CYCLE 5

RESOLUTION DE PROBLEMES PAR UTILISATION DE L'INGENIERIE NUMERIQUE OU L'APPRENTISSAGE AUTOMATISE

TP2 - PSI

TP 2.2

PELLETEUSE – APPRENTISSAGE PAR RENFORCEMENT

CORRECTION

1.1 Actions possibles

Quels sont les axes/vérins liés au positionnement du godet ?

Pour positionner le godet, seuls les vérins de flèche et de pénétration sont nécessaires. Le vérin de cavage ne sert qu'à orienter le godet.

En déduire quelques mouvements élémentaires possibles du godet.

Les mouvements élémentaires du godet sont donc liés à la conception des chaînes géométriques / cinématiques (voir schéma cinématique du bras dans la documentation)

- rotation de centre O1 si mouvement de la flèche uniquement
- rotation de contre O2 si mouvement du balancier uniquement
- mouvement « simple » de translation verticale et horizontale
- mouvement de translation suivant la diagonale

En analysant le travail réalisé par un conducteur de pelleteuse, on se rend compte que le déplacement du godet est toujours réalisé en « jouant » simultanément sur les vérins de flèche et de pénétration. Les actions possibles sont donc limitées aux 5 déplacements dans le plan : Haut, Gauche-Haut, Gauche, Gauche-bas et Bas.

La position d'arrivée ne pouvant pas se situer à droite de la position de départ dans la simulation, le déplacement vers la droite n'est pas proposé.

Justifier, d'un point de vue énergétique, que le déplacement vers la droite ne semble pas cohérent avec la condition de placement des points d'arrivée et de départ.

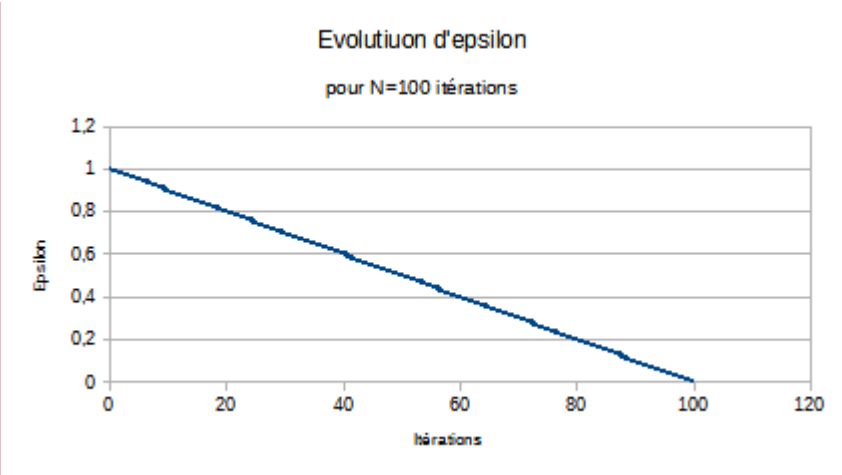
De manière assez logique, le chemin le plus court est le plus rapide, et on pourrait aussi dire le moins énergétique. Et donc, si le point d'arrivée se situe à gauche du point de départ, faire un mouvement vers la droite entraînerait forcément un retour vers la gauche pour atteindre le but, soit un rallongement du chemin et donc une plus grande consommation. Comme la trajectoire optimale doit minimiser l'énergie, il semble donc logique de ne pas prendre en compte les mouvements vers la droite.

Dans chacun des cas « initiaux » présentés sur les figures 3 et 4, proposer les actions possibles à retenir qui vous semblent le plus cohérent.

- obstacle triangulaire : il faut retenir toutes les actions (gauche, haut et bas sont nécessaires)
- obstacle rectangulaire : les mouvements vers le bas ne sont pas nécessaires, on retient donc gauche, haut et gauche-haut.

2 APPRENTISSAGE PAR RENFORCEMENT Q-LEARNING

Après avoir représenté l'évolution du facteur aléatoire en fonction de l'itération courante, proposer une description de l'évolution de l'apprentissage en utilisant les termes EXPLOITATION et EXPLORATION.



