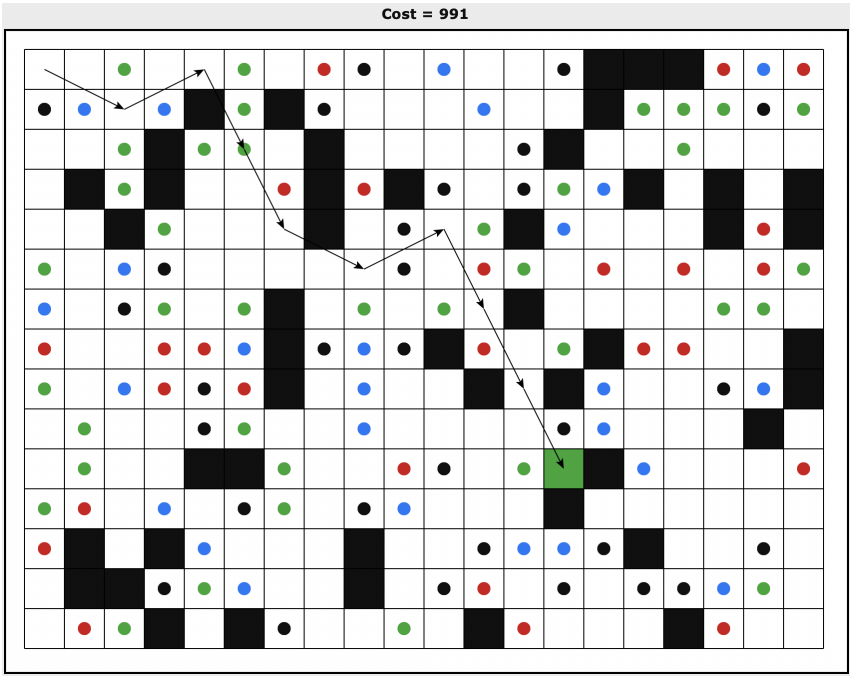

FIGURE 1.3 – nb_lines = 15, nb_columns = 20, gamma = 1, 8 itérations, Temps de calcul : 0.00930 secondes

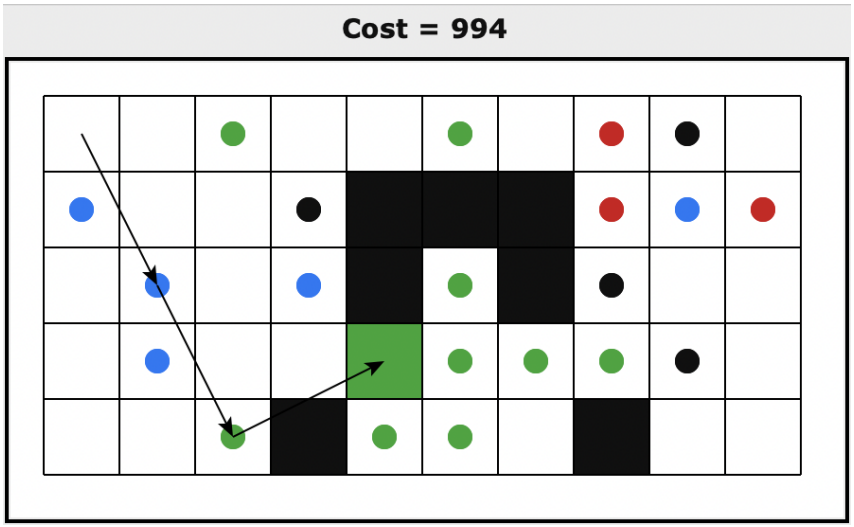

FIGURE 1.4 – nb_lines = 5, nb_columns = 10, gamma = 0.5, 5 itérations, Temps de calcul : 0.00078 secondes

