CYCLE 5

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

TP2 - PSI

TP 2.2

PELLETEUSE – APPRENTISSAGE PAR RENFORCEMENT

1 INTRODUCTION

1.1 *Présentation*

La pelleteuse électrique autonome est un système didactisé qui s'inspire des évolutions actuelles des engins de chantier, d'une part dans leur changement de pratique en terme d'empreinte environnementale (abandon des moteurs thermiques) et d'autres part dans l'automatisation des tâches (véhicules autonomes).

Le système didactisé reprend uniquement l'ensemble **bras articulé** d'une pelleteuse constitué des 3 sous-ensembles (la **flèche**, le **balancier** et le **godet**) disposant chacun d'un actionneur (vérin électrique) et d'un système de transmission/transformation de mouvement. Le but de la pelleteuse étant de pouvoir remplir le godet d'une charge (terre, graviers, gravats, etc.), la déplacer et enfin la décharger tout en respectant le milieu environnant. De manière traditionnelle, ces actions se font manuellement par une opérateur qui agit sur deux joysticks permettant chacun de commander 2 actionneurs. Dans le cas du système didactisé, l'actionneur lié au mouvement de rotation vertical de la tourelle n'est pas utilisé.

Le banc didactisé dispose de deux mode de fonctionnement :

- **mode réel** : permet d'agir directement sur le bras de la pelleteuse
- **mode simulé** : permet d'agir sur le jumeau numérique dont le comportement est simulé sur ordinateur

Les deux modes peuvent se piloter par le biais des joysticks ou via l'interface de commande.

1.2 Problématique

Dans le cas d'un travail autonome d'une pelleteuse, la seule tâche imposée par l'utilisateur est un volume délimité que la pelleteuse doit « retirer ». En plus de la gestion autonome de son positionnement dans l'espace, non étudiée ici, la pelleteuse doit être capable de positionner et d'orienter son godet mais aussi de gérer les trajectoires de celui-ci dans son environnement.

Cette étude se concentrera sur la détermination autonome d'une trajectoires du bras pour aller d'une position initiale à une position finale en évitant un obstacle tout en minimisant l'énergie consommée.

L'objectif de ce TP est de décrire une méthode d'apprentissage par renforcement et d'analyser l'influence des paramètres sur l'apprentissage.

1.3 Prise en main du système

Après avoir mis en marche le système et placer le système sous mode de commande par joystick depuis le logiciel de commande MyViz.

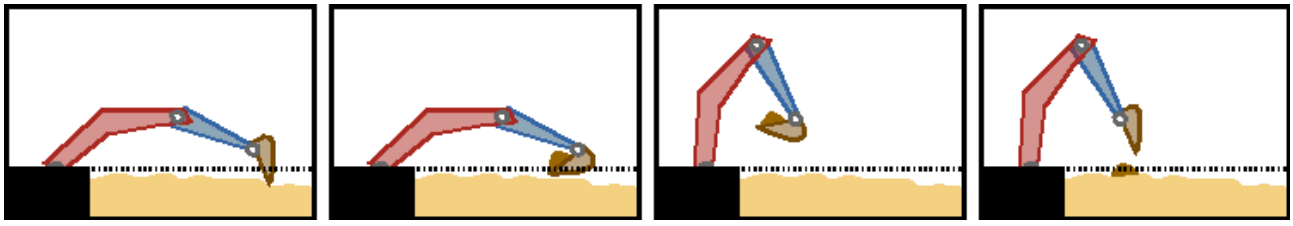
Actionner les joysticks suivant différentes inclinaisons et observer les effets sur le mouvement des pièces. Quelle est l'influence du degré d'inclinaison du joystick sur la vitesse de déplacement des vérins électriques ?

1.4 Réalisation d'opérations classiques

Lors de l'utilisation de la pelleteuse, l'opérateur est amené à réaliser différentes opérations telles que le creusement ou le nivellement.

1.4.1 Chargement-déplacement-déchargement

Lors des opérations de creusement (excavation, réalisation d'un tranchée, etc.), l'opérateur procède de manière cyclique : mise en position du godet, chargement, déplacement puis déchargement.