Dans les premières itérations, le facteur aléatoire est proche de 1, ainsi, il y a aura de fortes chances pour que le choix de l'action sera aléatoire. À mi parcours, ici pour 50 itérations, le facteur aléatoire vaut 0,5, soit autant de chance d'avoir un choix d'action aléatoire ou déterministe via la matrice Q. Enfin, plus l'indice d'itération augmente plus le facteur aléatoire diminue ce qui rend un choix d'action de moins en moins aléatoire. Ainsi, au début, l'apprentissage est basé sur de l'exploration, à la fin sur de l'exploitation.

Remarque : la méthode epsilon-greedy est implantée dans l'apprentissage et n'est pas modifiable.

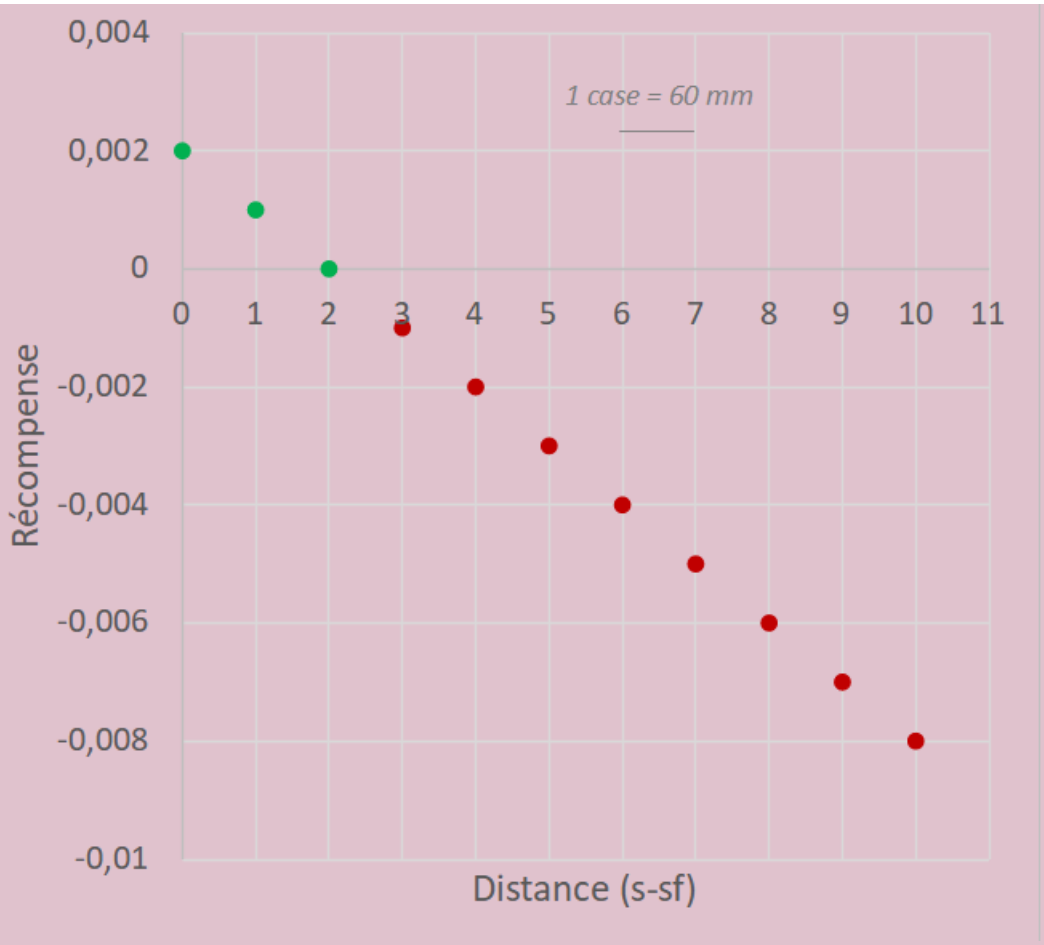
2.1 Récompense

La difficulté dans la méthode d'apprentissage basé sur les récompenses et de « bien » choisir la récompense.
Dans l'algorithme proposé ici, la récompense $\boldsymbol{r}$ est définie par

avec d la distance entre l'état actuel $\boldsymbol{s}$ et l'état final $\boldsymbol{s_f}$. Dans la discrétisation de la zone de travail, chaque case carrée a un côté de longueur 60 mm.
On définit aussi deux valeurs de récompense

- quand l'état $\boldsymbol{s}$ atteint correspond à l'état final
- quand l'état $\boldsymbol{s}$ atteint correspond à un obstacle ou s'il est hors domaine.