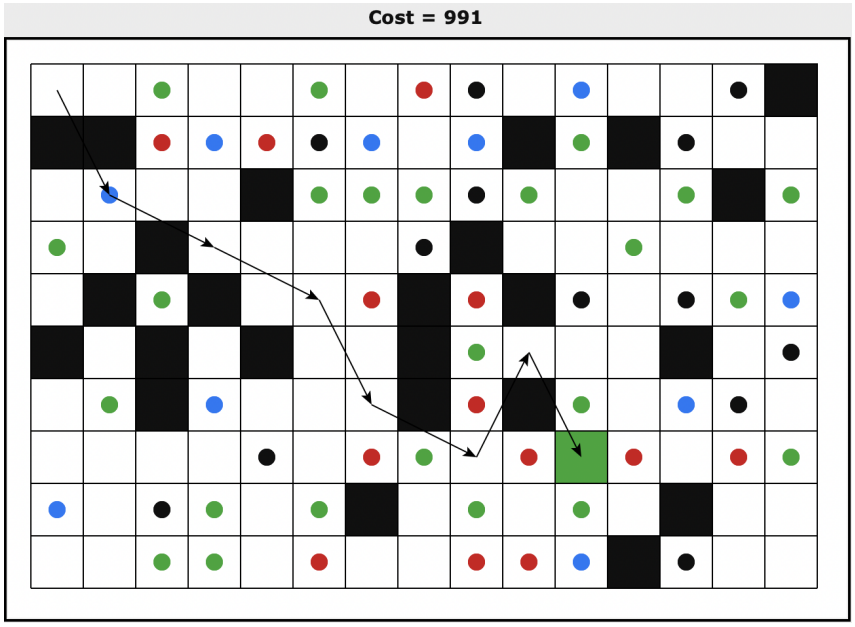

FIGURE 1.5 – nb_lines = 10, nb_columns = 15, gamma = 0.5, 6 itérations, Temps de calcul : 0.00323 secondes

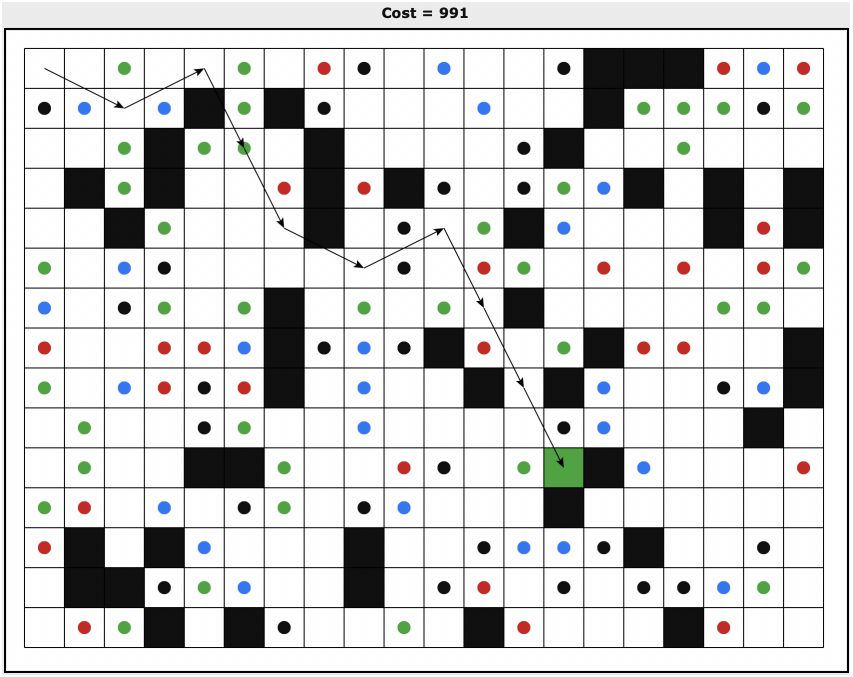

FIGURE 1.6 – nb_lines = 15, nb_columns = 20, gamma = 0.5, 8 itérations, Temps de calcul : 0.00865 secondes

