



gz-unitree: Reinforcement learning en robotique avec
validation par moteurs de physique multiples pour le
H1v2 d'Unitree

Gwenn Le Bihan
gwenn.lebihan7@gmail.com
ENSEEIH

16 Novembre 2025

1 Remerciements

Table des matières

1	Remerciements	2
2	Contexte	3
2.1	Bases théoriques du <i>Reinforcement Learning</i>	3
2.1.1	L'entraînement	4
2.1.2	Deep Reinforcement Learning	5
2.1.3	Tendances à la « tricherie » des agents	6
2.1.3.1	Sous-spécification de la fonction coût	6
2.1.3.2	Bug dans l'implémentation de l'environnement	6
2.1.3.3	La validation comme méthode de mitigation	6
2.2	Fonctions coût	6
2.3	Mise à jour	6
2.3.1	<i>Q-learning</i>	6
2.3.2	Évaluation de la performance d'une politique	7
2.3.2.1	Chemins d'états possibles \mathcal{C}	7
2.3.2.2	Récompense attendue η	9
2.3.2.3	Avantage A	9
2.3.2.4	Lien entre η et A	10
2.3.2.5	<i>Surrogate advantage</i> \mathcal{L}	10
2.3.3	<i>Trust Region Policy Optimization</i>	11
2.3.3.1	Distance entre politiques	11
2.3.3.2	Pourquoi faire le maximum sur chaque $s \in S$?	12
2.3.3.3	Région de confiance	12
2.3.4	<i>Proximal Policy Optimization</i>	12
2.3.4.1	Avec pénalité (<i>PPO-Penalty</i>)	13
2.3.4.2	Par <i>clipping</i> (<i>PPO-Clip</i>)	13
2.4	Application en robotique	14
2.4.1	Inventaire des simulateurs en robotique	14
2.4.1.1	Isaac	14
2.4.1.2	MuJoCo	14
2.4.1.3	Gazebo	14
2.4.2	Inventaire des moteurs de simulation physique	14
2.4.2.1	DART	14
2.4.2.2	Bullet	14
2.4.2.3	Bullet avec Featherstone	15
2.5	Le H1v2 d' <i>Unitree</i>	15
2.6	Reproductibilité logicielle	15
3	Packaging reproductible avec Nix	15
3.1	Reproductibilité	15
3.1.1	État dans le domaine de la programmation	15
3.1.2	Contenir les effets de bords	15
3.1.3	État dans le domaine de la robotique	16
3.1.4	Environnements de développement	16

3.2	Nix, le gestionnaire de paquets pur	16
3.2.1	Un <i>DSL</i> ¹ fonctionnel	16
3.2.2	Un écosystème de dépendances	18
3.2.3	Une compilation dans un environnement fixé	18
3.2.3.1	Un complément utile: compiler en CI	18
3.3	NixOS, un système d'exploitation à configuration déclarative	18
3.4	Packaging Nix pour <i>gz-unitree</i>	19
4	Étude du SDK d'Unitree et du bridge SDK \Leftarrow MuJoCo	19
4.1	Une base de code partiellement open-source	19
4.2	Canaux DDS bas niveau	19
4.3	Rétroingénierie des binaires	19
4.4	Un autre bridge existant: <code>unitree_mujoco</code>	19
5	Développement du bridge SDK \Leftarrow Gazebo	19
5.1	Établissement du contact	19
5.2	Réception des commandes	19
5.3	Émission de l'état	19
5.4	Essai sur des politiques réelles	19
5.5	Amélioration des performances	19
5.6	Enregistrement de vidéos	19
5.6.1	Contrôle programmatique de l'enregistrement	19
5.7	Mise en CI/CD	19
5.7.1	Une image de base avec Docker	19
5.7.2	Une pipeline Github Actions	19
	Bibliographie	19
A	Preuves	23
A.1	Cas dégénéré de $D_{\text{KL}}(Q, Q') = 0$ sans utilisation de max	23
A.2	$\eta(p, r)$ comme une espérance	23
A.3	Simplification de l'expression de $L(s, a, \mathcal{P}, \mathcal{P}', R)$ dans PPO-Clip	25

2 Contexte

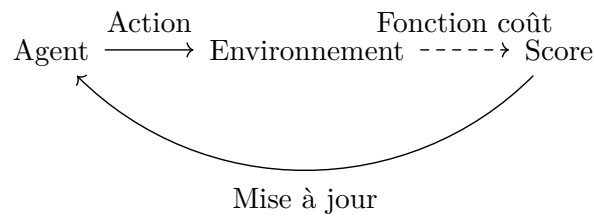
2.1 Bases théoriques du *Reinforcement Learning*

L'apprentissage par renforcement, ou *Reinforcement Learning*, permet de développer des programmes sans expliciter leur logique: on décrit plutôt quatre choses, qui vont permettre à la logique d'émerger pendant la phase d'entraînement:

- Un *agent*: c'est le programme que l'on souhaite créer
- Des *actions* que l'agent peut choisir d'effectuer ou pas
- Un *environnement*, que les actions viennent modifier
- Un *score* (*coût* s'il doit être minimisé, *récompense* inversement) qui dépend de l'état pré- et post-action de l'environnement ainsi que de l'action qui a été effectuée

La phase d'apprentissage consiste à trouver, par des cycles d'essai/erreur, quelles sont les meilleures actions à prendre en fonction de l'environnement actuel, avec meilleur défini comme « qui minimise le coût » (ou maximise la récompense):

¹Domain-Specific Language



Cette technique est particulièrement adaptée aux problèmes qui se prêtent à une modélisation type « jeu vidéo », dans le sens où l'agent représente le personnage-joueur, et le coût un certain score, qui est condition de victoire ou défaite.

En robotique, on a des correspondances claires pour ces quatre notions :

Agent	Robot pour lequel on développe le programme de contrôle (appelée une <i>politique</i>)
Actions	Envoi d'ordres aux moteurs
Environnement	Le monde réel. C'est de loin la partie la plus difficile à simuler informatiquement. On utilise des moteurs de simulation physique, dont la multiplicité des implémentations est importante, voir Chapitre 2.1.3.3
Coût	un ensemble de contraintes (« ne pas endommager le robot »), dont la plupart dépendent de l'objectif de la politique

2.1.1 L'entraînement

(TODO: Expliquer exploration vs exploitation et γ)

Une fois que ce cadre est posé, il reste à savoir *comment* l'on va trouver la fonction qui associe un état de l'environnement à une action.

Une première approche naïve, mais suffisante dans certains cas, consiste à faire une recherche exhaustive et à stocker dans un simple tableau la meilleure action à faire en fonction d'un état de l'environnement :

État actuel (x , retour)	Meilleure action +1 ou -1	Coûts associés
(0, C'est plus)		
(1, C'est plus)		
(3, C'est moins)		
(4, C'est moins)		
(5, C'est moins)		

Tableau 1. – Exemple d'agent à mémoire exhaustive pour un « C'est plus ou c'est moins » dans $\{0, 1, 2\}$, avec pour solution 2

L'entraînement consiste donc ici en l'exploration de l'entièreté des états possibles de l'environnement, et, pour chaque état, le calcul du coût associé à chaque action possible.