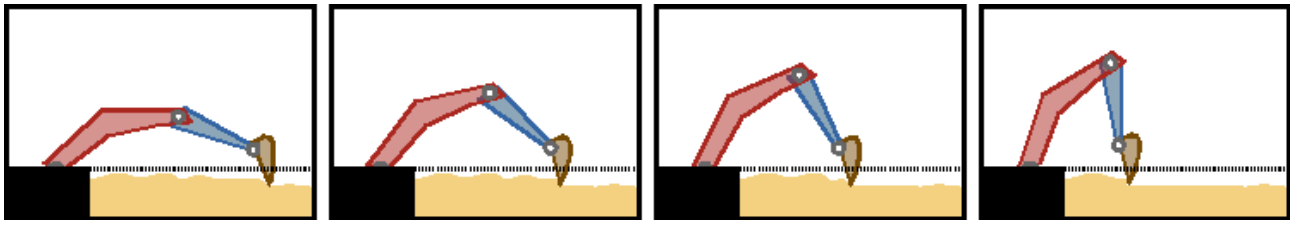


Entraînez vous à réaliser une procédure de chargement, déplacement puis déchargement du godet en évitant de perdre le chargement en route.

Après avoir réalisé quelques cycles, quelles sont les difficultés rencontrées ?

1.4.2 Nivellement

L'opération de nivellement consiste en un déplacement vertical ou horizontal du godet permettant de définir des contours plans de la zone en travaux. Seul le fond ou la lame du godet est alors en contact avec le sol.



Essayer de déplacer le godet suivant une trajectoire rectiligne horizontale ou verticale, quelles sont les difficultés rencontrées ?

2 MISE EN PLACE DE L'APPRENTISSAGE

2.1 *Environnement*

2.1.1 Zone de travail

On considère que le calcul de trajectoire se fait lorsque le godet est chargé. Lors du déplacement du bras, celui-ci doit donc garder une orientation fixe par rapport au bâti (sol). À cause de la géométrique du bras, toutes les positions possibles du godet ne sont pas compatibles avec une orientation « horizontale » du godet. C'est pourquoi, la zone d'apprentissage est volontairement limitée.

Dans MyViz, tableau de bord « ApprentissageRenforcement », observer cette zone (gris clair) et la valider approximativement avec le tableau de bord « Prise en main » et les joysticks.

2.1.2 Obstacles

Afin de rendre l'apprentissage le plus proche possible des situations réelles, deux formes d'obstacles sont proposées, chacun pouvant être associé à un type de travail.

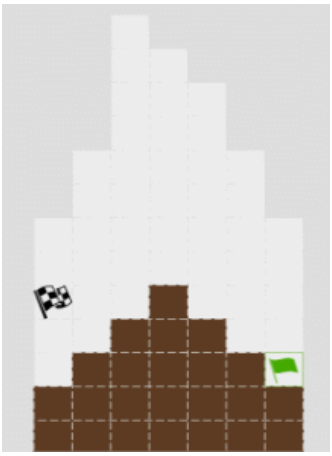


Figure 3: obstacle triangulaire

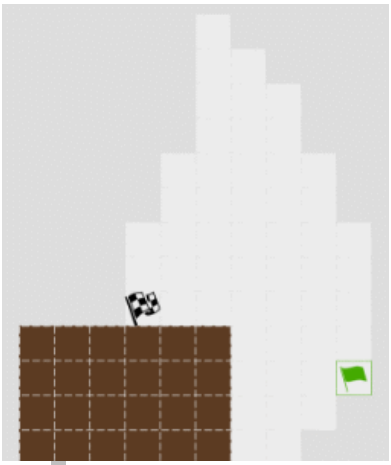


Figure 4: obstacle rectangulaire

Travail de prise de déchargement de terre dans une benne en passant au dessus de la ridelle	*Travail de creusement avec nivellement vertical*

2.1.3 Points de départ et d'arrivée

Comme l'illustre la figure précédente, la trajectoire calculée partira du point initial (drapeau vert) et arrivera au point final (drapeau à damier).

La position de ces points peut être modifiée en les glissant dans la fenêtre (maintient du clic gauche pendant le déplacement), mais sous la contrainte d'une abscisse du point de départ supérieure ou égale à celle de du point d'arrivée (voire paragraphe Actions)

2.2 .Actions possibles

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

En déduire quelques mouvements élémentaires possibles du godet.