1.3 Question C Transitions aléatoires

On se place maintenant dans le cas de transitions aléatoires définies comme expliqué ci-dessus. On souhaite déterminer une politique optimale par itération de la valeur et effectuer des essais numériques de résolution de grilles de différentes tailles. Pour chaque taille on tirera 10 instances de grilles aléatoirement et on donnera dans un tableau le temps moyen de résolution et le nombre moyen d'itérations. On pourra également étudier l'impact de γ en faisant varier $\gamma = 0.9$, $\gamma = 0.7$, $\gamma = 0.5$.

Voici le tableau donnant les valeurs moyennes sur 10 exécutions.

	Taille	Itération	Temps (seconde)	std Itération	Std Temps
$\gamma = 0.9$	(5,5)	10.9	0.0028	0.7	0.0006
	(10,10)	11.9	0.0144	1.044	0.0021
	(10,15)	12.9	0.0224	0.8307	0.0021
	(15,25)	15.8	0.0767	1.99	0.0102
	(20,20)	15.9	0.0848	2.5865	0.0149
	(25,50)	22.5	0.4184	5.6436	0.1057
	(40,40)	21.5	0.5093	4.653	0.1008
	(50,50)	22.9	0.8632	3.9862	0.1251
	Taille	Itération	Temps (seconde)	std Itération	Std Temps
$\gamma = 0.7$	(5,5)	7.2	0.0018	0.4	0.0003
	(10,10)	8.1	0.0097	0.9434	0.0012
	(10,15)	8.1	0.0154	1.044	0.0018
	(15,25)	10.4	0.0533	1.8	0.0085
	(20,20)	11.2	0.0625	1.4	0.0072
	(25,50)	13.1	0.2458	2.1656	0.0403
	(40,40)	12.7	0.312	1.9	0.0434
	(50,50)	14.4	0.5647	0.9165	0.032
	Taille	Itération	Temps (seconde)	std Itération	Std Temps
$\gamma = 0.5$	(5,5)	5.3	0.0011	0.4583	0.0002
	(10,10)	6.1	0.0075	0.8307	0.0012
	(10,15)	6.1	0.0122	0.7	0.0021
	(15,25)	7.7	0.0414	0.6403	0.0034
	(20,20)	7.4	0.0438	0.6633	0.0051
	(25,50)	8.	0.1558	0.	0.0033
	(40,40)	7.4	0.1917	1.0198	0.0231
	(50,50)	8.	0.3247	0.	0.0071

Par ailleurs, on observe un résultat intéressant avec les grilles de grande taille où la récompense est suffisamment éloignée quand le gamma est petit. Ainsi, comme on peut le voir sur la figure ci-dessous, la politique n'arrive pas à converger vers la solution optimale.

