PROBLÈME DU CHARIOT ET DU BÂTON

Guillaume Gagné-Labelle

RÉSUMÉ

La validité de l'approche Q-Learning, une méthode d'apprentissage par renforcement en situation de contrôle, a été étudiée et comparée à une méthode de résolution logique et physique à travers un problème de dynamique classique. Le but du problème était de minimiser de l'angle d'un bâton de 1m maintenu en équilibre sur un véhicule en mouvement unidimensionnel sans friction. L'apprentissage par renforcement d'une direction à laquelle appliquer une force 10N à tout intevervalle de 0.02s a permis d'atteindre des performances suprahumaines, mais il n'est généralement pas optimal d'appliquer cette méthode aveuglément puisque l'étude du problème d'un point de vue physique tend à être favorable: la méthode est plus stable, moins énergivore, mais surtout explicable.

1 INTRODUCTION: PHYSIQUE RÉGISSANT LE SYSTÈME (1) (2)

Le problème du chariot et du bâton est défini de la sorte (1): un chariot de masse $M = 1\text{kg}$ est libre de se déplacer selon l'axe $\hat{x}$ et est contrôlé par l'uilisateur via une force de 10N appliquée à chaque intervalle de 0.02s (cinquantième de seconde). Sur ce chariot est fixé une bâton de longueur $2l = 1\text{m}$ et de masse $m = 0.1\text{kg}$. Ce bâton est libre de se déplacer autour de l'axe $\hat{z}$, i.e. dans le plan XY et est soumis à l'accélération gravitationnelle ($g = 9.8\text{m/s}^2$) en direction $-\hat{y}$. Dans le référentiel du centre de masse du bâton, le mouvement latéral du chariot inflige un moment de force sur le bâton en son extrémité. Dans le référentiel du chariot, par la 3e loi de Newton, le bâton inflige au chariot un moment de force opposé qui se traduit en une force latérale en $\hat{x}$ étant donné que le chariot est fixé à la rail. Le but du problème est de minimiser l'angle θ, soit l'angle formé par le bâton et l'axe $\hat{y}$, ou équivalemment de maintenir le bâton en équilibre sur le chariot, tout en restant à l'intérieur des limites latérale de $0 \pm 2.4\text{m}$.

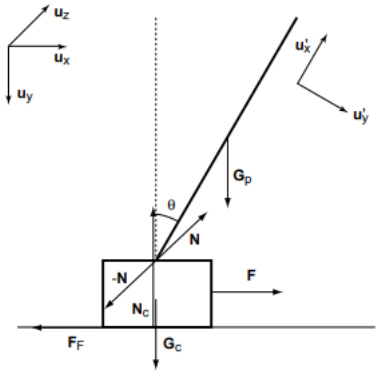


Figure 1: Visualisation du système chariot-bâton. Le chariot est ponctuel et considéré fixé sur la rail sans frottement.

Considérons que le chariot agit avec une force $\mathbf{N}$ sur le bâton, repoussant ce dernier. Réciproquement, le bâton pousse le chariot avec une force $-\mathbf{N}$. Ainsi, en appliquant la 2e loi de Newton au chariot:

$$\mathbf{F} + \mathbf{G_c} - \mathbf{N} + \mathbf{N_c} = m_c \mathbf{a_c}$$

avec $\mathbf{a_c} = \ddot{x}$. En faisant de même pour le bâton:

$$\mathbf{N} + \mathbf{G_b} = m_b \mathbf{a_b}$$

Toutefois, on peut considérer que l'accélération a_b du bâton en son centre de masse est due à l'effet composé de l'accélération du chariot ainsi qu'à son accélération centripète. On peut alors écrire:

$$\mathbf{a_b} = \mathbf{a_c} + \alpha \times \mathbf{r_b} + \omega \times (\omega \times \mathbf{r_b})$$

où $\alpha = \ddot{\theta}\mathbf{u_z}$ est l'accélération angulaire, $\omega = \dot{\theta}\mathbf{u_z}$ est la vitesse angulaire et $\mathbf{r_b} = l(\sin\theta\mathbf{u_x} - \cos\theta\mathbf{u_y})$

En résolvant le système d'équations pour $\ddot{x}$ et $\ddot{\theta}$ (2), on trouve:

$$\ddot{\theta} = \frac{g\sin\theta + \cos\theta\left(\frac{-F - m_b l\dot{\theta}^2\sin\theta}{m_c + m_b}\right)}{l\left(\frac{4}{3} - \frac{m_b\cos^2\theta}{m_c + m_b}\right)} \tag{1}$$

$$\ddot{x} = \frac{F + m_b l\left(\dot{\theta}^2\sin\theta - \ddot{\theta}\cos\theta\right)}{m_c + m_p} \tag{2}$$

À partir d'ici, la physique du système est pratiquement achevée. Seuls la position du chariot (x), sa vitesse ($\dot{x}$), l'angle du bâton (θ) et sa vitesse angulaire ($\dot{\theta}$) nous intéressent. Ce sont les variables d'entrée à chaque instant temporel afin de prendre un décision. En prenant un pas de temps assez petit, soit un cinquantième de seconde tel qu'énoncé, nous pouvons donc décrire l'évolution temporelle de ces variables en approximant leur intégrale temporelle par la méthode du rectangle:

$$\begin{aligned}
\dot{\theta}_{t+1} &\approx \dot{\theta}_t + \Delta t \cdot \ddot{\theta}_t \\
\theta_{t+1} &\approx \theta_t + \Delta t \cdot \dot{\theta}_t \\
\dot{x}_{t+1} &\approx \dot{x}_t + \Delta t \cdot \ddot{x}_t \\
x_{t+1} &\approx x_t + \Delta t \cdot \dot{x}_t
\end{aligned}$$