Il faut définir la fonction de coût, souvent appelée L pour *loss*:

$$L : E \rightarrow S \quad (1)$$

avec E l'ensemble des états possibles de l'environnement, et S un ensemble muni d'un ordre total (on utilise souvent $[0, 1]$). Ces fonctions coût, qui ne dépendent que de l'état actuel de l'environnement, représente un domaine du RL² appelé *Q-Learning* [1]

On remplit la colonne « Action à effectuer » avec l'action au coût le plus bas:

État actuel (x , retour)	Meilleure action +1 ou -1	Coûts associés avec $L = (x, \text{retour}) \mapsto x - 2 $
(0, C'est plus)	+1	$L(x + 1,) = 2$ $L(x - 1,) = 2$
(1, C'est plus)	+1	$L(x + 1,) = 1$ $L(x - 1,) = 2$
(3, C'est moins)	-1	$L(x + 1,) = 2$ $L(x - 1,) = 3$
(4, C'est moins)	-1	$L(x + 1,) = 3$ $L(x - 1,) = 4$
(5, C'est moins)	-1	$L(x + 1,) = 4$ $L(x - 1,) = 5$

Tableau 2. – Entraînement terminé, avec pour fonction coût L la distance à la solution

Ici, cette approche exhaustive suffit parce que l'ensemble des états possibles de l'environnement, E , possède 6 éléments

Cependant, ces ensembles sont bien souvent prohibitivement grands (e.g. $x \in \llbracket 0, 10^{34} \rrbracket$), infinis ($x \in \mathbb{N}$) ou indénombrables ($x \in \mathbb{R}$)

Dans le cas de la robotique, E est une certaine représentation numérique du monde réel autour du robot, on imagine donc bien qu'il y a beaucoup trop d'états possibles.

2.1.2 Deep Reinforcement Learning

Une façon de remédier à ce problème de dimensions est de remplacer le tableau exhaustif par un réseau de neurones:

État actuel	devient la couche d'entrée
Meilleure action	devient la couche de sortie
Coûts associés	deviennent les neurones des couches cachées
Le remplissage du tableau	devient la rétropropagation pendant l'entraînement

²Reinforcement Learning

2.1.3 Tendances à la « tricherie » des agents

Expérimentalement, on sait que des tendances « tricheuses » émergent facilement pendant l'entraînement [Réf. nécessaire]: l'agent découvre des séries d'actions qui causent un bug avantageux vis à vis du coût associé, soit parce qu'il y a un bug dans le calcul de l'état de l'environnement post-action, soit parce que la fonction coût ne prend pas suffisamment bien en compte toutes les possibilités de l'environnement (autrement dit, il manque de contraintes).

Sous-spécification de la fonction coût

(Note: Bof cette partie)

Un exemple populaire est l'expérience de pensée du Maximiseur de trombones [2]: un agent avec pour environnement le monde réel, pour actions « prendre des décisions »; « envoyer des emails »; etc. et pour fonction récompense (une fonction à maximiser au lieu de minimiser) « le nombre de trombones existant sur Terre », finirait possiblement par réduire en escalavage tout être vivant capable de produire des trombones: la fonction coût est sous-spécifiée

Bug dans l'implémentation de l'environnement

Bien évidemment, pour l'agent, tant qu'un bug n'est pas explicitement découragé par sa prise en compte dans la fonction coût. Si une action est favorable à l'amélioration du score, l'agent la prendra.

La validation comme méthode de mitigation (Note: ça se dit mitigation en français?)

Comme ces bugs sont des comportements non voulus, il est très probables qu'ils ne soient pas exactement les mêmes d'implémentation à implémentation du même environnement.

Il convient donc de se servir de *plusieurs* implémentations: un sert à la phase d'entraînement, pendant laquelle l'agent développe des « tendances à la tricherie », puis une phase de *validation*.

Cette phase consiste en le lancement de l'agent dans une autre implémentation, avec les mêmes actions mais qui, crucialement, ne comporte pas les mêmes bugs que l'environnement ayant servi à la phase d'apprentissage.

Les « techniques de triche » ainsi apprises deviennent inefficace, et si le score (le coût ou la récompense) devient bien pire que pendant l'apprentissage, on peut détecter les cas de triche.

On peut même aller plus loin, et multiplier les phases de validation avec des implémentations supplémentaires, ce qui réduit encore la probabilité qu'une technique de triche se glisse dans l'agent final

(Note: Rien à voir mais je me dis, c'est en fait un moyen de trouver des bugs dans un physics engine ! ça me fait penser au Fuzzing un peu, mais avec un NN plutôt que du hasard contrôlé)

2.2 Fonctions coût

2.3 Mise a jour

2.3.1 Q-learning

Le score associé à un état s_t et une action a_t , appelée $Q(s_t, a_t)$ ici pour « quality » [3], est mis à jour avec cette valeur [4]:

$$(1 - \alpha) \underbrace{Q(s_t, a_t)}_{\text{valeur actuelle}} + \alpha \left(\underbrace{R_{t+1}}_{\substack{\text{récompense} \\ \text{pour cette action}}} + \gamma \underbrace{\max_a Q(S_{t+1}, a)}_{\substack{\text{récompense de la meilleure} \\ \text{action pour l'état suivant}}} \right) \quad (2)$$

L'expression comporte deux hyperparamètres:

Learning rate α contrôle à quel point l'on favorise l'évolution de Q ou pas.

Discount factor γ contrôle l'importance que l'on donne aux récompenses futures. Il est utile de commencer avec une valeur faible puis l'augmenter avec le temps [5].

2.3.2 Évaluation de la performance d'une politique

Théoriquement, le « score » associé à un couple état/action est souvent réduit à l'intervalle $[0, 1]$ et assimilé à une distribution de probabilité: Q est une fonction de $S \times A$ vers $[0, 1]$ qui renvoie la probabilité qu'a l'agent à choisir une action en étant dans un état de l'environnement.

On note dans le reste de cette section:

A	l'ensemble des actions
S	l'ensemble des états possibles de l'environnement
$\rho_0 : S \rightarrow [0, 1]$	la distribution de probabilité de l'état initial de l'environnement. Si l'on initialise l'environnement de manière uniformément aléatoire, ρ_0 est une équiprobabilité ³
$M : S \times A \rightarrow S$	le moteur de simulation physique, qui applique l'action à un état de l'environnement et envoie le nouvel état de l'environnement
$\mathcal{P} : S \rightarrow A$	une politique
$\mathcal{P}^* : S \rightarrow A$	la meilleure politique possible, celle que l'on cherche à approcher
$R : S \rightarrow \mathbb{R}^+$	sa fonction de récompense
$Q_p : S \times A \rightarrow [0, 1]$	sa distribution de probabilité, qu'on suppose Markovienne (elle ne dépend que de l'état dans lequel on est). $Q_p(s_t, a_t)$ est la probabilité que p choisisse a_t quand on est dans l'état s_t (s_t est l'état pré -action, et non post-action)
Q et Q^*	$Q_{\mathcal{P}}$ et $Q_{\mathcal{P}^*}$, pour alléger les notations

On suppose A et S dénombrables⁴.

Pour alléger les notations, on surchargera les fonctions récompenses pour qu'elle puissent prendre en entrée des éléments de $S \times A$, en ignorant simplement l'action choisie:

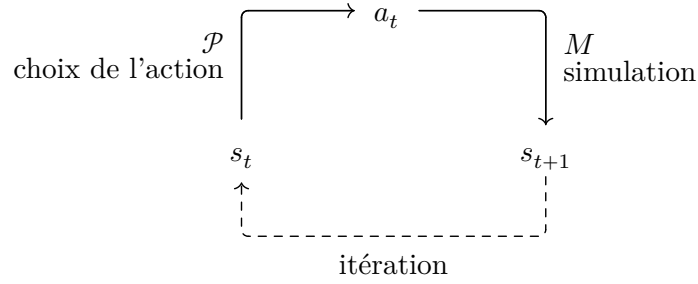
$$\forall (s, a) \in S \times A, \forall r \in \text{récompenses}, r(s, a) := r(s) \quad (3)$$

Chemins d'états possibles \mathcal{C}

³i.e. $\text{card } \rho_0(S) = 1$

⁴En pratique, \mathbb{R} est discrétisé dans les simulateurs numérique, donc cette hypothèse ne pose pas de problèmes à l'application de la théorie au domaine de la robotique