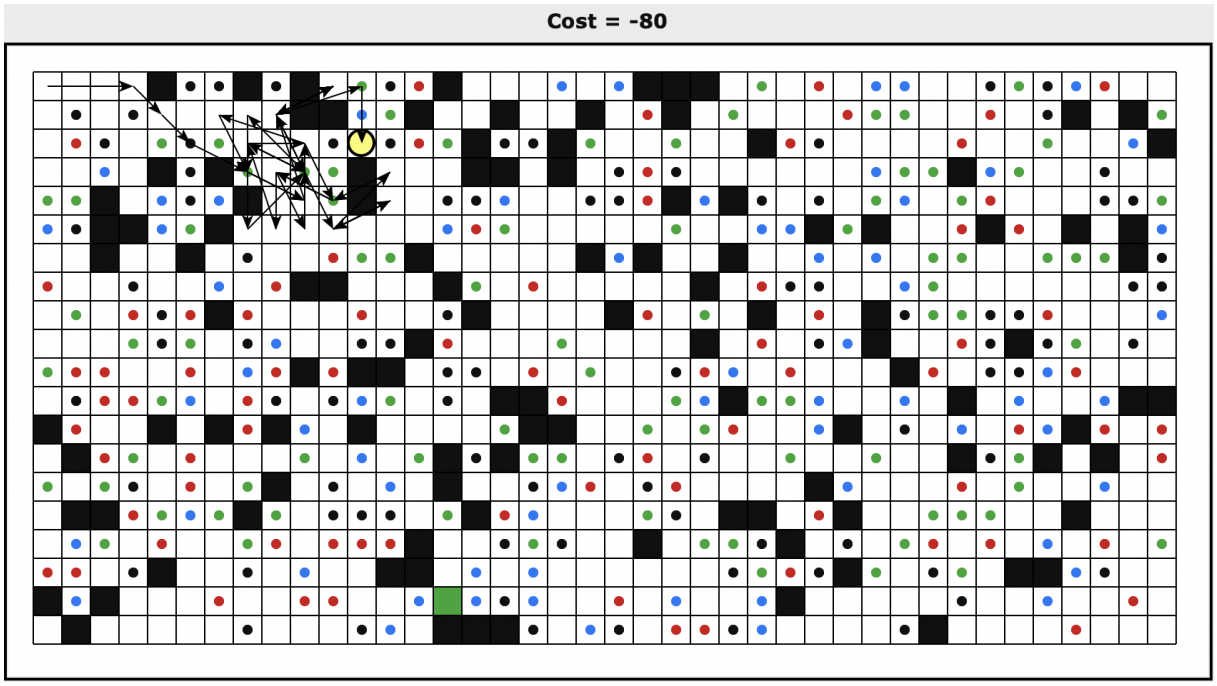

FIGURE 1.7 – nb_lines = 20, nb_columns = 40, gamma = 0.3

En effet, cela est dû au critère d'actualisation γ qui détermine l'importance relative de la récompense future par rapport à la récompense actuelle. Plus le critère d'actualisation est petit, plus la récompense future est peu prise en compte dans le calcul de la valeur de chaque état.

Ainsi, si le critère d'actualisation est trop petit par rapport à la taille de la grille, l'algorithme d'itération de la valeur sera principalement guidé par les récompenses immédiates et ne tiendra pas compte de l'impact à long terme de ses décisions. Cela entraîne ici une politique suboptimale.

Il est important de trouver un compromis entre la récompense immédiate et la récompense à long terme en choisissant un critère d'actualisation approprié. Si le critère d'actualisation est trop petit, la politique obtenue peut être suboptimale, mais si le critère d'actualisation est trop grand, l'algorithme peut mettre trop de temps à converger vers la politique optimale.

1.4 Question D Test de la politique optimale

Pour une grille donnée, on essayera de piloter à la main quelques trajectoires menant à la case but et on comparera leur coût au coût moyen d'une politique mixte optimale observé sur 20 exécutions successives de cette politique sur la même grille. On pourra aussi comparer les coûts moyens observés avec la valeur espérée de la politique mixte optimale testée.

Voici les différentes pertes sur 20 itérations avec la politique optimale :

— 62, 135, 121, 140, 88, 147, 244, 71, 93, 38, 120, 58, 22, 44, 82, 51, 91, 29, 23, 6 ;

— en moyenne cela donne une perte de 83.25 ;

— avec une variance de 55.1

Voici nos performances sur 3 essais :

— 19,38,66

— en moyenne cela donne une perte de 41

On remarque que l'on fait mieux. Cependant la variance est très élevée, ce qui montre que l'aléatoire est déterminant dans les résultats de nos essais. Ainsi, il aurait fallu faire plus d'essais pour obtenir une moyenne plus représentative de la performance de la politique appliquée en tant qu'utilisateur.

