Apprentissage par renforcement causal

Othmane Baddou
Etudiant à CentraleSupélec
othmane.baddou@student-cs.fr

Ayoub Bakkoury
Etudiant à CentraleSupélec
ayoub.bakkoury@student-cs.fr

Ismail Benaija
Etudiant à CentraleSupélec
ismail.benaija@student-cs.fr

Victor Gauthier
Etudiant à CentraleSupélec
victor.gauthier@student-cs.fr

Benoît Giniès
Etudiant à CentraleSupélec
benoit.ginies@student-cs.fr

Abstract

Dans ce papier il est question d'apprentissage causal et d'apprentissage par renforcement, et du lien que l'on peut essayer d'établir entre les deux. Nous avons cherché à réaliser un état de l'art des méthodes de causal reinforcement learning (CRL) qui sont développées à ce jour. Il y est aussi question des possibles implémentations qu'on peut en faire et des applications qu'on pourrait lui trouver.

1 Introduction

1.1 Contexte

L'apprentissage par renforcement (RL) est une des branches de l'apprentissage machine qui est le plus en vogue de nos jours. C'est son développement qui a permis le développement des intelligences artificielles capables de battre les humains aux échecs. Sa grande force réside dans sa capacité à explorer des environnements méconnus ou trop vastes pour être décrits précisément, et à exploiter les propriétés découvertes au fil de l'exploration pour construire des algorithmes d'intelligence artificielle robustes et efficaces.

Cette capacité exploratoire du RL fait fortement écho à une autre branche de l'apprentissage machine : l'apprentissage causal. Sous cette dénomination se cachent l'ensemble des algorithmes d'apprentissage dont le but est de repérer, au sein des données, des structures de causalité qui permettent in fine de mieux connaître les mécanismes en jeu dans l'environnement qu'on étudie, donc de mieux le simuler et d'affiner les décisions prises par l'intelligence artificielle. Ce qui la distingue des algorithmes classiques d'apprentissage machine, c'est que, quand les algorithmes ne repèrent que des liens de corrélation, l'apprentissage causal tente de donner du sens aux liens rencontrés, et repère ainsi des causalités.

Il n'y a alors qu'un pas entre la formulation de ce parallèle et l'idée de faire se rencontrer les deux branches. C'est là l'idée qu'ont eue quelques chercheurs, et que nous allons étudier maintenant : qu'est ce que l'apprentissage par renforcement causal ?

1.2 Principe général

Comme nous l'avons déjà souligné, les deux approches d'apprentissage par renforcement et d'apprentissage causal se distinguent par leur forte dimension exploratoire. Il convient dès lors de

réaliser un mélange des deux approches exploratoires et d'en déduire un processus d'apprentissage qui présenterait les avantages des deux branches. C'est la tâche à laquelle se sont pliés les membres du "Causal Reinforcement Learning Laboratory" au sein du département de Computer Science de Columbia university, menés par Elias Bareinboim (voir [1] pour plus d'informations).

L'idée derrière le travail présenté dans [1], a été de partir du RL, en s'inspirant des méthodes existantes telles que BANDIT, en y ajoutant de nouvelles procédures tirées de l'apprentissage causal. Le problème duquel nous partons est modélisé d'une manière légèrement différente dans le cas causal. On garde la représentation selon laquelle un agent interagit avec un environnement pour en apprendre les caractéristiques à travers des récompenses qu'il collecte au cours de ses expériences. Cependant, on considère désormais que l'environnement qu'on étudie est muni d'une matrice de causalité qui représente les liens de causalité entre les différentes variables qui décrivent cet environnement. Le but de l'agent est donc désormais de découvrir le comportement de l'environnement et ses réactions aux différentes actions que l'agent peut appliquer, tout en déterminant une estimation de cette matrice de causalité.