2 Méthodologie: Apprentissage par renforcement (3) (4)

L'apprentissage par renforcement est un modèle d'intelligence artificiel permettant d'apprendre des actions de contrôle sur un système afin d'optimiser une récompense quantitative au cours du temps. Un agent (ou acteur) est plongé dans un environnement, typiquement markovien, où il apprend à agir en fonction de l'état courant du système au cours du temps. Cette méthode d'apprentissage est active plutôt que passive en ce sens que l'agent apprend par l'expérimentation. De plus, elle a l'avantage de trouver des solutions même dans le cas où la solution optimale au problème est inconnue analytiquement. Survolons les concepts fondamentaux définissant l'apprentissage par renforcement:

2.1 Agent

À chaque pas de temps t, l'agent:

- Reçoit une observation O_t provenant de l'environnement, i.e. une représentation (typiquement) partielle ou totale de ce dernier.
- Reçoit une récompense R_t
- Exécute une action A_t conformément à une politique π_t

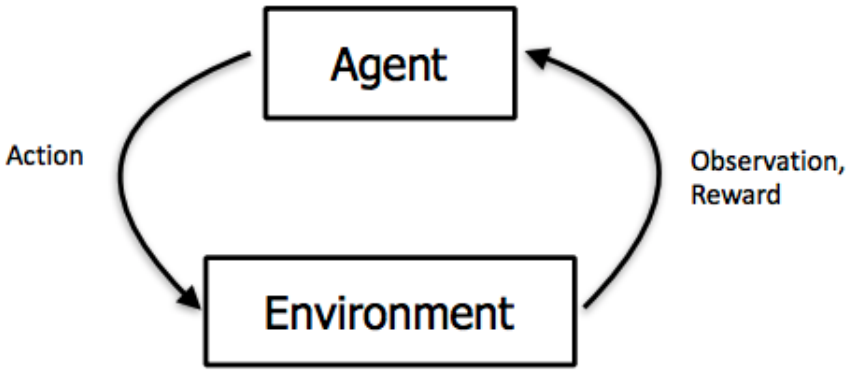


Figure 2: Séquence typique de l'interaction d'un agent d'apprentissage par renforcement avec son environnement, faisant évoluer sont état.

Dans notre cas, l'observation correspond à un vecteur à quatre composantes $O_t = (x, \dot{x}, \theta, \dot{\theta})$. Selon cette observation, l'agent décide de la direction ($\pm \mathbf{u_x}$) de la force qui est appliquée sur le chariot; il s'agit de son action. Cette force est constante, de module 20N et dure pendant $\Delta t = 0.02$s. Suite a cet intervalle de temps où le système évolue, selon sa stabilité, l'agent recevra une récompense et une nouvelle observation. Il sera alors demandé à l'agent de prendre une nouvelle action, et ainsi de suite advitam eternam. C'est lors de ces itérations que le processus de décision de l'agent se développera afin d'aiguiser la précision des actions choisies.

2.2 Environnement

L'environnement est quant à lui sujet aux actions de l'agent, mais aussi aux conditions du système régissant son évolution déterministe ou stochastique. En d'autres mots, l'agent peut faire évoluer l'environnement partiellement ou totalement. Formellement, l'environnement:

- Reçoit une action A_t provenant de l'agent.
- Émet une observation partielle ou totale O_{t+1}.
- Émet une récompense R_{t+1}

Dans le cas du problème du chariot et du bâton, l'environnement est caractérisé par la dynamique des objets, mais aussi par ses frontières. Considérant que le chariot se déplace autour de l'origine, les frontières sur l'axe x sont symétriquement et arbitrairement placées à $\pm 2,4$m, tandis que les limites angulaires sont placés à $\pm 30^o$ autour de l'axe verticale. Lorsque l'agent pose une action faisant en sorte que le système chariot-bâton sort des limites de l'environnement, cette action est considérée comme un échec, l'épisode s'arrête et se réinitialise, et la récompense (nulle) est envoyée accordément.

2.3 Récompense

La récompense R_t est une scalaire mesurant la performance de l'agent au temps t. Il définit le but. Le rôle de l'agent est de maximiser la récompense cumulative (ou retour) sur le temps:

$$G_t = \sum_{t'=t}^{\infty} R_{t'}$$

L'apprentissage par renforcement est basé sur ce concept de maximisation le retour au cours du temps. Pour le problème la récompense vaut 1 à chaque itération où le système est à l'intérieur des limites et 0 dans le cas contraire.

Valeur et Valeur-Action La valeur est définie comme la récompense cumulative estimée à partir d'un agent dans un état s.

$$V(s) = \mathbb{E}(G_t|S_t = s) = \mathbb{E}(R_{t+1} + R_{t+2} + \ldots|S_t = s)$$

Le but de l'agent est alors de maximiser la valeur en choisissant des actions cohérentes. La récompense et la valeur définissent le niveau de désirabilité d'un état ou d'une action. Il est à noter que le retour et la valeur peuvent être définie récursivement:

$$\begin{aligned} G_t &= R_{t+1} + R_{t+2} + \ldots \\ &= R_{t+1} + G_{t+1} \end{aligned}$$

La valeur-action $Q(s, a)$ est quant à elle définie comme la récompense cumulative estimée à partir d'un agent dans un état s effectuant une action a.

$$Q(s, a) = \mathbb{E}(G_t|S_t = s, A_t = a) = \mathbb{E}(R_{t+1} + R_{t+2} + \ldots|S_t = s, A_t = a)$$

Au fil de l'apprentissage, la valeur-action Q est estimée avec de plus en plus de précision et le processus de décision est alors choisi accordément.

2.4 Politique

Finalement, une politique π d'un agent est une fonction définissant le comportement de l'agent. Pour tout état ou observation O, elle cartographie cet état à une action:

$$\pi : O \mapsto A$$

L'apprentissage par renforcement évolue fondamentalement autour de l'apprentissage de cette politique. Pour chaque observation, il s'agit de trouver un moyen de quantifier la valeur de cette dernière et de trouver quelle action parmi celles disponibles maximise la valeur espérée à l'instant $t + 1$.