M et \mathcal{P} forment en fait tout se qui se passe pendant un pas de temps, c'est cette boucle que l'on répète pour soit entraîner l'agent (si l'on met \mathcal{P} à jour à chaque tour de boucle) ou l'utiliser:



Quand on « déroule » \mathcal{P} en en partant d'un certain état initial s_0 , on obtient une suite d'états et d'actions:

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} \dots$$

Pour tout pas de temps $t \in \mathbb{N}$, on a:

$$\begin{cases} a_t = \mathcal{P}(s_t) \\ s_{t+1} = M(s_t, a_t) \end{cases} \quad (4)$$

Un chemin se modélise aisément par une suite d'éléments de $S \times A$. Ainsi, on note

(Note: p-ê Expliquer pourquoi une suite de S en fait ça marche pas, en gros on choppe pas tt les chemins possible psk faut trouver a en fonction de p donc ya pas tout. Si on prend $p(a)$ c'est que le chemin que la politique prendrait)

$$\mathcal{C}_p := \left\{ (s_t, a_t)_{t \in \mathbb{N}} \text{ avec } \begin{cases} a_0 = p(s_0) \\ \forall t \in \mathbb{N} \quad a_{t+1} = p(s_{t+1}) \\ \forall t \in \mathbb{N} \quad s_{t+1} = M(s_t, a_t) \end{cases} \middle| s_0 \in S \right\} \quad (5)$$

l'ensemble des chemins possibles avec la politique p . C'est tout simplement l'ensemble de tout les « déroulements » de la politique p en partant des états possibles de l'environnement.

On définit également l'ensemble de *tout* les chemins d'états possibles, peut importe la politique, \mathcal{C} :

$$\mathcal{C} := \left\{ \left\{ \begin{cases} c_0 = (s_0, a_0) \\ \forall t \in \mathbb{N} \quad c_{t+1} = M(c_t) \end{cases} \middle| (s_0, a) \in S \times A^{\mathbb{N}} \right\} \right\} \quad (6)$$

On notera que, selon M , on peut avoir $\mathcal{C} \subsetneq (S \times A)^{\mathbb{N}}$: par exemple, certains états de l'environnement peuvent représenter des « impasses », où il est impossible d'évoluer vers un autre état, peu importe l'action choisie.

On note aussi que \mathcal{C} (et donc \mathcal{C}_p aussi) est dénombrable, étant construit à partir de $(S \times A)^{\mathbb{N}}$ et S , A et \mathbb{N} étant aussi dénombrables⁵

*Cette formalisation est utile par la suite,
pour proprement définir certaines grandeurs.*

(Note: pas sûre de cette phrase)

Récompense attendue η

η représente la récompense moyenne à laquelle l'on peut s'attendre pour une politique p avec fonction de récompense r .

Elle prend en compte le *discount factor* γ : les récompenses des actions deviennent de moins en moins⁶ importantes avec le temps. η est définie ainsi [6]

$$\eta(p, r) = \underbrace{\sum_{(c_t)_{t \in \mathbb{N}} \in \mathcal{S}} \underbrace{\rho_0(s_0) \prod_{t=0}^{\infty} Q_p(c_t)}_{\text{probabilité du chemin}} \underbrace{\sum_{t=0}^{\infty} \gamma^t r(c_t)}_{\text{récompense associée}}}_{\text{pour tout chemin possible}} \quad (7)$$

On peut également exprimer $\eta(p, r)$ comme une espérance. Soit C une variable aléatoire de \mathcal{S} . On a (cf preuve en A.2)

$$\eta(p, r) = \mathbb{E} \left(\sum_{t=0}^{\infty} \gamma^t r(C_t) \right) \quad (8)$$

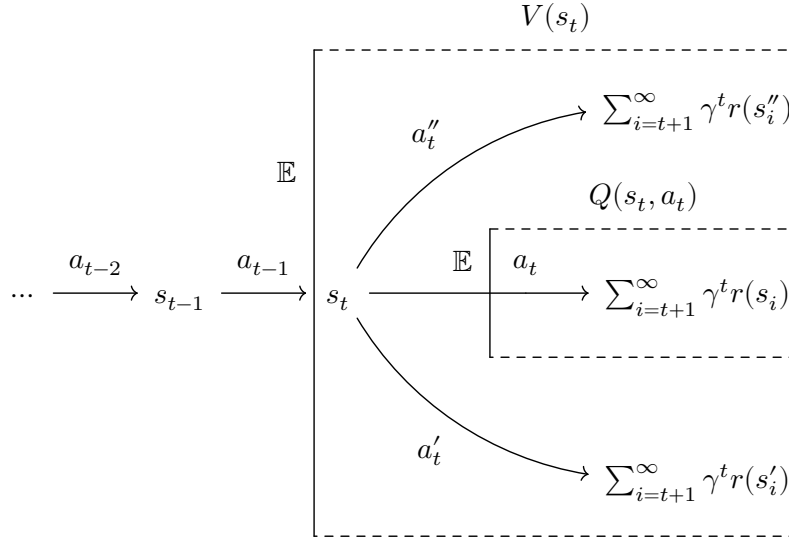
Avantage A

L'avantage $A_{p,r}(s, a)$ mesure à quel point il est préférable de choisir l'action a quand on est dans l'état s (pour la politique p , avec « préférable » au sens de $(r(S), \geq)$)

On peut visualiser ce calcul ainsi:

⁵On a $\text{card } \mathcal{C} \leq \text{card}((S \times A)^{\mathbb{N}}) = \text{card}(S \times A)^{\text{card } \mathbb{N}} = (\text{card } S \text{ card } A)^{\text{card } \mathbb{N}} \leq (\aleph_0)^{\text{card } \mathbb{N}} = 2^{\aleph_0} = \aleph_0$

⁶En supposant $\gamma < 1$, ce qui est souvent le cas [Réf. nécessaire] (TODO: Mettre dans la def de γ)



Pour calculer $A_{p,r}(s, a)$, on regarde l'espérance des récompenses cumulées pour tout chemin commençant par s , et on la compare à celle pour tout chemin commençant par $M(s, a)$

$$A_{p,r}(s, a) := \underbrace{\mathbb{E}_{\substack{(s_t, a_t)_{t \in \mathbb{N}} \sim p \in \mathcal{S} \\ s_0 = s \\ s_1 = M(s_0, a)}}}_{Q(s, a)} \sum_{t=0}^{\infty} \gamma^t r(s_t) - \underbrace{\mathbb{E}_{\substack{(s_t, a_t)_{t \in \mathbb{N}} \sim p \in \mathcal{S} \\ s_0 = s}}}_{V(s)} \sum_{t=0}^{\infty} \gamma^t r(s_t) \quad (9)$$

On considère tout les chemins à partir de l'état s_t , et l'on regarde l'espérance...

pour $V(s_t)$ de tout les chemins

pour $Q(s_t, a_t)$ du chemin où l'on a choisi a_t

En suite, il suffit de faire la différence, pour savoir l'*avantage* que l'on a à choisir a_t par rapport au reste.