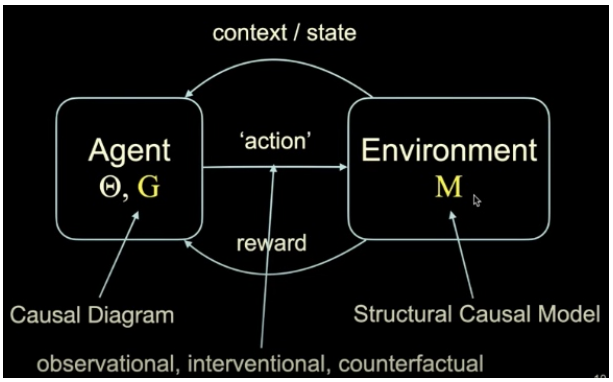


Figure 1: Représentation schématique du problème d'apprentissage causal.

Cette modélisation du problème est commune à l'intégralité des méthodes de RLC que nous avons rencontrées au cours de ce projet.

2 Méthodes d'apprentissage causal

L'approche qui est présentée par Elias Bareinboim ([1]) est à la fois très générale, mais présente quelques spécificités que nous soulignerons plus tard en la comparant à d'autres méthodes.

Les différentes étapes de l'exploration d'un algorithme de RL sont d'abord disséquées et théorisées, avant d'être remaniées. Plusieurs types d'apprentissage par exploration sont présentés. Un premier type dans lequel l'agent apprenant, effectue lui même des expériences et ajuste son modèle de prédiction en conséquence. Un autre type dans lequel l'agent observe les expériences d'un autre agent (qui lui n'apprend pas nécessairement) et base son ajustement dessus. Enfin, un dernier type dans lequel l'agent apprenant base son ajustement sur les observations d'autres agents, peu importe leur provenance.

Pour l'apprentissage en lui même, les chercheurs distinguent trois types de raisonnements possibles:

- Le raisonnement par association, qui se base sur l'observation, c'est le raisonnement propre à l'apprentissage profond par exemple.
- Le raisonnement par intervention, qui s'appuie sur l'expérience, c'est celui qu'on utilise pour le RL.
- Le raisonnement contre-factuel, qui consiste à modifier par la pensée les conditions d'une expérience, et en déduire le résultat conséquent. C'est ce dernier type de raisonnement que les chercheurs ont cherché à mêler au RL.

A partir de cette théorisation des grands aspects de l'apprentissage par renforcement, les chercheurs ont cherché à modifier les procédures appliquées au cours d'un processus d'apprentissage (comme pendant un BANDIT par exemple) développant ainsi les méthodes alternatives suivantes.

2.1 Generalized policy learning

L'Online learning n'est généralement pas une méthode acceptable en raison de contraintes financières, techniques ou éthiques. Prenons l'exemple du développement d'un nouveau traitement pour le cancer. Certaines actions à répétition peuvent avoir des conséquences catastrophiques et irréversibles sur les patients.

En général, on souhaite exploiter les données recueillies dans différentes conditions pour accélérer l'apprentissage, sans avoir à repartir de zéro. C'est ce que l'on appelle *l'offline learning*.

Cependant, ce type d'apprentissage demande que certaines conditions strictes soient vérifiées, ce qui est rarement le cas. C'est pourquoi ce cas d'usage est très intéressant.

En effet, l'objectif est de combiner ces deux approches pour des méthodes plus réalistes et appropriées aux cas réels (même lorsque les conditions prouvées nécessaires ne sont pas respectées).

Prenons l'exemple du médical encore une fois. Nous cherchons à concevoir une expérimentation optimale à partir de données d'observation.

Nous avons à notre disposition un jeu de données observées. À partir de la distribution

$$P(x, y) \tag{1}$$

l'objectif est d'apprendre une politique pour

$$P(y|do(x)). \tag{2}$$

Cependant, l'hypothèse que le contexte selon lequel ces données ont été collectées est le même que celui de notre expérimentation n'est pas valide. Théoriquement, on ne peut donc pas utiliser les données déjà relevées pour notre problème et la seule solution serait d'utiliser l'*online learning*. Évidemment, ceci n'est pas un résultat acceptable et optimal.

Pour illustrer comment la prise en compte des effets de causalité permet de répondre à ce type de problématiques, considérons tout d'abord que les contextes des deux expérimentations sont les mêmes, et appelons ce cas le "*Naive TS*".