Différents algorithmes existent pour maximiser la récompense cumulative espérée en troquant parfois la maximisation instantanée pour une meilleure exploration de la plage des actions. En particulier, pour ce projet, une approche Monte-Carlo avec un algorithme ϵ-glouton a été mis en place. Cet algorithme est défini de la sorte:

- Avec probabilité $1 - \epsilon$, on choisit au temps t la meilleure action estimée: $a = \underset{a \in A_t}{\operatorname{argmax}}(Q_t(a))$
- Avec probabilité ϵ, on choisit au temps t une action aléatoire.

Mathématiquement,

$$\pi_t(a|s) = \begin{cases} 1 - \epsilon + \epsilon/|A| & \text{si } Q_t(a) = \max\limits_b Q_t(b) \\ \epsilon/|A| & \text{sinon.} \end{cases}$$

Et la performance de test a été évaluée en sélectionnant continuellement la meilleure action plutôt que d'introduire une action aléatoire.

Les définitions présentées s'appliquent très bien au problème choisit. Toutefois, on remarque que l'ensemble des états est infini puisque les positions et vitesses sont des valeurs continues et de même pour le temps. Pour la modélisation numérique, on veut cependant éviter que cet ensemble soit infini parce qu'on veut pouvoir garder une trace de chaque état. En effet, l'idée est de tabuler la fonction de valeur-action et d'itérativement améliorer l'estimation du système pour maximiser les actions.

Pour ce faire, les valeurs possibles de position, de vitesse, d'angle, de vitesse angulaire et de temps ont été discrétisées en intervalles Δx (4 intervalles), $\Delta \dot{x}$ (2 intervalles), $\Delta \theta$ (32 intervalles) et $\Delta \dot{\theta}$ (16 intervalles) respectivement. Ensuite, un tenseur d'états en 5 dimensions a été créé où les quatres premières composantes représentent un état et la cinquième composante représente l'action à prendre (bidimensionnelle, gauche ou droite). Plus simplement, la situation peut être vue comme une fonction qui prend en entrée une observation, la discrétise et retourne une valeur dans l'ensemble

$\{\text{gauche}, \text{droite}\}$. Cette action est alors effectuée pour un intervalle de temps Δt. Ensuite, l'entrée de la matrice sera mise à jour selon la récompense obtenue.

En pratique, plutôt que de simplement pondérer linéairement les récompenses, un taux d'apprentissage adaptatif a été utilisé (4). Ce taux d'apprentissage adaptatif incarne l'idée où l'agent devrait apprendre plus fortement de ses actions au début de l'entraînement, mais, au fil du temps, les nouvelles actions devraient avoir de moins en moins d'importance. En effet, bien que les deux méthodes se valent, l'utilisation d'un taux d'apprentissage adaptatif à décroissance logarithmique permet un apprentissage plus rapide en pratique ainsi qu'une plus grande stabilité à long terme.

Ceci étant dit, le tenseur multi-dimensionnelle est alors une approximation de la fonction valeur-action. En effet, chaque entrée est une moyenne empirique pondérée du retour espéré pour chaque état et chaque action. Le tenseur fût alors utilisé pour l'algorithme ϵ-glouton en choisissant l'action maximisant l'espérance du retour selon l'entrée.

Étant donné qu'une action peut mener à un échec de différentes façons, c'est-à-dire en sortant des limites de l'axe $\hat{x}$ établies ou lorsque l'angle θ est trop élevé, l'approche par renforcement est d'autant plus intéressante parce que l'agent aprend à éviter toutes les conditions d'échec en même temps. En effet, en segmentant l'espace angulaire et linéaire, pour un même angle, l'agent différencie une action posée en bordure latérale à celle posée au centre, permettant une optimisation multidimensionnelle. Cette approche est généralement très difficile de façon analytique lorsque les variables sont indépendantes.

3 VALIDATION

	Système libre **Énergie totale**	**Force vers la gauche** **Énergie cinétique**	**Force vers la droite** **Énergie cinétique**
R^2	$(4.44 \pm 0.18)E{-}17$	$0.99999999640 \pm 1.8E-10$	$0.99999999614 \pm 1.8E-10$
EQM	$(3,6 \pm 0.5)E{-}7$	0.225 ± 0.010	0.224 ± 0.010

Table 1: Régression de l'énergie totale ou cinétique du système soumis à différentes conditions permettant de démontrer la conservation de l'énergie du système.

Tel que présenté, le système est difficilement anaylisable. Effectivement, la présence de la force externe variable dans le système produit un système non-conservatif instable d'un point de vue énergétique. Conséquemment, une approche en 3 étapes a été élaborée où, à chaque étape, la complexité du système augmentait en tendant vers le problème abordé.

En premier lieu, la force externe et les conditions aux frontières ont été complètement supprimées. Ainsi, en initialisant le système en une position aléatoire, l'énergie potentielle, cinétique et rotationnelle du système se devaient d'osciller, mais la somme de ces énergies devait quant à elle être conservée. Le respect de ce sous-système fût alors un témoin de la physique conservative du problème. Il est à noter que le calcul de l'énergie doit être fait par rapport au centre de masse du système, mais que les équations 1 sont les équations de l'accélération de chacun des objet. La dérivation de l'énergie de rotation de ce système à deux corps, de même que celle de l'énergie cinétique du centre de masse sont reportés en annexe B et C respectivement. En ce qui à trait à l'énergie potentielle, seul le bâton possède une hauteur non nulle et les calculs en sont simplifiés et omis.

On observe à la figure 3 un comportement généralement linéaire de l'énergie totale. Le taux de corrélation linéaire est d'ailleurs très bas: $R^2 = (4.44 \pm 0.18)E{-}17$ en répétant 5 fois l'expérience. Ceci est contre-intuitif, mais toute à fait vraisemblable. En effet, lorsqu'une variable y est constante par rapport à une variable x, les deux variables sont complètement décorrélées, tel qu'attendu. D'ailleurs l'erreur quadratique moyenne par rapport à la droite de pente nulle optimisée: $EQM = (3,6 \pm 0.5)E{-}5\%$, témoigne aussi de la constance de l'énergie totale du système. On peut alors conclure que, lorsque la force externe n'agit pas sur le système, ce dernier suit les lois de la dynamique classique.