Lien entre η et A

Pour une fonction de récompense r donnée, A permet de calculer η pour une politique p' en fonction de la valeur de η pour une autre politique p' [7]

$$\begin{aligned} \eta(p', r) &= \eta(p, r) + \mathbb{E}_{(c_t)_{t \in \mathbb{N}} \sim p' \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t A_{p,r}(c_t) \\ &\text{Qui se simplifie en [6]} \\ &= \eta(p, r) + \sum \end{aligned} \quad (10)$$

Surrogate advantage \mathcal{L}

Il est théoriquement possible d'utiliser A pour optimiser une politique, en maximisant sa valeur à un état donné:

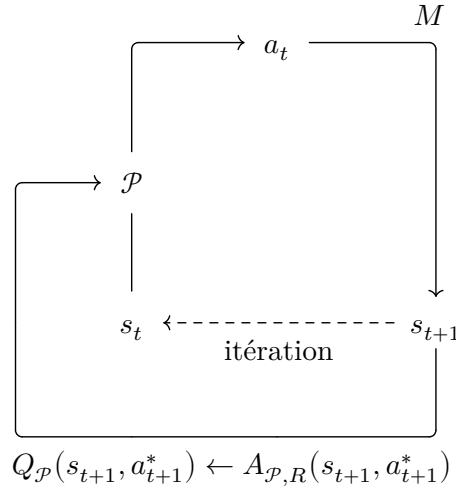


Fig. 5. – Boucle d'entraînement

Avec

$$a_{t+1}^* := \operatorname{argmax}_{a \in A} A_{\mathcal{P},R}(s_{t+1}, a) \quad (11)$$

Mais, en pratique, des erreurs d'approximations peuvent rendre $A_{\mathcal{P},R}(s_{t+1}, a_{t+1}^*)$ négatif, ce qui empêche de s'en servir pour définir une valeur de $Q_{\mathcal{P}}$ [6]

Le *surrogate advantage* détermine la performance d'une politique par rapport à une autre

$$\mathcal{L}_r(p', p) := \mathbb{E}_{(s_t, a_t)_{t \in \mathbb{N}} \in \mathcal{C}} \sum_{t=0}^{\infty} \frac{Q_p(s_t, a_t)}{Q_{p'}(s_t, a_t)} A_{p,r}(s_t, a_t) \quad (12)$$

2.3.3 Trust Region Policy Optimization

La méthode TRPO définit la mise à jour de Q avec un Q' qui maximise le *surrogate advantage* [8], sous une contrainte limitant l'écart entre Q et Q'

L'idée de la *TRPO* est de maximiser le *surrogate advantage* du nouveau Q tout en limitant l'ampleur des modifications apportées à Q , ce qui procure une stabilité à l'algorithme, et évite qu'un seul « faux pas » dégrade violemment la performance de la politique.

$$Q' = \begin{cases} \operatorname{argmax}_q \mathcal{L}_r(q, Q) \\ \text{s.c. distance}(Q', Q) < \delta \end{cases} \quad (13)$$

Avec δ une limite supérieure de distance entre Q' , la nouvelle politique, et Q , l'ancienne.

Distance entre politiques

Il existe plusieurs manières de mesurer l'écart entre deux distributions de probabilité, dont notamment la *divergence de Kullback-Leibler*, aussi appelée entropie relative [9], [10]:

$$D_{\text{KL}}(P \parallel P') := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{P'(x)} \quad (14)$$

Avec \mathcal{X} l'espace des échantillons et P, P' deux distributions de probabilité sur celui-ci. Dans notre cas, $\mathcal{X} = S \times A$,

Pour évaluer cette distance, on regarde la plus grande des distances entre des paires de distributions de probabilité de politiques $Q_{\mathcal{P}}$ et $Q_{\mathcal{P}'}$ pour $s \in S$ fixé [6]

$$\max_{s \in S} D_{\text{KL}}(Q_{\mathcal{P}'}(s, \cdot) \parallel Q_{\mathcal{P}}(s, \cdot)) < \delta \quad (15)$$

En notant $Q_p(s, \cdot) := a \mapsto Q_p(s, a)$. On a donc ici « $\mathcal{X} = A$ » dans la définition de D_{KL}

Pourquoi faire le maximum sur chaque $s \in S$?

Ce maximum revient à limiter non pas la simple distance entre les deux politiques, mais *limiter la modification de la politique sur chacune de ses actions*.

(Note: C'est ma théorie ça, faudrait être sûr que le papier ne donne pas d'explications)

Ceci permet d'éviter d'avoir deux politiques jugées similaires par D_{KL} à cause de modifications se « compensant » [Réf. nécessaire]. Par exemple, avec

$$\forall s \in S, Q(s, 1) = Q(s, 2) \quad (16)$$

et

$$Q' := (s, a) \mapsto \begin{cases} Q(s, a) \cdot 2 & \text{si } a = 1 \\ Q(s, a) \cdot \frac{1}{2} & \text{si } a = 2 \\ Q(s, a) & \text{sinon} \end{cases} \quad (17)$$

On a $D_{\text{KL}}(Q, Q') = 0$ (cf preuve en A.1), alors qu'il y a eu une modification très importante des probabilités de choix de l'action 1 et 2 dans tous les états possibles : si on imagine $Q(s, 1) = Q(s, 2) = 1/4$, on a après modification $Q'(s, 1) = 1/2$ et $Q'(s, 2) = 1/8$.

Région de confiance

Cette contrainte définit un ensemble réduit de \mathcal{P}' acceptables comme nouvelle politique, aussi appelé une *trust region* (région de confiance), d'où la méthode d'optimisation tire son nom [6].

En pratique, l'optimisation sous cette contrainte est trop demandeuse en puissance de calcul, on utilise plutôt l'espérance [6]

$$\overline{D_{\text{KL}}} := \mathbb{E}_{s \in S} D_{\text{KL}}(Q(s, \cdot) \parallel Q'(s, \cdot)) \quad (18)$$

2.3.4 Proximal Policy Optimization

La *PPO* repose sur le même principe de stabilisation de l'entraînement par limitation de l'ampleur des changements de politique à chaque pas.

Cependant, les méthodes *PPO* préfèrent changer la quantité à optimiser, pour limiter intrinsèquement l'ampleur des modifications, en résolvant un problème d'optimisation sans contraintes [11]

$$\begin{aligned} \arg\max_{\mathcal{P}'} \quad & \mathbb{E}_{(s,a) \in \mathcal{S}} L(s, a, \mathcal{P}, \mathcal{P}', R) \\ \text{s.c. } & \top \end{aligned} \quad (19)$$

Avec pénalité (*PPO-Penalty*)

PPO-Penalty soustrait une divergence K-L pondérée à l'avantage:

$$L(s, a, \mathcal{P}, \mathcal{P}', R) = \frac{Q_{\mathcal{P}'}(s, a)}{Q_{\mathcal{P}}(s, a)} A_{\mathcal{P}', R}(s, a) - \beta D_{\text{KL}(\mathcal{P} \parallel \mathcal{P}')} \quad (20)$$

Avec β ajusté automatiquement pour

Par *clipping* (*PPO-Clip*)

PPO-Clip utilise une limitation du ratio de probabilités (en minimum et en maximum) [12]

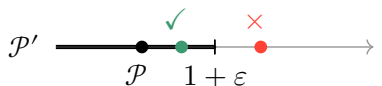
$$\begin{aligned} L(s, a, \mathcal{P}, \mathcal{P}', R) = \min & \left(\frac{Q_{\mathcal{P}'}(s, a)}{Q_{\mathcal{P}}(s, a)} A_{\mathcal{P}', R}(s, a), \right. \\ & \left. \text{clip} \left(\frac{Q_{\mathcal{P}'}(s, a)}{Q_{\mathcal{P}}(s, a)}, 1 - \varepsilon, 1 + \varepsilon \right) A_{\mathcal{P}', R}(s, a) \right) \end{aligned} \quad (21)$$

Avec $\varepsilon \in \mathbb{R}_+^*$ un paramètre indiquant à quel point l'on peut s'écarter de la politique précédente, et

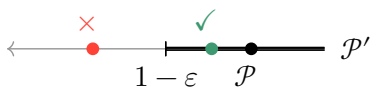
$$\text{clip} := (x, m, M) \mapsto \begin{cases} m & \text{si } x < m \\ M & \text{si } x > M \\ x & \text{sinon} \end{cases} \quad (22)$$

La complexité de l'expression, et la présence d'un min au lieu de simplement un clip est due au fait que l'avantage $A_{\mathcal{P}', R}(s, a)$ peut être négatif. L'expression se simplifie en séparant les cas (cf preuve en A.3)

Si l'avantage est positif a est un meilleur choix que $\mathcal{P}(s)$.

$$L(s, a, \mathcal{P}, \mathcal{P}', R) = \min \left(\frac{Q_{\mathcal{P}'}(s, a)}{Q_{\mathcal{P}}(s, a)}, 1 + \varepsilon \right) A_{\mathcal{P}', R}(s, a)$$


Si l'avantage est négatif choisir a est pire que garder $\mathcal{P}(s)$.

$$L(s, a, \mathcal{P}, \mathcal{P}', R) = \max \left(1 - \varepsilon, \frac{Q_{\mathcal{P}'}(s, a)}{Q_{\mathcal{P}}(s, a)} \right) A_{\mathcal{P}', R}(s, a)$$


2.4 Application en robotique

Dans le contexte de la robotique, le calcul de l'état post-action de l'environnement est le travail du *moteur de physique*.