Ainsi, le *naive TS* tentera d'utiliser les données d'observations comme *prior*. La figure 2 ci-dessous présente les résultats de cette expérimentation (le graphique est tiré du travail d'Elias Bareinboim du laboratoire de Columbia University).

On observe aussi dans ce graphique le résultat d'un modèle qui ne prendrait aucunement en compte les données à disposition : appelons le "traditional TS" (figure 2).

Les résultats présentés sont contradictoires à toute intuition statistique. En effet on pourrait se demander comment l'ajout de données d'observation peut affecter négativement les performances du modèle ?

L'explication réside dans la nature de la distribution des données. Prenons le cas où :

$$E(Y|X = 1) > E(Y|X = 0) \tag{3}$$

Tandis qu'en réalité, corrélation n'implique pas causalité :

$$E(Y|do(X = 1)) < E(Y|do(X = 0)) \tag{4}$$

Ainsi, dans ce type de situation, le *naive TS* aura de très mauvaises performances. Malheureusement en pratique, nous n'avons jamais accès à cette information, c'est d'ailleurs là où réside toute la complexité et l'objectif de ces approches.

De plus, on ne peut pas savoir à l'avance si notre donnée comporte ce type de pattern. Cependant, ne pas utiliser toute la donnée disponible et considérer qu'il n'y a aucune information intéressante dans celle ci ne serait pas raisonnable.

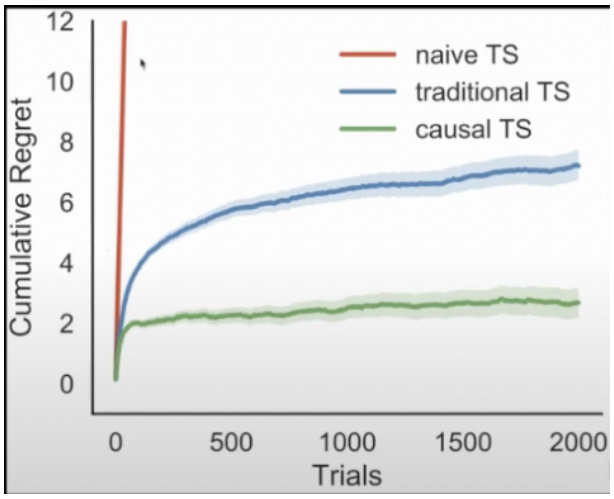


Figure 2: Comparaison du naive, traditional et causal TS

Pour résoudre ce problème, il est possible de borner l'espérance de l'effet de causalité. Le théorème discuté dans les papiers en référence [2], [3] et [4] stipule que pour toute observation provenant d'une distribution P(x,y), l'espérance de l'effet de causalité est bornée par :

$$E(Y|x)P(x) < E(Y|do(X) < E(Y|x)P(x) + 1 - P(x) \tag{5}$$

Ainsi, il est possible d'intégrer ces frontières au processus d'apprentissage. L'algorithme ci-dessous, tiré du travail d'Elias Bareinboim [1] explique comment intégrer au modèle les frontières

$$Ix = E(Y|x)P(x) \tag{6}$$

et

$$hx = Ix + 1 - P(x). \tag{7}$$

Algorithm 1: Introduction des bornes causales dans le processus d'apprentissage online x offline

Input: α, β : les paramètres du prior / l_x, h_x : les bornes causales calculées pour chaque bras x
$\forall x, S_x \leftarrow 0$;
$\forall x, F_x \leftarrow 0$;
for $t = 1, ..., T$ **do**
  foreach x **do**
    repeat
      Draw $\theta_x \sim Beta(S_x + \alpha, F_x + \beta)$
    until $\theta_x \in [l_x, h_x]$;
  end
  Play $do(x_t)$ where $X_t = argmax_x \theta_x$;
  Observe Y_t and update F_{xt} and S_{xt};
end

La figure 2 montre les résultats de l'ajout de ces frontières au modèle. Comme on peut le voir, cette méthode améliore considérablement les performances.