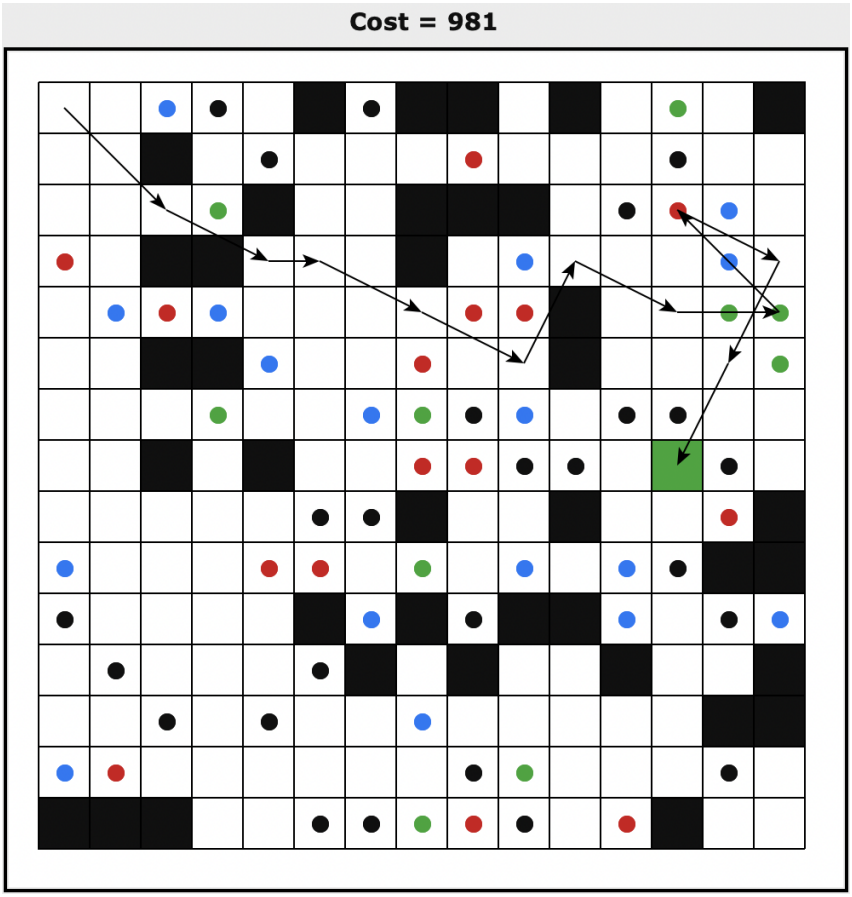

FIGURE 1.8 – nb_lines = 15, nb_columns = 15, exemple de trajectoire réalisée à la main

Chapitre 2

Recherche d'une politique équilibrée

2.1 Question A formalisation du PL

La méthode de l'itération de la valeur ne peut pas être utilisée pour résoudre ce problème, car elle vise à trouver une politique qui minimise l'espérance de coût global à chaque état. Alors que dans notre cas, on cherche à trouver une politique qui minimise le coût maximal attendu sur les deux composantes du coût.

Il existe plusieurs raisons pour lesquelles l'itération de la valeur ne peut pas optimiser un MDP multi-objectif :

L'itération de la valeur utilise une mesure de performance unique pour évaluer chaque état. Cette valeur est optimisée en choisissant la politique qui maximise cette valeur. Dans un MDP multi-objectif, il est difficile de définir une seule valeur qui puisse refléter tous les critères de performance. De plus elle utilise une boucle itérative pour améliorer la valeur d'état en prenant en compte l'ensemble des actions possibles à chaque étape. Dans un MDP multi-objectif, il peut être difficile voir impossible de trouver une action qui améliore simultanément tous les critères de performance.