Bien évidemment, ce sont des programmes complexes avec souvent des numériques souvent numériques d'équation physiques; il est presque inévitable que des bugs se glissent dans ces programmes.

Un environnement de RL⁷ ne se résume pas à son moteur de physique: il faut également charger des modèles 3D, le modèle du robot (qui doit être contrôlable par les actions), et également, pendant les phases de développement, avoir un moteur de rendu graphique, une interface et des outils de développement.

Cet ensemble s'appelle un *simulateur*.

2.4.1 Inventaire des simulateurs en robotique

Isaac

Un simulateur développé par NVIDIA [13], utilisant son propre moteur de rendu, PhysX [14]

MuJoCo

Un simulateur initialement propriétaire. Il a été rendu gratuit puis open source par Google DeepMind [15].

Bien que MuJoCo est décrit comme un moteur de simulation physique et non un simulateur, il embarque une commande `simulate` qui le rend fonctionnellement équivalent à un simulateur [16].

Gazebo

Les intérêts de Gazebo [17] sont multiples:

- C'est un logiciel open-source *communautaire*, qui ne dépend pas du financement d'une grande entreprise
- Son architecture modulaire permet notamment d'utiliser plusieurs moteurs de simulation physique différents [18], à l'inverse de MuJoCo.

Gazebo possède des plugins officiels pour:

DART Plugin `gz-physics-dartsim-plugin`, c'est l'implémentation principale, et celle par défaut [18].

Bullet Plugin `gz-physics-bulletsim-plugin`. En beta [18].

Bullet Featherstone Plugin `gz-physics-bullet-featherstone-plugin`, également en beta [18].

2.4.2 Inventaire des moteurs de simulation physique

DART

DART, pour Dynamic Animation and Robotics Toolkit [19],

Bullet

⁷Reinforcement Learning

Bullet [20], [21]

Bullet avec Featherstone

L'algorithme de Featherstone [22], servant d'implémentation alternative à Bullet [23]

2.5 Le H1v2 d'*Unitree*

2.6 Reproductibilité logicielle

La reproductibilité est particulièrement complexe dans le champ du reinforcement learning [24]

3 Packaging reproductible avec Nix

3.1 Reproductibilité

3.1.1 État dans le domaine de la programmation

La différence entre une fonction au sens mathématique et une fonction au sens programmatique consiste en le fait que, par des raisons de praticité, on permet aux **fonctions** des langages de programmation d'avoir des *effets de bords*. Ces effets affectent, modifient ou font dépendre la fonction d'un environnement global qui n'est pas explicitement déclaré comme une entrée (un argument) de la fonction en question [25].

Cette liberté permet, par exemple, d'avoir accès à la date et à l'heure courante, interagir avec un système de fichier d'un ordinateur, générer une surface pseudo aléatoire par bruit de Perlin, etc.

Mais, en contrepartie, on perd une équation qui est fondamentale en mathématiques:

$$\forall E, F, \forall f : E \rightarrow F, \forall (e_1, e_2) \in E^2, e_1 = e_2 \Rightarrow f(e_1) = f(e_2) \quad (23)$$

En programmation, on peut très facilement construire un f qui ne vérifie pas ceci:

```
from datetime import date

def f(a):
    return date.today().year + a
```

Selon l'année dans laquelle nous sommes, $f(0)$ n'a pas la même valeur.

De manière donc très concrète, si cette fonction f fait partie du protocole expérimental d'une expérience, cette expérience n'est plus reproductible, et ses résultats sont donc potentiellement non vérifiables, si le papier est soumis le 15 décembre 2025 et la *peer review* effectuée le 2 janvier 2026.

3.1.2 Contenir les effets de bords

En dehors du besoin de vérifiabilité du monde de la recherche, la reproductibilité est une qualité recherchée dans certains domaines de programmation [26]

Il existe donc depuis longtemps des langages de programmation dits *fonctionnels*, qui, de manière plus ou moins stricte, limite les effets de bords. Certains langages font également la distinction entre une fonction *pure*⁸ et une fonction classique [27]. Certaines fonctions, plutôt appelées *procédures*, sont uniquement composées d'effet de bord puisqu'elle ne renvoie pas de valeur [28]

⁸sans effets de bord

3.1.3 État dans le domaine de la robotique

En robotique, pour donner des ordres au matériel, on interagit beaucoup avec le monde extérieur (ordres et lecture d'état de servo-moteurs, flux vidéo d'une caméra, etc), souvent dans un langage plutôt bas-niveau, pour des questions de performance et de proximité abstraitionnelle au matériel.

De fait, les langages employés sont communément C, C++ ou Python⁹ [29], des langages bien plus impératifs que fonctionnels [30].

L'idée de s'affranchir d'effets de bords pour rendre les programmes dans la recherche en robotique reproductibles est donc plus utopique que réaliste.

3.1.4 Environnements de développement

Cependant, ce qui fait un programme n'est pas seulement son code: surtout dans des langages plus anciens sans gestion de dépendance intégrée au langage, les dépendances (bibliothèques) du programme, ainsi que l'environnement et les étapes de compilation de ce dernier, représentent également une partie considérable de la complexité du programme (par exemple, en C++, on utilise un outil générant des fichiers de configuration pour un autre outil qui à son tour configure le compilateur de C++ [31])

C'est cette partie que Nix, le gestionnaire de paquet, permet d'encapsuler et de rendre reproductible. Dans ce modèle, la compilation (et de manière plus générale la construction, ou *build*) du projet est la fonction que l'on veut rendre pure. L'entrée est le code source, et le résultat de la fonction est un binaire, qui ne doit dépendre que du code source.

$$\forall \text{src}, \text{bin}, \forall f \in \text{bin}^{\text{src}}, \forall (P_1, P_2) \in \text{src}^2, P_1 = P_2 \Rightarrow f(P_1) = f(P_2) \quad (24)$$

Ici, P_1 et P_2 sont deux itérations du code source (src) du programme. Si le code source est identique, les binaires résultants de la compilation (f) sont égaux, au sens de l'égalité bit à bit.

On a la proposition (1), avec $E = \text{src}$, l'ensemble des code source possibles pour un langage, et $F = \text{bin}$, l'ensemble des binaires exécutable

Nix ne peut pas garantir que le programme sera sans effets de bords au *runtime*, mais vise à le garantir au *build-time*.

3.2 Nix, le gestionnaire de paquets pur

3.2.1 Un *DSL*¹⁰ fonctionnel

Une autre caractéristique que l'on trouve souvent dans la famille de langages fonctionnels est l'omniprésence des *expressions*: quasi toute les constructions syntaxiques forment des expressions valides, et peuvent donc servir de valeur

<pre>def g(x, y): if y == 5: x = 6 else: x = 8 return f(x)</pre>	<pre>let g x y = f (if y = 5 then 6 else 8)</pre>
--	---

⁹Il arrive assez communément d'utiliser Python, un langage haut-niveau, mais c'est dans ce cas à but de prototypage, et le code contrôlant les moteurs est écrit dans un langage bas niveau puis appelé par Python par FFI.