En second lieu, la force externe sur le chariot a été ajoutée, mais de façon constante, c'est-à-dire en pointant toujours dans la même direction. En changeant légèrement les conditions aux frontières de l'axe $\hat{x}$ pour que ces dernières soient circulaires, une augmentation l'énergie devait être observée.

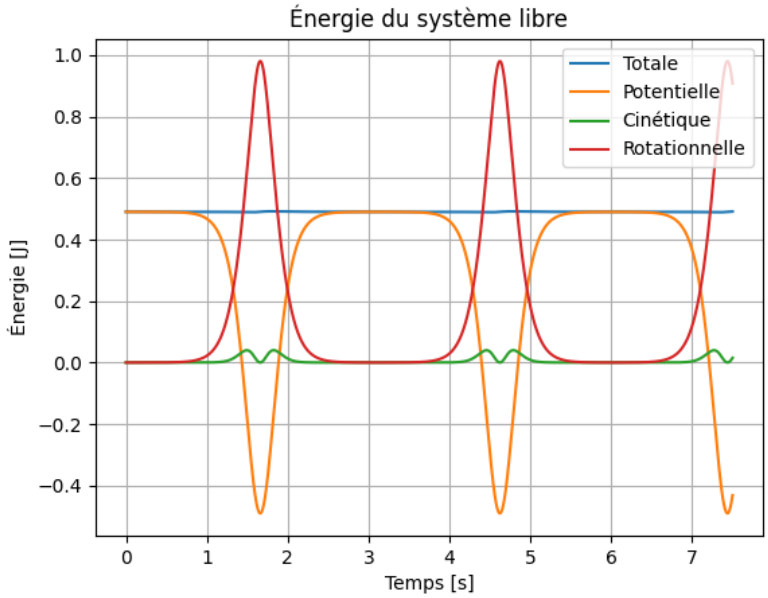


Figure 3: Évolution temporelle de l'énergie du système sans force externe. La configuration fût initialisée par un vecteur aléatoire uniforme à 4 dimensions dans l'intervalle $[-.05, 0.05]$.

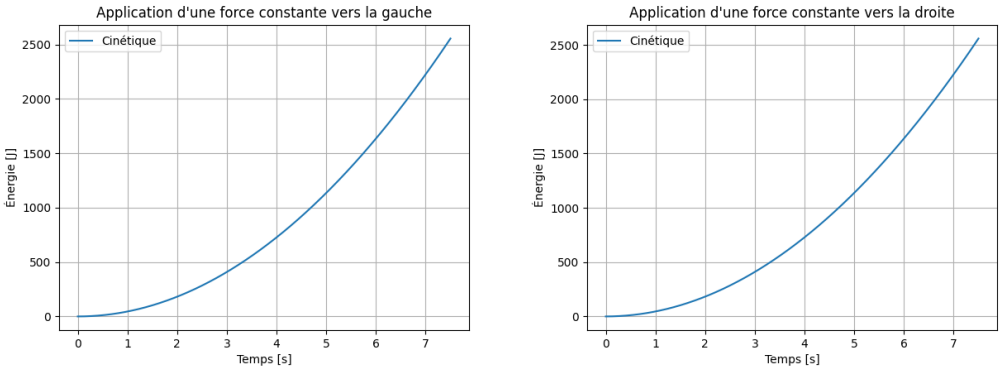


Figure 4: Évolution temporelle de l'énergie cinétique du système chariot-bâton lors de l'application d'une force directionnelle à gauche ou à droite de 10N sur le chariot.

En effet, la force ajoutée est conservative, on peut la voir comme une force de gravité latérale, mais le système se trouvait alors en chute libre sur son axe unidemensionnel. La variation de vitesse étant alors proportionnelle à cette force par la deuxième loi de Newton, l'augmentation constante de la vitesse signifie alors qu'une augmentation quadratique de l'énergie cinétique du système devait être observée (puisque l'énergie est proportionnelle à la vitesse au carré, qui est elle-même linéairement proportionnelle à la force).

Dans les deux cas, les coefficients de corrélation linéaire par rapport à une parabole de forme canonique sont pratiquement parfaits. Dans le cas de la force appliquée vers la gauche: $R^2 = 0.99999999640 \pm 0.00000000018$, tandis que dans le cas de la force appliquée à droite: $R^2 = 0.99999999614 \pm 0.00000000018$. Équivalemment $EQM = 0.225 \pm 0.010$ à gauche et $EQM = 0.224 \pm 0.010$ à droite. Avec confiance, il est alors possible, étant donné qu'une seule force à la fois est appliquée à chaque instant, de conclure que l'application des forces externes sur le système

respecte la physique du problème. Il ne s'agit que d'alterner leur application à divers instants temporels.

Enfin, le système a été résolu de façon logique afin de borner les performances de l'algorithme d'apprentissage. Intuitivement, si le bâton tend vers la droite du chariot et que sa vitesse angulaire est en sens horaire, le chariot devrait évidemment se déplacer vers la droite (de façon analogue pour la gauche). De plus, si le bâton se trouve très près de l'angle nul (moins de 3 degrés), alors la force devrait pointer dans le même sens que la vitesse angulaire du bâton (anti-horaire à gauche, horaire à droite) afin d'anticiper le mouvement. Cette approche basée sur les données physiques du système a permis des performances quasi-parfaites pour la résolution du problème, bornant ainsi supérieurement les performances de l'algorithmes d'apprentissage. Ces dernières sont présentées dans la section 4.

Incidemment, si le système respecte les deux premières conditions et que les résultats de l'algorithme d'apprentissage sont bels et bien bornés par la 3e, alors nous pouvons conclure à la validité du code.