Objectif

$$\text{maximize } z$$

Contraintes

$$\sum_{a\in A} x_{i,j,a} - \gamma \sum_{(i',j')\in S} \sum_{a\in A} x_{i',j',a} T((i',j'),a,(i,j)) = \frac{1}{|S|}, \forall (i,j) \in S$$

$$z \leq \sum_{(i,j)\in S} \sum_{a\in A} R_0((i,j)) x_{i,j,a}$$

$$z \leq \sum_{(i,j)\in S} \sum_{a\in A} R_1((i,j)) x_{i,j,a}$$

variables

$$x_{i,j,a} \geq 0, \forall a \in A, \forall (i,j) \in S$$

2.2 Question B recherche de trajectoire équilibrée par un PL avec politique mixte

Comme pour la question 1.3, nous sommes dans le cas où les transitions sont aléatoires. Cependant, comme les résultat des PLs sont toujours identiques, nous n'effectuons qu'une itération par couple de paramètres γ et taille. De plus la version étudiante de gurobi ne permet pas de résoudre des grilles de plus de 10x20 à cause d'une limite de nombre de variables.

Voici donc le tableau donnant les valeurs pour une exécution.

	Taille	Itération	Temps (seconde)
$\gamma = 0.9$	(5,5)	89	0.0065
	(10,10)	174	0.0125
	(10,15)	324	0.0112
	(10,20)	365	0.0143
	Taille	Itération	Temps (seconde)
$\gamma = 0.7$	(5,5)	64	0.0051
	(10,10)	158	0.0075
	(10,15)	299	0.0115
	(10,20)	359	0.0141
	Taille	Itération	Temps (seconde)
$\gamma = 0.5$	(5,5)	47	0.0051
	(10,10)	154	0.0079
	(10,15)	268	0.0098
	(10,20)	423	0.0150

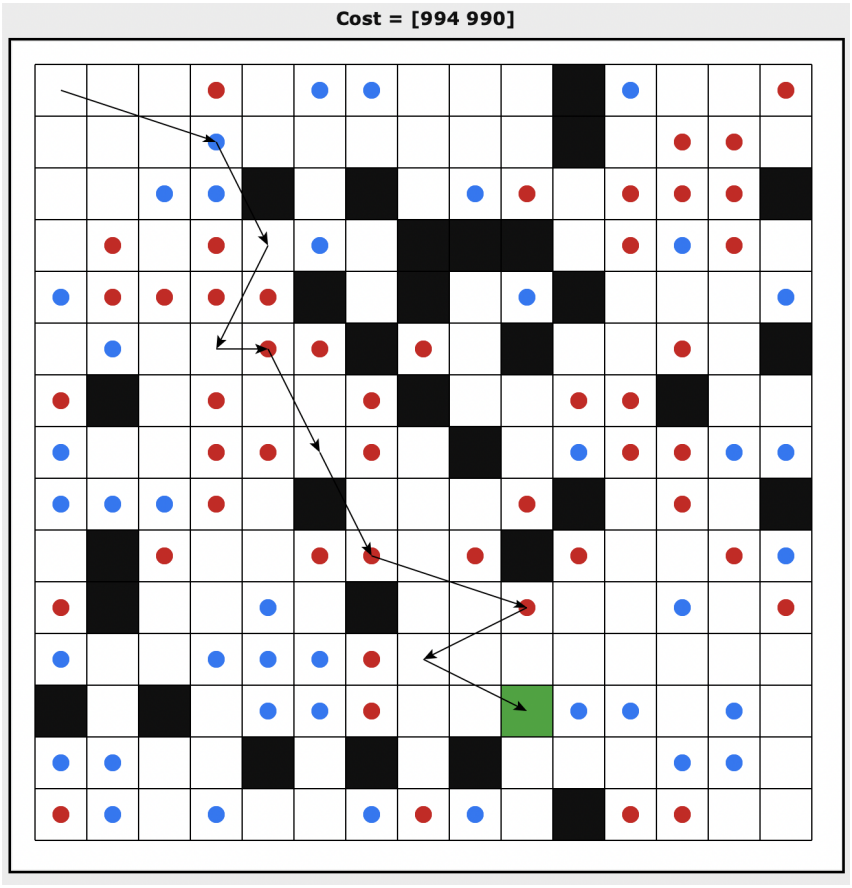

FIGURE 2.1 – nb_lines = 15, nb_columns = 15, exemple de trajectoire réalisée par un PL en politique mixte

Pour une grille donnée de 10x10, on calcule le coût moyen de la politique mixte optimale observée sur 20 exécution.