¹⁰Domain-Specific Language

Python (<code>if</code> et <code>else</code> sont des instructions)	OCaml (<code>if</code> et <code>else</code> forment une expression)
---	---

Afin de décrire les dépendances d'un programme, l'environnement de compilation, et les étapes pour le compiler (en somme, afin de définir le $f \in \text{bin}^{\text{src}}$), Nix comprend un langage d'expressions [32]. Un fichier `.nix` définit une fonction, que Nix sait exécuter pour compiler le code source.

Expression d'une fonction en Python	En Nix
<code>lambda f(a): a + 3</code>	<code>{ a }: a + 3</code>

Voici un exemple de définition d'un programme, appelée *dérivation* dans le jargon de Nix:

```
{
  src-odri-masterboard-sdk,

  lib,
  stdenv,
  jrl-cmakemodules,
  cmake,
  python3Packages,
  catch2_3,
}:

stdenv.mkDerivation {
  pname = "odri_master_board_sdk";
  version = "1.0.7";

  src = src-odri-masterboard-sdk;

  preConfigure = ''
    cd sdk/master_board_sdk
  '';

  doCheck = true;

  cmakeFlags = [
    (lib.cmakeBool "BUILD_PYTHON_INTERFACE" stdenv.hostPlatform.isLinux)
  ];

  nativeBuildInputs = [
    jrl-cmakemodules
    python3Packages.python
    cmake
  ];

  buildInputs = with python3Packages; [ numpy ];

  nativeCheckInputs = [ catch2_3 ];

  propagatedBuildInputs = with python3Packages; [ boost ];
}
```

La dérivation ici prend en entrée le code source (`src-odri-masterboard-sdk`), ainsi que des dépendances, que ce soit des fonctions relatives à Nix même (comme `stdenv.mkDerivation`) pour simplifier la définition de dérivation, ou des dépendances au programmes, que ce soit pour sa compilation ou pour son exécution (dans ce dernier cas de figures, les dépendances sont incluses ou reliées au binaire final)

3.2.2 Un écosystème de dépendances

Afin de conserver la reproductibilité même lorsque l'on dépend de bibliothèques tierces, ces dépendances doivent également avoir une compilation reproductible: on déclare donc des dépendances à des *packages* Nix, disponibles sur *Nixpkgs* [33].

Parfois donc, écrire un paquet Nix pour son logiciel demande aussi d'écrire les paquets Nix pour les dépendances de notre projet, si celles-ci n'existent pas encore, et cela récursivement. On peut ensuite soumettre nos paquets afin que d'autres puissent en dépendre sans les réécrire, en contribuant à *Nixpkgs* [34]

Pour ne pas avoir à compiler toutes les dépendances soit-même quand on dépend de `.nix` de *nixpkgs*, il existe un serveur de cache, qui propose des binaires des dépendances, Cachix [35]

3.2.3 Une compilation dans un environnement fixé

Certains aspects de l'environnement dans lequel l'on compile un programme peuvent faire varier le résultat final. Pour éviter cela, Nix limite au maximum les variations d'environnement. Par exemple, la date du système est fixée au 0 UNIX (1er janvier 1990): le programme compilé ne peut pas dépendre de la date à laquelle il a été compilé.

Quand le *sandboxing* est activé, Nix isole également le code source de tout accès au réseau, aux autres fichiers du système (ainsi que d'autres mesures) pour améliorer la reproductibilité [36]

Un complément utile: compiler en CI

Pour aller plus loin, on peut lancer la compilation du paquet Nix en *CI*¹³, c'est-à-dire sur un serveur distant au lieu de sur sa propre machine. On s'assure donc que l'état de notre machine de développement personnelle n'influe pas sur la compilation, puisque chaque compilation est lancée dans une machine virtuelle vierge [37].

3.3 NixOS, un système d'exploitation à configuration déclarative

Une fois le programme compilé avec ses dépendances, il est prêt à être transféré sur l'ordinateur ou la carte de contrôle embarquée au robot.

Lorsqu'il y a un ordinateur embarqué, comme par exemple une Raspberry Pi [38], il faut choisir un OS sur lequel faire tourner le programme.

Là encore, un OS s'accompagne d'un amas considérable de configuration des différentes parties du système: accès au réseau, drivers,...

Sur les OS Linux classiques tels que Ubuntu ou Debian, cette configuration est parfois stockée dans des fichiers, ou parfois retenue en mémoire, modifiée par l'exécution de commandes.

C'est un problème assez récurrent dans Linux de manière générale: d'un coup, le son ne marche plus, on passe ½h sur un forum à copier-coller des commandes dans un terminal, et le problème est réglé... jusqu'à ce qu'il survienne à nouveau après un redémarrage ou une réinstallation.

Ici, NixOS assure que toute modification de la configuration d'un système est *déclarée* (d'où l'adjectif « déclaratif ») dans des fichiers de configurations, également écrits dans des fichiers `.nix` [39].

Ici encore, cela apporte un gain en terme de reproductibilité: l'état de configuration de l'OS sur lequel est déployé le programme du robot est, lui aussi, rendu reproductible.

¹³Continuous Integration, lit. intégration continue

3.4 Packaging Nix pour *gz-unitree*

(TODO: Faire cette partie)

4 Étude du SDK d'Unitree et du bridge SDK \leftrightarrow MuJoCo

4.1 Une base de code partiellement open-source

4.2 Canaux DDS bas niveau

4.3 Rétroingénierie des binaires

4.4 Un autre bridge existant: `unitree_mujoco`

5 Développement du bridge SDK \leftrightarrow Gazebo

5.1 Établissement du contact

5.2 Réception des commandes

5.3 Émission de l'état

5.4 Essai sur des politiques réelles

5.5 Amélioration des performances

5.6 Enregistrement de vidéos

5.6.1 Contrôle programmatique de l'enregistrement

5.7 Mise en CI/CD

5.7.1 Une image de base avec Docker

5.7.2 Une pipeline Github Actions

Bibliographie

- [1] Shengbo Eben Li, *Reinforcement Learning for Sequential Decision and Optimal Control*. Springer Singapore, p. 1-460. doi: [10.1007/978-981-19-7784-8](https://doi.org/10.1007/978-981-19-7784-8).
- [2] Nick Bostrom, « Ethical Issues in Advanced Artificial Intelligence », 2003, *Int. Institute of Advanced Studies in Systems Research and Cybernetics*. Consulté le: 8 octobre 2025. [En ligne]. Disponible sur: <https://nickbostrom.com/ethics/ai>