Après avoir tracer l'évolution de la récompense en fonction de la distance d, juger de la cohérence de la récompense. Vous pourrez utiliser la notion de malus/bonus pour argumenter.



On remarque que plus l'état est proche du point d'arrivée, plus la récompense sera importante. Toutefois, pour une distance trop grande, la récompense est négative (malus), montrant ainsi que cet état n'est pas forcément cohérent à atteindre pour se diriger vers le point final, d'autant plus que cet état est éloigné du point d'arrivée (d'où un malus plus important).

Remarque : dans la suite de l'étude, la valeur 120 et le coefficient pourront être modifiés pour analyser leur influence.

3 ANALYSE DE L'APPRENTISSAGE

Dans MyViz, réaliser un apprentissage avec tous les paramètres par défaut, pour un obstacle rectangulaire et en choisissant uniquement les actions haut, gauche-haut et gauche comme action possible.

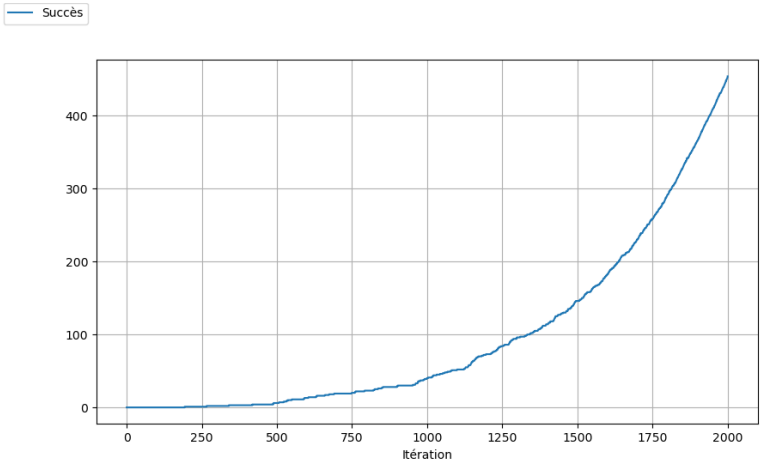
En analysant les différentes étapes du chemin à différentes itérations, déterminer à quoi correspondent les flèches oranges, vertes et violettes. Il est possible de modifier la vitesse de calcul pour bien représenter les différentes étapes.

Les flèches oranges désignent le parcours suivi par l'itération actuelle, les flèches vertes affichent la direction à privilégier dans l'état pour optimiser la récompense (valeur maximale de Q) et les flèches de couleur violette correspondent à la meilleure trajectoire trouvée (celle qui minimise l'énergie)

Observer l'évolution du choix des actions à prendre en fonction de l'avancement dans le calcul.

Au début du calcul, le choix des actions se fait de manière aléatoire puis plus on avance dans les itérations, plus on voit des choix d'action à partir de la matrice Q qui deviennent ensuite majoritaires sur la fin.

Tracer l'évolution du nombre de succès en fonction du nombre d'itération (bouton tracé dans le cadre haut-gauche de MyViz). Commenter.



On observer une augmentation « exponentielle » du nombre de succès avec le nombre d'itération. Cette évolution croissante est due au fait que plus on avance dans le calcul, plus l'exploitation prend le dessus sur l'exploration Vers la fin du calcul, le facteur aléatoire étant tellement nul, que chaque itération mène à un succès.

3.1 *Influence du facteur d'apprentissage*

Le choix de l'action ***a*** à prendre par l'agent qui va le faire passer de l'état ***s*** à l'état ***s'*** dépend de la matrice Q telle que

.

Dans MyViz, réaliser un apprentissage pour la valeur du facteur d'apprentissage α=0.

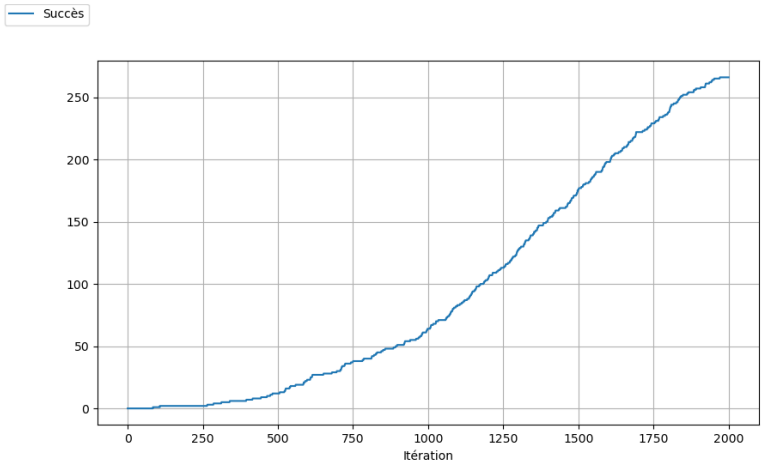
Comment évolue la matrice Q ? Les bonnes trajectoires sont-elles obtenues de manière aléatoire ou grâce à la matrice Q ?