La prise en compte de ces bornes dans l'apprentissage s'appelle l'échantillonnage causal de Thompson, nous en avons rencontré des implémentations au cours de nos recherches.

2.2 When and where to intervene ?

L'idée ici est de redéfinir l'espace des politiques dans lequel on cherche notre politique optimale (ou du moins la politique résultant de notre processus d'apprentissage). On fait ici l'hypothèse que l'agent apprenant n'est pas conscient des liens de causalité qui relient les différentes variables de l'environnement. Ainsi, le graphe de causalité n'étant pas connu, le choix des actions à appliquer à l'environnement au cours de l'apprentissage sont faits de manière arbitraire, et peuvent donc handicaper la convergence. Il est donc ici question de définir une nouvelle méthode d'expérimentation, qui permet de respecter les contraintes de causalité imposées par le graphe de causalité de l'environnement.

Ainsi, à partir de l'ensemble de toutes les interventions, ou actions, que l'on peut appliquer à l'environnement, on commence par repérer ce qu'on appelle les *Minimal Intervention Set*. L'idée ici est que, l'intégralité des interventions qu'on pourrait réaliser sont une combinaison des MIS. Un effet intéressant de cette étape, est qu'elle pousse à préférer des interventions nucléaires, plutôt que des interventions qui agissent sur plusieurs variables du système en même temps. Effectivement, pour en apprendre davantage sur la réaction des sols à différents facteurs agricoles, il vaudrait mieux faire varier chaque paramètre un à un et observer ses conséquences plutôt que de tout faire varier en même temps.

Au sein de ces MIS, on continue en sélectionnant les *Possibly Optimal MIS*. Il s'agit des MIS dont la récompense est supérieure à tous les autres MIS.

Enfin, on finit en inférant la récompense moyenne associée à un bras, en utilisant les données recueillies sur les autres bras. La conjonction de ces trois étapes, permet de réduire le nombre d'interventions que l'on peut appliquer, et aussi d'améliorer les résultats en termes de regret, par rapport à une approche où l'on teste toutes les combinaisons possibles.

2.3 Conterfactual decision-making

Ici, les chercheurs ont voulu introduire le raisonnement contrefactuel, dont nous avons parlé plus haut, dans le processus d'apprentissage par renforcement. L'idée qui se cache derrière est la suivante : un agent qui n'a pas une démarche contrefactuel ne réfléchit pas aux raisons des choix d'interventions qu'il fait, et peut ainsi facilement être utilisé ou trompé par l'environnement qu'il étudie, passant à côté de causalités fondamentales. Il s'agit dès lors de doter l'agent apprenant d'une faculté de remise en question de ses choix de manière à "prendre l'ascendant" sur l'environnement et explorer plus en profondeur les liens de causalité qui le caractérisent.

Ainsi, supposons qu'un agent est disposé à jouer x_0 sur un bras donné. S'apprêtant à le jouer, on le stoppe. Alors ou bien :

- On le laisse reprendre là où il en est, et il joue x_0.
- On change complètement son état et on le dispose à jouer x_1.
- On ne change pas son état, mais on le force à jouer x_1, qu'il n'aurait ainsi pas joué s'il avait poursuivi son raisonnement.

C'est cette dernière démarche qui introduit vraiment le raisonnement contrefactuel. On est alors en mesure de berner l'environnement, et de déceler des causalités qui exploitent l'état de l'agent apprenant (du type : si l'agent est dans cet état là, alors les récompenses seront moindres).

Il est évident qu'implémenter une telle méthode n'est pas simple. Cela nécessite d'avoir une modélisation de l'environnement et de l'agent adéquate. Il existe d'ailleurs des cas dans lesquels il est plus facile de l'appliquer que d'autres (binary treatment, backdoor admissibility ...).