4 RÉSULTATS (5) (6)

Afin pouvoir recueillir des résultats, les performances des approches utilisées ont toutes été plafonnées à 20 secondes par épisode. En effet, un système convergent réussit en théorie à maintenir le bâton en équilibre à l'intérieur des frontières pour un temps infini, mais il a été considéré que, après 20s, le système était suffisamment convergent et qu'un autre épisode pouvait alors commencer. Il est à noter que ces 20 secondes sont numériques. C'est-à-dire que, en connaissant la physique du problème (incidemment son évolution temporelle), chaque épisode de 20s a pu être calculé en beaucoup moins de temps. De plus, les systèmes ont tous été testés pendant 750 épisodes.

Méthode	**Humain**	**Q-Learning (post-entraîn.)**	**Physique**
Performance $[s]$	1.3 ± 1.4	18 ± 4	19.91 ± 0.16

Table 2: Comparaison des méthodes expérimentales permettant de résoudre le problème de façon temporelle.

Différentes approches ont été mises en place afin de contraster avec la méthode développée à la section 2. De prime abord, tel que mentionné à la section précédente, une approche basée sur une analyse heuristique de la dynamique du problème a été mise en place pour borner supérieurement la performance de l'algorithme (5). Cette méthode a produit des résultats quasi-parfaits en ce sens que le bâton ne tombait jamais, mais, très rarement, il était possible que le chariot sorte des limites latérales permises (± 2.4m) lors d'un épisode et un échec était alors enregistré. Néanmoins, les performances ont plafonnées à (19.91 ± 0.16)s.

Ensuite, un interface graphique a été développé (6) à l'aide des librairies *pygame* et *keyboard* afin de développer le jeu vidéo associé au problème et ainsi comparer les performances humaines aux performances machines. Étant donné que la force externe appliquée au chariot durait 0.02s, 50 actions se produisait à chaque seconde et tout moniteur d'écran possédant la capacité d'affichage de 50 "fps" pouvait alors simuler le problème avec un rendu exact. Cette approche a produit de piètre performance de (1.3 ± 1.4)s, possédant même un co-domaine négatif.

Enfin, la méthode d'apprentissage principale par l'algorithme Q-Learning a performé autour de la moyenne (18 ± 4)s, soit de bonnes performances, mais aussi hautement instables. Différentes métriques ont été rapportées. Premièrement, la métrique d'entraînement représente la performance de l'agent lorsque ce dernier suit la politique ϵ-glouton. Il est alors contaminé par une vague d'actions aléatoires et la performance diminue. Parallèlement, la performance d'évaluation représente la performance de l'agent au cours de l'apprentissage, mais lorsque ce dernier choisit (selon lui) la meilleure action disponible plutôt que la politique stochastique. Enfin, la performance post-entraînement utilise un second agent pré-entraîné, c'est-à-dire ayant déjà itéré 750 fois, et il suit la meilleure action disponible selon la table Q de valeur-action.

De façon supplétive à la performance, l'analyse de la maximisation indirecte de l'énergie potentielle du système a aussi été étudiée. En effet, les contraintes sur le système imposent que le bâton reste droit et son que son énergie potentielle s'en voit maximiser. Certains tests ont d'ailleurs été fait

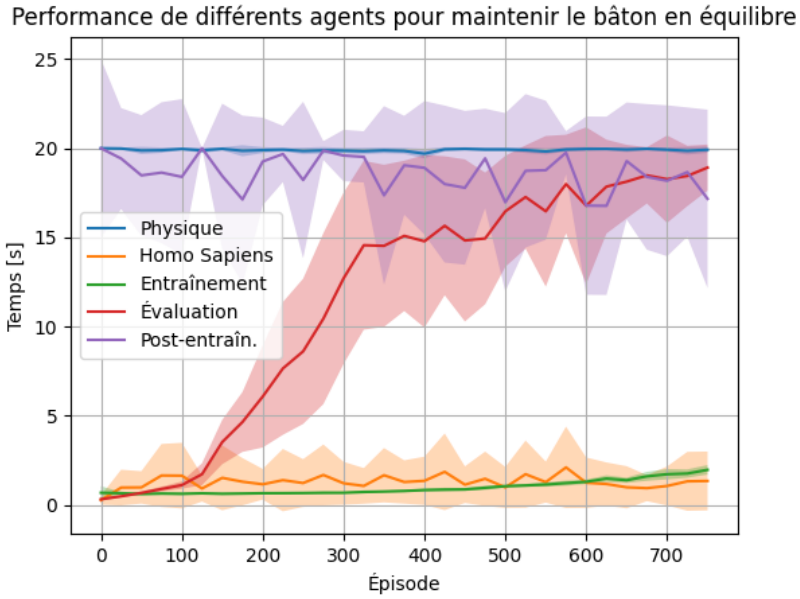


Figure 5: Comparaison de la performance d'un agent d'apprentissage automatique à différents stages de l'apprentissage en contraste avec d'autres approches de résolution.

afin que la récompense de l'agent soit proportionnelle à l'énergie potentielle du système, mais les résultats furent si décevant qu'ils en sont omis.

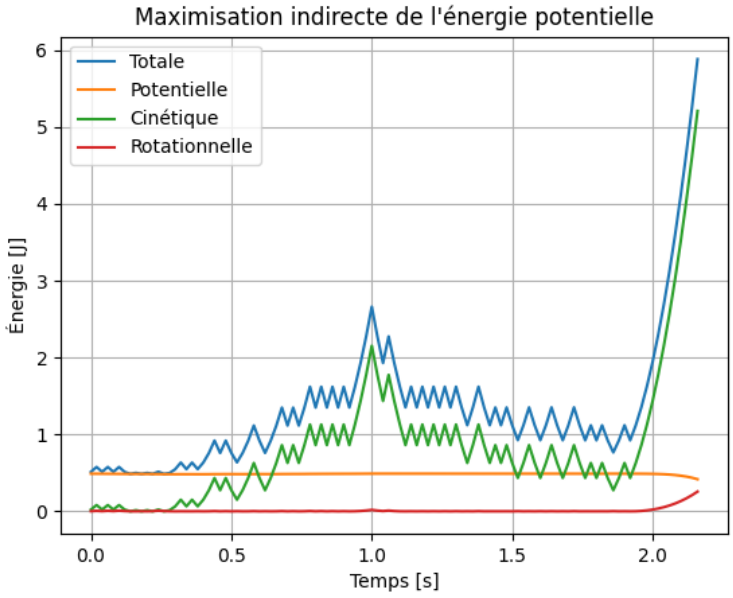


Figure 6: Évolution typique de l'énergie du système pour un épisode singulier échouant après 2.2s.