La matrice Q n'évolue pas et ne prend donc jamais en compte les récompenses. Toutes les actions sont donc prises de manière aléatoire, même lorsque l'exploitation domine sur l'exploration.

Pourquoi n'y a-t-il pas de flèche verte symbolisant la meilleure direction à prendre dans chaque état ?

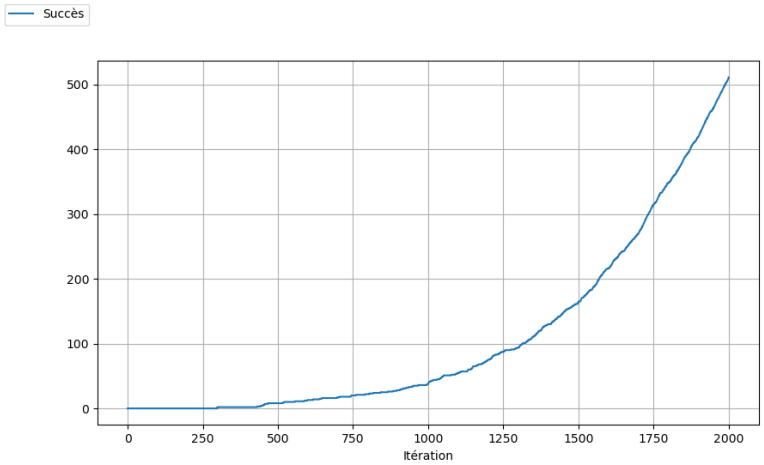
Comme la matrice Q reste constante, il n'y a pas de meilleures direction à choisir pour maximiser le nombre de récompenses.

Commenter l'évolution du nombre de succès en fonction du nombre d'itération.



On remarque que l'augmentation du nombre de succès n'est pas vraiment cohérente avec ce que l'on pourrait s'attendre lorsque le calcul présente une phase d'exploitation (i grand) puisque l'évolution tend à stagner. Ceci est lié par le fait que la matrice Q ne prend pas en compte les récompenses futures, et n'ai jamais réactualisée. Ainsi, la phase d'exploitation correspond toujours à une phase d'exploration car toutes les valeurs de Q sont identiques pour chaque état s.

Reprendre les 3 questions précédentes avec un facteur d'apprentissage α=1.



Dans ce cas, l'apprentissage se base sur une récompense qui ne prend en compte que les états futurs. On constate une évolution proche de celle avec une valeur par défaut du coefficient d'apprentissage (α=0,6) avec toutefois plus évolution un peu plus rapide.

3.2 *Influence du facteur de dépréciation*

Le facteur de dépréciation représente l'importance donnée aux actions futures.

Reprendre la même analyse que précédemment pour le facteur de dépréciation.

Pour γ=0, les actions ***a'*** dans les états futurs ***s'*** ne sont pas pris en compte de la réactualisation de la matrice **Q** pour l'état ***s***.Il n'y a donc pas propagation du meilleur chemin et seul la valeur de la récompense r est prise en compte. Ainsi, d'après la définition de la récompense, seuls le critère de distance entre l'état ***s*** et l'état final s_f , la présente d'obstacle et la réussite intervient. La matrice Q n'évolue pas vers des valeurs suffisamment élevées qui permettraient de proposer une chemin à privilégier (peu de flèches verte en dehors du parcours le plus cours). Au contraire, pour g=1, on

observe plus de flèche verte qui convergent vers la meilleure trajectoire signe d'une propagation des récompenses des états futurs.