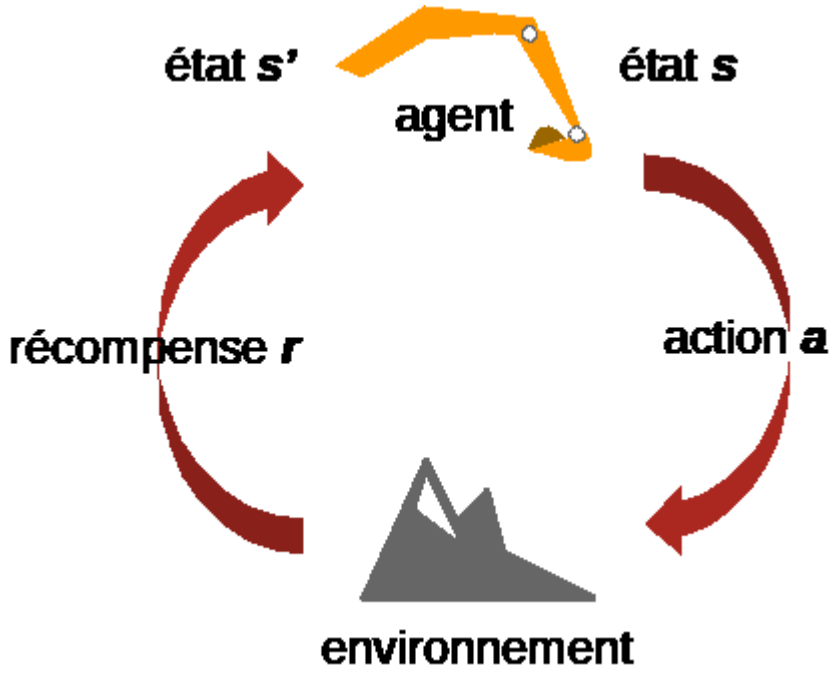
En analysant le travail réalisé par un conducteur de pelleteuse, on se rend compte que le déplacement du godet est toujours réalisé en « jouant » simultanément sur les vérins de flèche et de pénétration. Les actions possibles sont donc limitées aux 5 déplacements dans le plan : Haut, Gauche-Haut, Gauche,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

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

3 APPRENTISSAGE PAR RENFORCEMENT Q-LEARNING



3.1 Fonctionnement général

Ce modèle d'apprentissage est basé sur les récompenses acquises par le bras (agent) au cours de son évolution (action) dans son environnement. La meilleure trajectoire possible sera celle qui maximise le nombre ou la valeur des récompenses futures.

L'agent se trouve dans un état ***s***, il décide d'une action ***a*** en fonction de son état actuel ***s*** et d'une fonction ***Q***. L'agent se retrouve alors dans un état ***s'*** avec une récompense ***r*** qui dépend de la fonction ***Q*** et de l'environnement.

3.1.1 Matrice Q (Q-value)

Cette matrice ***Q*** mesure la qualité de l'action ***a*** exécutée par l'agent lorsqu'il se trouve dans l'état ***s*** pour passer à l'état ***s'***. Elle est définie par

$$Q[s,a] := (1-\alpha)Q[s,a] + \alpha\left(r + \gamma \max_{a'} Q[s',a']\right)$$

avec

- α *: facteur d'apprentissage (learning rate) avec* $0 \leqslant \alpha \leqslant 1$;
- *r : récompense ;*
- γ *: facteur de dépréciation (discount factor) avec* $0 \leqslant \gamma \leqslant 1$

3.2 Choix de l'action à prendre : Epsilon-greedy

Au terme de l'apprentissage, l'action ***a*** à prendre par l'agent est celle qui maximise les récompenses futures, « enregistrées » dans la matrice ***Q***.
Or, au début de l'apprentissage, la matrice ***Q*** ne reflète en rien l'environnement, l'action a à prendre par l'agent ne peut donc pas se baser sur la matrice ***Q*** mais doit être prise au hasard afin d'explorer l'environnement.
Dans la méthode du Q-learning, on retient très souvent la méthode simple ε-greedy (epsilon-greedy) décrite par le code python suivant

```
import numpy as np

def choix_action(i,N):
    eps=1-i/N
    p = np.random.random()
    if p < eps:
        alea = true
    else:
        alea = false
```

où N est le nombre d'itérations choisi pour l'apprentissage et i le numéro de l'itération courante.

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.

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

3.3 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 ***r*** est définie par

$$r=[120-d(s-s_f)].10^{-3}$$

avec *d* la distance entre l'état actuel ***s*** et l'état final 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

- *r=10* quand l'état ***s*** atteint correspond à l'état final
- *r=-10* quand l'état ***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.

Remarque : dans la suite de l'étude, la valeur 120 et le coefficient 10^{-3} pourront être modifiés pour analyser leur influence.

4 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.

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

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

4.1 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

$$Q[s,a]:=(1-\alpha)Q[s,a]+\alpha\left(r+\gamma \max_{a'} Q[s',a']\right)$$

.

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 ?

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

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

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

4.2 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.

4.3 Influence de la récompense.

On rappelle ici le calcul de la récompense $r=[120-d(s-s_f)].10^{-3}$ et les valeurs particulière 10 (respectivement -10) en cas de réussite (resp. de rencontre d'obstacle ou sortie de domaine)

4.3.1 Influence du facteur multiplicatif

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

4.3.2 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.

4.3.3 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.

5 CONCLUSIONS

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