- [3] Tabet Matiisen, « Demystifying deep reinforcement learning », 19 décembre 2015, *Computational Neuroscience Research Group at University of Tartu*. Consulté le: 13 octobre 2025. [En ligne]. Disponible sur: <https://web.archive.org/web/20180407053740/http://neuro.cs.ut.ee/demystifying-deep-reinforcement-learning/>
- [4] T. G. Dietterich, « Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition », *CoRR*, 1999, Consulté le: 2002. [En ligne]. Disponible sur: <https://arxiv.org/abs/cs/9905014>
- [5] V. François-Lavet, R. Fonteneau, et D. Ernst, « How to Discount Deep Reinforcement Learning: Towards New Dynamic Strategies », *CoRR*, 2015, Consulté le: 13 octobre 2025. [En ligne]. Disponible sur: <http://arxiv.org/abs/1512.02011>
- [6] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, et P. Abbeel, « Trust Region Policy Optimization », févr. 2015, Consulté le: 13 octobre 2025. [En ligne]. Disponible sur: <http://arxiv.org/abs/1502.05477v5>
- [7] J. Langford, « Approximately Optimal Approximate Reinforcement Learning », p. 267-274, 2002.
- [8] « Trust Region Policy Optimization — Spinning Up documentation ». Consulté le: 14 octobre 2025. [En ligne]. Disponible sur: <https://spinningup.openai.com/en/latest/algorithms/trpo.html#background>
- [9] David Pollard, *Asymptotia*, Ch. 3, "Distances and affinities between measures". 2000, p. 6-7. Consulté le: 13 octobre 2025. [En ligne]. Disponible sur: <https://web.archive.org/web/20150412031925/http://www.stat.yale.edu/~pollard/Books/Asymptopia/Metrics.pdf>
- [10] David J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003, p. 34. Consulté le: 13 octobre 2025. [En ligne]. Disponible sur: https://books.google.fr/books?id=AKuMj4PN_EMC&lpg=PA34&pg=PA34#v=onepage&q&f=false
- [11] Z. Xie, « Simple Policy Optimization », janv. 2024, Consulté le: 16 octobre 2025. [En ligne]. Disponible sur: <http://arxiv.org/abs/2401.16025v2>
- [12] « Proximal Policy Optimization — Spinning Up documentation ». Consulté le: 16 octobre 2025. [En ligne]. Disponible sur: <https://spinningup.openai.com/en/latest/algorithms/ppo.html>
- [13] NVIDIA Developer, « Isaac Sim - Robotics Simulation and Synthetic Data Generation ». Consulté le: 10 octobre 2025. [En ligne]. Disponible sur: <https://developer.nvidia.com/isaac/sim>
- [14] NVIDIA Developer, « PhysX SDK - Latest Features & Libraries ». Consulté le: 10 octobre 2025. [En ligne]. Disponible sur: <https://developer.nvidia.com/physx-sdk>
- [15] Consulté le: 16 juin 2025. [En ligne]. Disponible sur: <https://mujoco.org/>
- [16] « MuJoCo simulate tutorial ». Consulté le: 10 octobre 2025. [En ligne]. Disponible sur: <https://youtu.be/P83tKA1iz2Y>
- [17] Consulté le: 6 juin 2025. [En ligne]. Disponible sur: <https://gazebo-sim.org/>
- [18] « Gazebo Sim: Physics engines ». Consulté le: 10 octobre 2025. [En ligne]. Disponible sur: <https://gazebo-sim.org/api/sim/9/physics.html>

- [19] J. Lee *et al.*, « Dart: Dynamic animation and robotics toolkit », *The Journal of Open Source Software*, vol. 3, n° 22, p. 500, 2018.
- [20] Bullet Physics SDK, « bullet3 ». Consulté le: 10 octobre 2025. [En ligne]. Disponible sur: <https://github.com/bulletphysics/bullet3>
- [21] « Bullet Real-Time Physics Simulation | Home of Bullet and PyBullet: physics simulation for games, visual effects, robotics and reinforcement learning. ». Consulté le: 10 octobre 2025. [En ligne]. Disponible sur: <https://pybullet.org/wordpress/>
- [22] Roy Featherstone, « Robot Dynamics Algorithms », 1978, *Springer New York, NY*.
- [23] Erwin Coumans, « Bullet Physics Simulation Constraint Solving and Featherstone Articulated Body Algorithm ». International Conference and Exhibition on Computer Graphics and Interactive Technologies, 2015. Consulté le: 10 octobre 2025. [En ligne]. Disponible sur: https://docs.google.com/presentation/d/1wGUJ4neOhw5i4pQRfSGtZPE3CIIm7MfmqfTp5aJKuFYM/edit?slide=id.g644a5aa5f_0_16#slide=id.g644a5aa5f_0_16
- [24] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, et D. Meger, « Deep Reinforcement Learning that Matters », sept. 2017, Consulté le: 16 octobre 2025. [En ligne]. Disponible sur: <http://arxiv.org/abs/1709.06560v3>
- [25] Brian Lonsdorf, « Professor Frisby's Mostly Adequate Guide to Functional Programming », 2015, *Github*. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://github.com/MostlyAdequate/mostly-adequate-guide/blob/master/ch03.md>
- [26] « Reproducible Builds ». Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://reproducible-builds.org/>
- [27] Fortran 2015 Committee Draft (J3/17-007r2), *ISO/IEC JTC 1/SC 22/WG5/N2137*. International Organization for Standardisation, 2017, p. 336-338. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://wg5-fortran.org/N2101-N2150/N2137.pdf>
- [28] « Relationship Between Routines, Functions, and Procedures », 13 janvier 2025, *IBM*. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://www.ibm.com/docs/en/informix-servers/15.0.0?topic=statement-relationship-between-routines-functions-procedures>
- [29] « Different Types of Robot Programming Languages », 2015, *Plant Automation Technology*. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://www.plan-tautomation-technology.com/articles/different-types-of-robot-programming-languages>
- [30] « Imperative programming: Overview of the oldest programming paradigm », 21 mai 2021, *IONOS*. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://www.ionos.com/digitalguide/websites/web-development/imperative-programming/>
- [31] Bill Hoffman et Kenneth Martin, *The Architecture of Open Source Applications (Volume 1) CMake*. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://aosabook.org/en/v1/cmake.html>
- [32] Consulté le: 19 mai 2025. [En ligne]. Disponible sur: <https://nix.dev/manual/nix/2.17/language/>
- [33] Consulté le: 3 septembre 2025. [En ligne]. Disponible sur: <https://search.nixos.org/packages>

- [34] NixOS Wiki Authors, « Nixpkgs/Contributing ». Consulté le: 3 septembre 2025. [En ligne]. Disponible sur: <https://wiki.nixos.org/wiki/Nixpkgs/Contributing>
- [35] « Cachix — Nix binary cache hosting ». Consulté le: 3 septembre 2025. [En ligne]. Disponible sur: <https://www.cachix.org/>
- [36] « Nix (package manager) — Sandboxing ». Consulté le: 3 septembre 2025. [En ligne]. Disponible sur: [https://wiki.nixos.org/wiki/Nix_\(package_manager\)#Internals](https://wiki.nixos.org/wiki/Nix_(package_manager)#Internals)
- [37] « GitHub-hosted runners », *Github*. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://docs.github.com/en/actions/concepts/runners/github-hosted-runners>
- [38] Consulté le: 6 juin 2025. [En ligne]. Disponible sur: <https://www.raspberrypi.com/>
- [39] Fernando Borretti, « NixOS for the Impatient », 7 mai 2023. Consulté le: 4 septembre 2025. [En ligne]. Disponible sur: <https://borretti.me/article/nixos-for-the-impatient>

A Preuves

A.1 Cas dégénéré de $D_{\text{KL}}(Q, Q') = 0$ sans utilisation de \max

Soit S (resp. $A \subset \mathbb{N}$) l'espace des états (resp. actions) de l'environnement. Soit $Q : S \times A \rightarrow [0, 1]$ une distribution de probabilité du choix par l'agent d'une action dans un état tel que

$$\forall s \in S, Q(s, 1) = Q(s, 2) \quad (25)$$

Soit $Q' : S \times A \rightarrow [0, 1]$ définit ainsi:

$$\forall s \in S, Q'(s, 1) := 2Q(s, 1) \quad (26)$$

$$\forall s \in S, Q'(s, 2) := \frac{1}{2}Q(s, 2) \quad (27)$$

$$\forall s \in S, \forall a \in A - \{1, 2\}, Q'(s, a) := Q(s, a) \quad (28)$$

On a

$$D_{\text{KL}}(Q \parallel Q') = \sum_{(s,a) \in S \times A} Q(s, a) \log \frac{Q(s, a)}{Q'(s, a)}$$