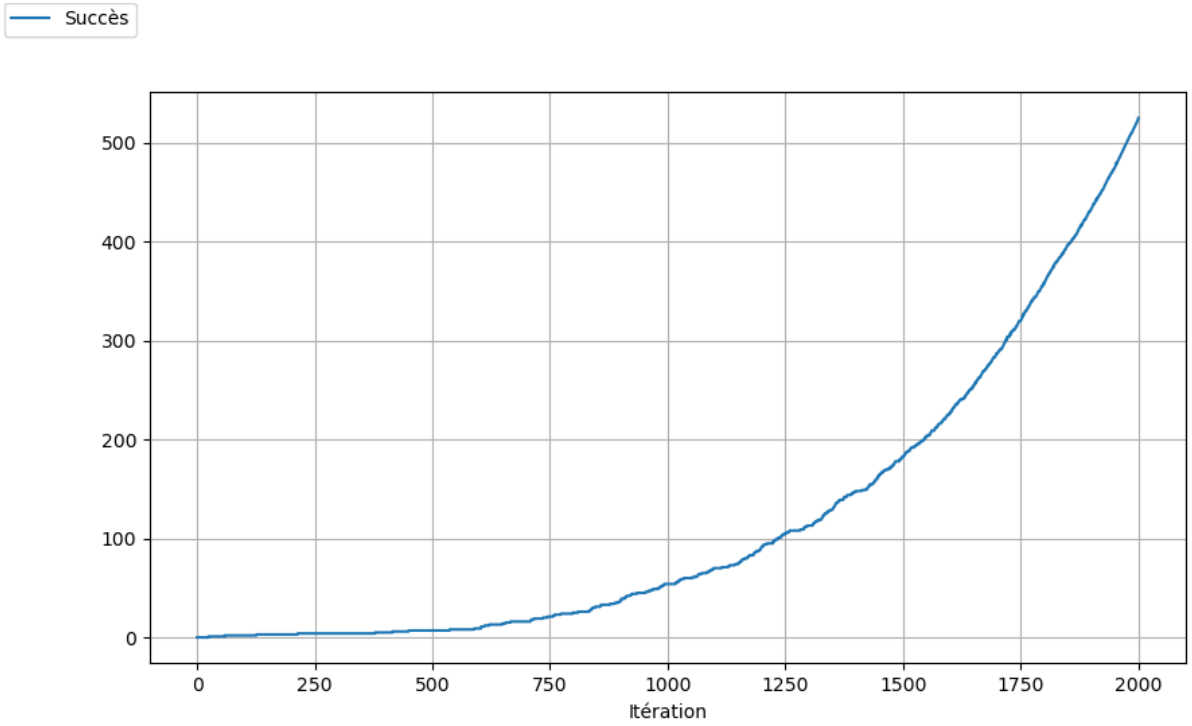


Figure 5: Evolution des succès pour γ=0

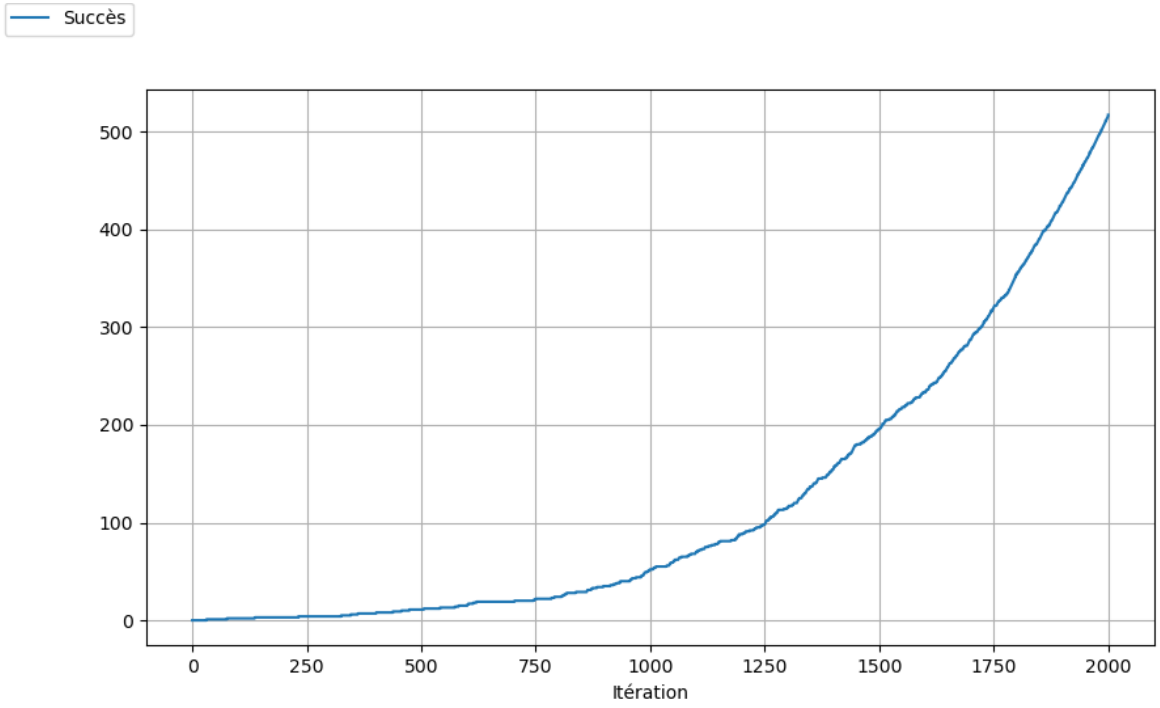


Figure 6: Evolution des succès pour γ=1

3.3 Influence de la récompense.

On rappelle ici le calcul de la récompense et les valeurs particulière 10 (respectivement -10) en cas de réussite (resp. de rencontre d'obstacle ou sortie de domaine)

3.3.1 Influence du facteur multiplicatif

Réaliser des simulations pour des cas extrême du facteur multiplicatif (valeur par défaut) et analyser le résultat.

Pour des valeurs plus élevées du facteur multiplicatif, la distance entre l'état s et l'état final s_f prend plus d'importance que la récompense issue d'une réussite. Ainsi lors du choix de l'action a à réaliser durant des phases d'exploitation, le chemin préférentiel pourra ne sera pas forcément le chemin vers l'état final.
Dans l'exemple donnée sur l'extrait vidéo (***Divergence_due_a_une_mauvaise_recompense.mp4***), on observe une divergence entre deux états. Pour ces deux états, les valeurs de la matrice Q associé au choix de l'action menant à l'état final est de 10 (normal car associé à la valeur de la récompense pour une victoire), alors qu'elle dépasse 20 pour la case d'au dessus (ou d'en dessous), ainsi, seule un choix d'action aléatoire permettrait d'avoir une chance de sortie de cet état, ce qui n'est quasiment plus possible vers la fin du calcul « à cause » du epsilon-greedy.

On constate également qu'il y a beaucoup de cases avec un chemin préférentiel proposé, ceci est du à des valeurs dans la matrice Q qui augmentent plus vite que pour un facteur multiplicatif faible.

Enfin, l'évolution du nombre de succès n'est pas vraiment impacté, par contre, le temps de calcul est plus long du aux divergences présentées précédemment.