2.4 Transportabilité des modèles causaux

La question de la généralisation des connaissances causales est centrale dans les inférences scientifiques puisque des expériences sont menées et que les conclusions obtenues en laboratoire sont transportées et appliquées ailleurs. Si l'environnement cible est radicalement différent de l'environnement d'étude, aucune relation causale ne peut être apprise. Cependant, le fait que l'expérimentation scientifique continue de fournir des informations utiles sur notre monde suggère

que certains environnements partagent des caractéristiques communes et que, du fait de ces points communs, les affirmations causales seraient valables même là où des expériences n'ont jamais été réalisées.

La classe de problèmes de généralisabilité causale est appelée transportabilité. Nous considérons l'exemple le plus général de transportabilité connu à ce jour qui est le problème du transport de connaissances expérimentales depuis des contextes hétérogènes vers une certaine cible spécifique.

Plus précisément, le problème de *mz-transportabilité* concerne le transfert de connaissances causales d'une collection hétérogène de domaines source $\Pi = \pi_1, ..., \pi_n$ vers un domaine cible π^*. Dans chaque domaine $\pi_i \in \Pi$, des expériences sur un ensemble de variables Z_i peuvent être réalisées et des connaissances causales rassemblées. En π^*, potentiellement différent de π_i, seules des observations passives peuvent être collectées (cette contrainte sera affaiblie). Le problème est d'inférer une relation causale R dans π^* à partir des connaissances obtenues dans Π.

Des conditions suffisantes pour la *mz-transportabilité* ont été données dans [5], mais ce traitement ne permet pas de garantir si ces conditions sont également nécessaires et si ils doivent être augmentées, voire remplacées par des conditions plus générales.

De récentes recherches [6] ont réussi à montrer ces résultats très importants:

- Une condition nécessaire et suffisante pour décider quand les effets causals dans le domaine cible peuvent être estimés à partir à la fois des informations statistiques disponibles et des informations causales transférées des expériences dans les domaines.
- Une preuve que l'algorithme proposé dans [6] est en fait complet pour calculer la formule de transport, c'est-à-dire que la stratégie conçue pour combiner les preuves empiriques pour synthétiser la relation cible ne peut pas être améliorée.
- Une preuve que le calcul est complet pour la classe de *mz-transportabilité*, ce qui signifie que trouver une preuve dans ce langage suffit pour résoudre le problème.

2.5 Apprentissage des modèles causaux

Le problème de l'apprentissage des relations causales sous-jacentes à un système complexe est d'un grand intérêt en IA et dans toutes les sciences empiriques. Les systèmes causaux sont généralement représentés par des graphes acycliques dirigés, où les sommets sont des variables aléatoires et une arête de la variable X à Y indique que la variable X est une cause directe de Y.

Pour découvrir les relations causales entre un ensemble de variables, s'il est limité de travailler uniquement avec des données d'observation des variables, on peut utiliser un algorithme basé sur des contraintes. De telles approches purement observationnelles reconstruisent le graphe causal jusqu'à la classe d'équivalence de Markov, et par conséquent, le chercheur se retrouve généralement avec quelques (ou dans certains cas plusieurs) relations causales non résolues. Bien que, sous certaines hypothèses supplémentaires, dans certains contextes convaincants, l'apprentissage de la structure complète en utilisant simplement des données d'observation est réalisable.

D'autre part, il est bien entendu que chaque fois que l'investigateur peut effectuer un nombre suffisant d'interventions, le graphe causal représentant le système sous-jacent peut être entièrement récupéré. Il existe un nombre croissant de recherches sur l'apprentissage des structures causales à l'aide de données interventionnelles dans des systèmes causalement suffisants et causalement insuffisants (avec des variables latentes) . Une approche d'apprentissage structurel interventionnel nécessite la réalisation d'un ensemble d'expériences, chacune intervenant sur un sous-ensemble de variables, puis la collecte de données à partir du système intervenu.

Dans ce cadre, deux questions se posent :

- Quel est le plus petit nombre d'expériences requis pour apprendre pleinement le graphe causal sous-jacent ?
- Pour un nombre fixe d'expériences (budget), quelle proportion du graphe causal est apprenable ?