De la figure 4 ci-haut, on remarque que l'énergie cinétique du système est libre de varier, mais le système échoue dès que l'énergie potentielle décroit de sa valeur maximale. Équivalemment, le système échoue lorsque l'énergie rotationnelle augmente par transfert d'énergie lorsque le bâton tombe. La figure personnalise très bien l'idée derrière l'algorithme: peu importe l'action, si elle

maximise l'énergie potentielle, il s'agit d'une bonne action et on récompense positivement l'agent. À l'inverse, dès que l'agent pose une action qui ne maximise pas l'énergie potentielle, l'épisode échoue et la récompense est nulle.

5 ANALYSE

Les différentes méthodes présentées comportent chacunes leurs avantages et leurs désavatanges et, généralement, le problème dicte la méthode à utiliser afin de choisir les désavantages qui sont acceptables. Le problème du chariot et du bâton exemplifie bien ce compromis.

D'un côté, l'avantage principal de l'approche par apprentissage par renforcement est l'agnostisme de la méthode par rapport aux types de données du problème. En aucun temps il n'a été question d'être au courant de la physique régissant le problème pour développer l'approche. Il ne s'agit que de développer et d'optimiser une fonction de 4 paramètres (l'observation) dans 1 (l'action). Il aurait été possible d'ajouter ou même de diminuer la caridinalité du domaine ou du co-domaine en quelques lignes de code. De la même façon, le problème aurait pu être tout autre, par exemple un problème économique ou même strictement mathématique, et un informaticien sans aucune connaissance à priori pourrait développer une approche semblable et obtenir des résultats intéressants. Certains problèmes sont fondamentalement mystérieux ou agissent comme des boîtes noires, mais d'autres sont aussi simplement trop difficiles: ils contiennent trop de variables à optimiser en même temps ou font partie de l'ensemble des problèmes NP-difficiles. Pour ce genre de problème, l'approche par renforcement est tout à fait adéquate. Elle trouve une solution minimale locale produisant des résultats adéquats et performants lorsqu'aucune approche analytique n'existe.

Toutefois, la méthode comporte aussi plusieurs désavantages. Principalement, elle est inexplicable et très difficilement interprétable. Par exemple, le système échoue à quelques occasions sur des instances du problèmes très similaires à d'autres instances où il performe très bien, sans donner d'information sur les causes de l'échec. Le comportement est erratique et sa fiabilité s'en voit affecté. La variance élevée des performances post-entraînement en est un bon témoin. En effet, bien que la moyenne de 18s de performance soit tout à fait respectable, la performance minimale pour une itération était de 3.26s, faisant grimper la variance à 4s. De plus, l'approche Q-Learning ne généralise pas bien à l'extérieur du domaine d'entraînement. En effet, c'est l'expérience et l'essai-erreur qui fait en sorte que l'agent performe bien. Tout repose sur la validité du tenseur des décisions. Pour n'importe quel état à l'extérieur de ce tenseur, le système n'a aucune idée de l'action à prendre et il doit être entraîner sur tout nouvel état avant de pouvoir prendre une décision sensée. Une problématique en amène une autre alors que le fléau des grandes dimensions se fait sentir. En effet, ici, le système possédait 4 variables (position, vitesse, position angulaire et vitesse angulaire) qui étaientt chacune discrétisée en 4, 2, 32, et 16 espaces. Le tenseur de décision possédait ainsi $4 \cdot 2 \cdot 32 \cdot 16$ entrées. Incidemment la croissance de la complexité de l'algorithme est exponentielle par rapport aux nombres de variables en entrée, ce qui est évidemment problématique pour certains problèmes complexes. Cette méthode ne passe pas à l'échelle.

De l'autre côté, les avantages et les désavantages de l'approche physique de résolution du problème sont tout à fait dichotomiques. Entre autres, la méthode se distingue par sa fiabilité et par le peu de mémoire informatique nécessaire à la résolution du problème, mais surtout, la solution est interprétable. En effet, pour chaque état, même si de nouveaux étaient introduits (par exemple une pente), l'action à prendre est prévisible et explicable. De plus, il est facile de comprendre la raison pour laquelle le système échoue: les limites latérales ne sont pas prises en compte et il arrive que le chariot dérive et en sorte. Ce genre de problème a tendance à arriver très tard lors d'un épisode parce que le chariot dérive lentement, et les résultats sont donc très bons et la variance est très petite.

Cependant, cette méthode contient aussi des lacunes non négligeables. Premièrement, elle se complexifie très rapidement avec le problème. Pour cette situation élémentaire où l'on cherche à poser une force unidimensionnelle, trouver une solution est facile. Toutefois, lorsque le problème d'optimisation se complexifie et que plusieurs buts sont exigés en même temps, les solutions analytiques sont de plus en plus difficiles à trouver. D'ailleurs, les limites latérales n'ont même pas été prises en ligne de compte dans le développement de cette approche. Finalement, pour ce genre d'approche, le développement d'une solution est unique au problème donnée, et le changement de même une seule variable peut signifier que le travail de développement de l'algorithme doit être repris du départ. C'est une autre forme de mauvaise généralisation de la méthode.

Un simple mot peut aussi être mentionné pour l'approche de résolution par un agent humain. Cette méthode recquiert généralement le développement parallèle d'un interface personne-machine complexe, coûteux et énergivore à développer. De plus, la rapidité de décision de l'encéphale est à des années-lumières de celle de la machine. En contre-partie, il s'agit de la seule méthode pour laquelle l'agent peut être tenu responsable de ses actions. En effet on ne peut "punir" un algorithme, mais on peut très bien le faire pour des humains. Cette caractéristique est parfois sous-estimée dans le monde de l'informatique et prime même parfois sur la qualité de l'algorithme. En effet, pour tout problème de décisions à haut risque où la décision peut influer sur la vie et la mort d'un individu, par exemple dans le réseau médical ou par le contrôle de véhicule autonome, il est intéressant d'un point de vue social qu'un algorithme de décision (ici l'humain) possède cette qualité pour une maintenance de l'ordre. C'est d'ailleurs un aspect important limitant l'implémentation des algorithmes d'apprentissage dans diverses sphères de la société.