3.3.2 Influence sur la distance à l'état final

Réaliser des simulations permettant d'identifier l'influence de la distance minimale bonus/malus (par défaut 2 cases = 120 mm) sur le résultat de la simulation.

Quelque soit la distance choisie, on n'observe pas de modifications notables sur l'évolution des succès ni même sur l'évolution des chemins préférentiels.

3.3.3 Influence des bonus/malus

Réaliser des simulations permettant d'identifier l'influence des bonus/malus (par défaut +10 et -10) sur le résultat de la simulation.

L'influence des bonus/malus est assez proche de l'influence du facteur multiplicatif. Les valeurs des bonus/malus doivent toujours avoir des valeurs suffisamment grandes pour ne pas être dépassées par la récompense basée sur la distance pour ne pas retomber sur les mêmes problèmes identifiés plus haut.

4 CONCLUSIONS

Comment expliquer que parfois, au cours du calcul, le meilleur chemin (violet) ne reprend pas les meilleurs choix possible d'un point de vue de la matrice Q ?

Le meilleur chemin peut-être trouvé de manière aléatoire (c'est ce qui se produit au début du calcul). Le fait qu'il n'y a pas eu assez de passage sur certains états fait que les coefficients de la matrice Q pour ces états n'ont pas encore convergés vers une valeurs données. Il faut attendre plusieurs succès et plusieurs passags sur le même états pour que les récompenses futures se propagent. Ainsi, au cours du calcul, pour un même état, la flèche verte va évoluer et au finale, devrait indiquer le « bon » chemin à suivre.
Toutefois, il se peut que deux chemins différents soient tous les deux optimaux, dans ce cas, les flèches vertes sur une des trajectoires optimales peuvent ne pas correspondre à l'orientation des flèches violettes.