On découpe la somme selon les valeurs de A :

$$\begin{aligned} &= \sum_{s \in S} \sum_{a \in A - \{1, 2\}} \left[Q(s, a) \log \frac{Q(s, a)}{Q'(s, a)} \right] + Q(s, 1) \log \frac{Q(s, 1)}{Q'(s, 1)} + Q(s, 2) \log \frac{Q(s, 2)}{Q'(s, 2)} \\ &= \sum_{s \in S} \underbrace{\sum_{a \in A - \{1, 2\}} Q(s, a) \log \frac{Q(s, a)}{Q(s, a)}}_{\text{d'après (28)}} + Q(s, 1) \log \underbrace{\frac{Q(s, 1)}{2Q(s, 1)}}_{\text{d'après (26)}} + Q(s, 2) \log \underbrace{\frac{Q(s, 2)}{\frac{1}{2}Q(s, 2)}}_{\text{d'après (27)}} \\ &= \sum_{s \in S} Q(s, 1) \left[\log Q(s, 1) - \log Q(s, 1) - \log 2 \right] + \quad (29) \\ &\quad Q(s, 2) \left[\log Q(s, 2) - \log Q(s, 2) - \log \frac{1}{2} \right] \\ &= \sum_{s \in S} -Q(s, 1) \log 2 + Q(s, 2) \log 2 \\ &= \sum_{s \in S} \log 2 \underbrace{(Q(s, 2) - Q(s, 1))}_{\text{d'après (25)}} \\ &= \sum_{s \in S} 0 = 0 \end{aligned}$$

A.2 $\eta(p, r)$ comme une espérance

Soit r une fonction récompense et p une politique. Soit C une variable aléatoire à valeurs dans \mathcal{S} , dont la loi de probabilité suit celle de p .

On a

$$\begin{aligned} \exp\left(\sum_{t=0}^{\infty} \gamma^t r(C_t)\right) &= \sum_{(c_t)_{t \in \mathbb{N}} \in \mathcal{S}} \left(\sum_{t=0}^{\infty} \gamma^t r(c_t)\right) \mathbb{P}\left(\sum_{t=0}^{\infty} \gamma^t r(C_t) = \sum_{t=0}^{\infty} \gamma^t r(c_t)\right) \\ &= \sum_{(c_t)_{t \in \mathbb{N}} \in \mathcal{S}} \left(\sum_{t=0}^{\infty} \gamma^t r(c_t)\right) \mathbb{P}(C = (c_t)_{t \in \mathbb{N}}) \end{aligned} \quad (30)$$

Soit S (resp. A) la suite des premiers (resp. deuxièmes) éléments de C , c'est-à-dire $\forall t \in \mathbb{N}, (S_t, A_t) := C_t$.

Étant donné la définition de \mathcal{S} :

- S_t dépend de A_{t-1} et S_{t-1}
- A_t dépend de S_t

On a alors, pour toute suite $(c_t)_{t \in \mathbb{N}} \in \mathcal{S}$:

$$\begin{aligned} P(C = (c_t)_{t \in \mathbb{N}}) &= \mathbb{P}(S_0 = s_0) \mathbb{P}(A_0 = a_0 \mid S_0 = s_0) \cdot \\ &\quad \prod_{t=1}^{\infty} \mathbb{P}(S_t = s_t \mid C_{t-1} = c_{t-1}) \mathbb{P}(A_t = a_t \mid S_t = s_t) \end{aligned} \quad (31)$$

On a

$$\begin{aligned} \mathbb{P}(S_0 = s_0) &= \rho_0(s_0) \\ \forall t \in \mathbb{N}, \quad \mathbb{P}(A_t = a_t \mid S_t = s_t) &= Q_p(s_t, a_t) \\ \forall t \in \mathbb{N}^*, \quad \mathbb{P}(S_t = s_t \mid C_{t-1} = c_{t-1}) &= \mathbb{P}(M(C_{t-1}) = M(c_{t-1}) \mid C_{t-1} = c_{t-1}) \\ &= \mathbb{P}(C_{t-1} = c_{t-1} \mid C_{t-1} = c_{t-1}) = 1 \end{aligned} \quad (32)$$

Donc on a

$$\begin{aligned} P(C = (c_t)_{t \in \mathbb{N}}) &= \rho_0(s_0) Q_p(s_0, a_0) \prod_{t=1}^{\infty} Q_p(s_t, a_t) \\ &= \rho_0(s_0) \prod_{t=0}^{\infty} Q_p(s_t, a_t) \end{aligned} \quad (33)$$

Et ainsi

$$\begin{aligned} \exp\left(\sum_{t=0}^{\infty} \gamma^t r(C_t)\right) &= \sum_{(c_t)_{t \in \mathbb{N}} \in \mathcal{S}} \left(\sum_{t=0}^{\infty} \gamma^t r(c_t)\right) \mathbb{P}(C = (c_t)_{t \in \mathbb{N}}) \\ &= \sum_{(c_t)_{t \in \mathbb{N}} \in \mathcal{S}} \left(\sum_{t=0}^{\infty} \gamma^t r(c_t)\right) \rho_0(s_0) \prod_{t=0}^{\infty} Q_p(s_t, a_t) \\ &= \eta(p, r) \quad \blacksquare \end{aligned} \quad (34)$$

A.3 Simplification de l'expression de $L(s, a, \mathcal{P}, \mathcal{P}', R)$ dans PPO-Clip

Soit $(s, a) \in S \times A$, et \mathcal{P}' une politique. Posons $\alpha := A_{\mathcal{P}', R}(s, a)$, $q/q' := Q_{\mathcal{P}}(s, a)/Q_{\mathcal{P}'}(s, a)$.

Cas $\alpha > 0$	Cas $\alpha < 0$
$L(s, a, \mathcal{P}, \mathcal{P}', R)$ $= \min\left(\frac{q}{q'}\alpha, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\alpha\right)$ $= \min\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ <p style="text-align: center; margin-top: 0;">car $\alpha > 0$</p>	$L(s, a, \mathcal{P}, \mathcal{P}', R)$ $= \min\left(\frac{q}{q'}\alpha, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\alpha\right)$ $= \max\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ <p style="text-align: center; margin-top: 0;">car $\alpha < 0$</p>
...et $q/q' \in [1 - \varepsilon, 1 + \varepsilon]$	
$= \min\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ $= \min\left(\frac{q}{q'}, \quad \frac{q}{q'}\right)\alpha$ $= \min\left(\frac{q}{q'}, 1 + \varepsilon\right)\alpha$	$= \max\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ $= \max\left(\frac{q}{q'}, \quad \frac{q}{q'}\right)\alpha$ $= \max\left(\frac{q}{q'}, 1 - \varepsilon\right)\alpha$
...et $q/q' > 1 + \varepsilon$	
$= \min\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ $= \min\left(\frac{q}{q'}, \quad 1 + \varepsilon\right)\alpha$	$= \max\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ $= \max\left(\frac{q}{q'}, \quad 1 + \varepsilon\right)\alpha$ $= \max\left(\frac{q}{q'}, \quad 1 - \varepsilon\right)\alpha$ <p style="text-align: center; margin-top: 0;">car $1 - \varepsilon < 1 + \varepsilon < \frac{q}{q'}$</p>
...et $q/q' < 1 - \varepsilon$	
$= \min\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ $= \min\left(\frac{q}{q'}, \quad 1 - \varepsilon\right)\alpha$ $= \min\left(\frac{q}{q'}, \quad 1 + \varepsilon\right)\alpha$ <p style="text-align: center; margin-top: 0;">car $1 + \varepsilon > 1 - \varepsilon > \frac{q}{q'}$</p>	$= \max\left(\frac{q}{q'}, \quad \text{clip}\left(\frac{q}{q'}, 1 - \varepsilon, 1 + \varepsilon\right)\right)\alpha$ $= \max\left(\frac{q}{q'}, \quad 1 - \varepsilon\right)\alpha$