6 Conclusion

La comparaison de trois méthodes de résolution du problème du chariot et du bâton a été étudiée. Chaque approche possède ses avantages et ses défauts, et c'est le travail du scientifique de choisir la méthode appropriée compte tenu du problème. Les méthodes d'apprentissage automatiques ont tendances à être performantes, mais difficilement interprétable. Ici l'algorithme Q-Learning a permis de maintenir le bâton en équilibre pendant 18 ± 4s: une bonne performance, mais hautement variable. Les approches ciblées sont généralement les plus performantes, mais aussi plus difficiles à développer. L'approche physique a permis les meilleures performances de 19.91 ± 0.16s. Enfin, bien que la rapidité de décision humaine est inférieure à celle de la machine, produisant de piètre performance de 1.3 ± 1.4s, elle possède une qualité pour la résolution de résolution de problème qui est inexistante de quelconque algorithme: la responsabilité. Plusieurs angles différents auraient pu être traités pour la résolution du problème. Entre autres, le fléau des hautes dimensions de l'algorithme Q-Learning aurait pu être réduit grandement en utilisant un réseau de neurones, puisque l'augmentation de la complexité de ces derniers augmente linéairement avec le nombre de variables en entrée. Une autre approche intéressante aurait été d'entraîner l'algorithme avec l'énergie du système en entrée (plutôt que les positions et les vitesses) afin de vérifier si l'on peut obtenir les mêmes résultats et ainsi vérifier l'agnostisme et la flexibilité de la méthode Q-Learning.

References

[1] S. Barto, *Reinforcement Learning: An Introduction*, p. Section 3. MIT Press, 2014. https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf.

[2] R. V. Florian, "Correct equations for the dynamics of the cart-pole system," 2005. https://coneural.org/florian/papers/05_cart_pole.pdf.

[3] "Deepmind x ucl rl lecture series - introduction to reinforcement learning,". https://www.youtube.com/watchv=TCCjZe0y4Qc&list=PLqYmG7hTraZDVH599EItlEWsUOsJbAodm.

[4] S. Sudhakar, "Learning rate scheduler,". https://towardsdatascience.com/learning-rate-scheduler-d8a55747dd90.

[5] "How to beat the cartpole game in 5 lines - a simple solution without artificial intelligence,". https://towardsdatascience.com/how-to-beat-the-cartpole-game-in-5-lines-5ab4e738c93f.

[6] Y. Omar, "cartpole.py," *GitHub repository* (2022) . https://github.com/openai/gym/blob/master/gym/envs/classic_control/cartpole.py.

[7] "Gym documentation: Cart pole,". https://www.gymlibrary.dev/environments/classic_control/cart_pole/.

[8] H. Benson, M. Lachance, and M. Séguin, *Physique 1: Mécanique*. ERPI, 2015. https://www.pearsonerpi.com/fr/collegial-universitaire/physique/physique-1-mecanique-dition-en-ligne-monlab-xl-tudiant-6-mois-a37888.

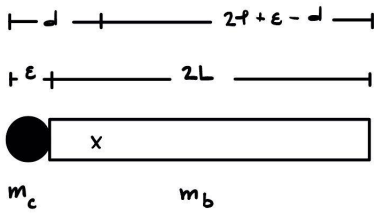
A PSEUDO-CODE (7) (6)

Algorithm 1: Problème Chariot-pôle

Initialiser la matrice Q
for 750 épisodes **do**
  $t \leftarrow 0$
  while $t < 20$ **do**
    if $t = 0$ **then**
      Initialiser le bâton dans une position aléatoire θ_t près du centre avec une vitesse initiale aléatoire ω_t
    end
    $\epsilon \leftarrow \max\{0.01, \min\{1 - \log 5(t+1)\}\}$
    taux $\leftarrow \max\{0.01, \min\{1 - \log 5(t+1)\}\}$
    Selon l'état, choisir une des deux actions (gauche ou droite) possibles $A_t(\theta_t, \omega_t)$ en accord avec la politique ϵ-glouton dans Q.
    Appliquer A_t pendant un temps Δt sur le système pour générer $x_{t+1}, v_{t+1}, \theta_{t+1}$ et ω_{t+1}.
    $R \leftarrow 1$ si $((\theta_{t+1}) \leq seuil$ ou $(x_{t+1}) \leq seuil)$ et 0 sinon
    $Q[x_t, v_t, \theta_t, \omega_t, A_t] \leftarrow (1 - \text{taux})\, Q[x_t, v_t, \theta_t, \omega_t, A_t] + \text{taux} \cdot R \cdot \max\{Q[x_t, v_t, \theta_t, \omega_t]\}$
    $\theta_t \leftarrow \theta_{t+1}$
    $\omega_t \leftarrow \omega_{t+1}$
    $\mathrm{x}_t \leftarrow x_{t+1}$
    $\mathrm{v}_t \leftarrow v_{t+1}$
    $t \leftarrow t + 0.02$
    if $(\theta_t) \geq seuil$ ou $(x_t) \geq seuil$ **then**
      $t \leftarrow 0$
    end
  end
end

B ÉNERGIE DE ROTATION DU SYSTÈME CHARIOT-PÔLE (8)

L'énergie de rotation du système est donnée par $T = \frac{1}{2} I\omega^2$ où I est le moment inertiel du système et ω est la vitesse angulaire dur système ($\omega = \dot{\omega}$). Or, la vitesse angulaire du centre de masse est la même que la vitesse angulaire du bâton parce que le CM est positionné sur le bâton. La complexité réside donc dans la dérivation du moment inertiel de ce système à deux corps. Considérons le chariot comme une masse ponctuelle (de longeur $\epsilon \to 0$) de masse m_c située au bout du bâton de longueur $2L$ et de masse m_b. Soit d, la distance à laquelle se trouve le centre de masse à partir du chariot:



Puisque le moment interniel des corps rigide est donné (8) par

$$I = \int r^2 dm$$

où r est la distance d'un élément infinitésimal de masse dm à l'axe de rotation du système (le CM). Donc,

$$I = \int r^2 dm_c + \int r^2 dm_b$$

On peut ensuite exprimer dm_c et dm_b en fonction de dr:

$$dm_c = \frac{m_c}{\epsilon} dr \qquad\qquad dm_b = \frac{m_b}{2L} dr$$

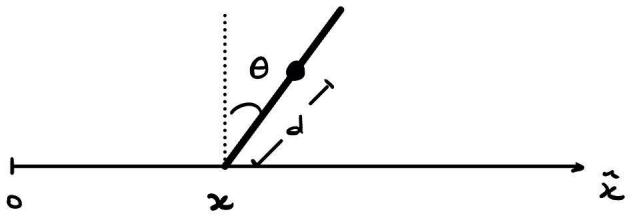
et ainsi résoudre les intégrales:

$$\begin{aligned}
I &= \frac{m_c}{\epsilon} \int_{d-\epsilon}^{d} r^2 dr + \frac{m_b}{2L} \int_0^{d-\epsilon} r^2 dr + \frac{m_b}{2L} \int_0^{2L+\epsilon-d} r^2 dr \\
&= \frac{m_c}{\epsilon} \frac{r^3}{3}\Big|_{d-\epsilon}^{d} + \frac{m_b}{2L} \left(\frac{r^3}{3}\Big|_0^{d-\epsilon} + \frac{r^3}{3}\Big|_0^{2L+\epsilon-d} \right) \\
&= \frac{m_c}{3\epsilon} \left(d^3 - (d-\epsilon)^3 \right) + \frac{m_b}{6L} \left((d-\epsilon)^3 + (2L - (d-\epsilon))^3 \right) \\
&= \frac{m_c}{3\epsilon} \left(d^3 - d^3 - 3d^2\epsilon + 3d\epsilon^2 + \epsilon^3 \right) + \frac{m_b}{6L} \left((2L)^3 - 3(2L)^2(d-\epsilon) + 3(2L)(d-\epsilon)^2 + (d-\epsilon)^3 \right) \\
&= \frac{m_c}{3} \left(-3d^2 + 3d\epsilon + \epsilon^2 \right) + \frac{m_b}{6L} \left(8L^3 - 12L^2(d-\epsilon) + 6L(d-\epsilon)^2 + (d-\epsilon)^3 \right)
\end{aligned}$$

En négligeant tous les termes d'ordre 1 ou supérieur en ϵ:

$$\begin{aligned}
I &= -m_c d^2 + \frac{m_b}{6L} \left(8L^3 - 12L^2 d + 6Ld^2 + d^3 \right) \\
I &= -m_c d^2 + \frac{4}{3} m_b L^2 - 2m_b L d + m_b d^2 + \frac{1}{6} m_b \frac{d^3}{L}
\end{aligned}$$

L'analyse dimensionnelle de chaque terme en kg m^2 correspond à un moment inertiel. Tous les termes sont connus à l'exception de la distance d du centre de masse par rapport au chariot. Comme nous connaissons en tout temps la position du chariot, la longueur et l'angle du bâton, on peut trouver la position du centre de masse en fonction de x, de L et de θ.



Soit $x_c = x$ et $y_c = 0$, la position dans le plan du chariot. Soit x_b et y_b, la position du centre de masse du bâton (attention, ce n'est pas le centre de masse du système). Sachant que le bâton a une longueur de $2L$, de la figure ci-dessus, on a que

$$x_b = x + L\sin\theta \qquad\qquad y_b = L\cos\theta$$

La position du centre de masse est alors:

$$x_{CM} = \frac{m_c x_c + m_b x_b}{m_c + m_b} \qquad y_{CM} = \frac{m_c y_c + m_b y_b}{m_c + m_b}$$
$$x_{CM} = \frac{m_c x + m_b (x + L\sin\theta)}{m_c + m_b} \qquad y_{CM} = \frac{m_b L\cos\theta}{m_c + m_b}$$

où toutes les valeurs sont connues. Ainsi,

$$d = \sqrt{x_{CM}^2 + y_{CM}^2}$$

et le moment inertiel I est ainsi complètement défini. L'énergie en rotation est alors:

$$E_{rot} = \frac{1}{2} I\omega^2$$
$$E_{rot} = \frac{1}{2} I\dot{\theta}^2$$

C ÉNERGIE CINÉTIQUE DU SYSTÈME CHARIOT-PÔLE

L'énergie cinétique est donnée par $E_{cin} = \frac{1}{2} m_{CM} v_{CM}^2$ où $m_{CM} = m_b + m_c$. Quant à la vitesse du centre de masse, rappelons que la position de ce dernier à été calculée en annexe B. En prenant la dérivée temporelle de chaque composante:

$$\dot{x}_{CM} = \frac{d}{dt}\frac{m_c x + m_b (x + L\sin\theta)}{m_c + m_b} \qquad \dot{y}_{CM} = \frac{d}{dt}\frac{m_b L\cos\theta}{m_c + m_b}$$
$$\dot{x}_{CM} = \frac{m_c \dot{x} + m_b (\dot{x} + L\dot{\theta}\cos\theta)}{m_c + m_b} \qquad \dot{y}_{CM} = -\frac{m_b L\dot{\theta}\sin\theta}{m_c + m_b}$$

Et on a finalement que

$$E_{cin} = \frac{1}{2} m_{CM} v_{CM}^2$$
$$E_{rot} = \frac{1}{2}(m_c + m_b)(\dot{x}_{CM}^2 + \dot{y}_{CM}^2)$$