Voici les coût : (21,21), (5,9), (3,5), (8,4), (3,5), (3,5), (4,2), (6,6), (5,11), (5,9), (11,11), (13,15), (2,4), (3,5), (2,4), (3,3), (8,10), (6,6), (8,5), (2,6), (20, 16)

Ce qui nous donne un coût moyen de (6.71, 7.71)

2.3 Question C Politique Pur

Voici l'actualisation du PL lorsqu'on l'on force la politique pure ; en gras ce qui a été rajouté à l'ancien PL.

Objectif

$$\text{maximize } z$$

Contraintes

$$\sum_{a\in A} x_{i,j,a} - \gamma \sum_{(i',j')\in S} \sum_{a\in A} x_{i',j',a} T((i',j'),a,(i,j)) = \frac{1}{|S|}, \forall (i,j) \in S$$

$$\sum_{(i,j)\in S} \sum_{a\in A} R_0((i,j)) x_{i,j,a} \geq z$$

$$\sum_{(i,j)\in S} \sum_{a\in A} R_1((i,j)) x_{i,j,a} \geq z$$

$$(\mathbf{1}-\gamma)\cdot \mathbf{x_{i,j,a}} \leq \mathbf{d_{i,j,a}} \quad \forall \mathbf{a} \in \mathbf{A}, \forall (\mathbf{i},\mathbf{j}) \in \mathbf{S}$$

$$\sum_{\mathbf{a}\in\mathbf{A}} \mathbf{d_{i,j,a}} \leq \mathbf{1} \quad \forall (\mathbf{i},\mathbf{j}) \in \mathbf{S}$$

variables

$$x_{i,j,a} \geq 0, \forall a \in A, \forall (i,j) \in S$$
$$\mathbf{d_{i,j,a}} \in \{\mathbf{0},\mathbf{1}\} \quad \forall \mathbf{a} \in \mathbf{A}, \forall (\mathbf{i},\mathbf{j}) \in \mathbf{S}$$

L'ajout de nouvelles variables réduit encore la taille maximale des grilles de test, en nous faisant descendre à 10x15

Voici donc le tableau donnant les valeurs pour une exécution du PL en politique pure.

	Taille	Itération	Temps (seconde)
$\gamma = 0.9$	(5,5)	110	0.0438
	(10,10)	648	0.0470
	(10,15)	528	0.0396
	Taille	Itération	Temps (seconde)
$\gamma = 0.7$	(5,5)	108	0.0306
	(10,10)	355	0.0289
	(10,15)	570	0.0384
	Taille	Itération	Temps (seconde)
$\gamma = 0.5$	(5,5)	114	0.0198
	(10,10)	248	0.0251
	(10,15)	781	0.0446

On remarque que la résolution du PL avec une politique pure est 2 à 7 fois plus lent que la politique mixte et qu'il demande plus d'itérations.

Comme pour question précédente, on exécute la politique 20 fois pour avoir l'espérance de coût empirique. Nous obtenons les valeurs suivantes : (7,3), (2,6), (6,2), (3,5), (6,6), (3,5), (10,12), (2,4), (17,11), (3,3), (4,4), (7,9), (3,3), (4,6), (8,8), (6,8), (12,14), (4,4), (9,9), (2,6) Soit un coût moyen de (5.9, 6.4), ce qui est légèrement inférieur à la politique mixte, et avec les 2 coûts plus proches (0.5 de différence contre 1).