Le premier problème a été abordé dans la littérature sous différentes hypothèses. Dans [7], les auteurs ont pris en compte les coûts d'intervention sur chaque variable et ont dérivé un algorithme de conception d'expérience avec un coût total minimum qui reconstruit l'ensemble de la structure.

Le problème de trouver la meilleure cible d'intervention peut être formulé comme un problème d'optimisation qui vise à maximiser le nombre moyen d'arêtes dont les directions sont découvertes. Les résultats dans [8] montrent qu'une partie importante des systèmes causaux peut être apprise par seulement un petit nombre d'interventions.

2.6 Imitation causale

L'apprentissage par imitation se concentre sur l'apprentissage de politiques avec des performances appropriées à partir de démonstrations générées par un expert, avec une mesure de performance non spécifiée et un signal de récompense non observé.

Les méthodes populaires d'apprentissage par imitation commencent soit par imiter directement la politique de comportement d'un expert (clonage de comportement), soit par l'apprentissage d'une fonction de récompense qui donne la priorité aux trajectoires d'experts observées (apprentissage par renforcement inverse).

Cependant, ces méthodes reposent sur l'hypothèse que les covariables utilisées par l'expert pour déterminer ses actions sont pleinement observées.

Lorsqu'il existe des covariables non observées, imiter naïvement la politique de l'expert nominal ne conduit pas nécessairement à une performance satisfaisante, même lorsque l'expert lui-même se comporte de manière optimale.

Dans [9], ils introduisent un critère graphique complet pour déterminer la faisabilité de l'imitation à partir de données de démonstration et de connaissances qualitatives sur le processus de génération de données représenté sous forme de graphe causal. Ensuite, ils développent un algorithme suffisant pour identifier une politique d'imitation lorsque le critère donné ne tient pas, en exploitant les connaissances quantitatives dans la distribution observationnelle et enfin fournissent une procédure efficace et pratique pour trouver une politique d'imitation grâce à une paramétrisation explicite du modèle causal.

3 Les applications du CRL

La première section de ce document établit les bases théoriques du CRL. Nous allons maintenant présenter certains cas d'applications intéressantes de ces méthodes.

Il reste important de préciser que ce domaine de recherche est encore jeune et très vaste, les domaines d'application sont donc nombreux. La suite de ce document ne fait pas figure de liste exhaustive des cas d'applications mais plutôt d'une sélection que les auteurs ont jugé comme étant pertinente. Une étude des articles Finn Lattimore et al. [11] et Sachidanda et al. [10] nous permet de réaliser une étude comparative de bandits classiques avec des modèles prenant en compte d'éventuels graphes de causalité. Nous nous sommes également appuyés sur un template de comparaison de certains modèles par Shah et al. [12] en modifiant certaines de leurs configurations.

3.1 Cas d'usage

Le modèle de "Causal Bandit" a en effet été formulé la première fois en 2016 dans Lattimore et al. [11] dans lequel on suppose que le graphe causal est connu en amont de l'apprentissage, ce qui n'est pas toujours le cas dans des problèmes de la vie courante. Nous conservons cette hypothèse dans notre étude comparative, et analyserons des résultats pour différentes structures de graphes possibles. Une instance de bandit causal peut être décrite comme suit. On nous donne un ensemble de variables aléatoires $\mathcal{X} = \{X_1, X_2...X_N\}$, une variable de récompense Y et un ensemble d'actions autorisées A. Les dépendances entre $\mathcal{X}$ et Y sont représentées à l'aide d'un graphe causal G, et on peut faire varier la quantité d'informations dont on dispose au préalable sur G. Chaque action a de A est une intervention de la forme

$$do(X = x_t) \tag{8}$$

. L'objectif des algorithmes est donc de minimiser le regret cumulatif R défini de la sorte :

$$R = E[Y = 1|a] * T - E[\sum_{t=1}^{T} y_t]. \tag{9}$$

.

Sachidananda et Brunskill [10] étendent le Thompson Sampling au cadre du bandit causal de Lattimore et al. en divisant la procédure en deux parties. Étant donnée une intervention $a = do(X = x)$, ils essaient d'abord d'estimer une distribution sur les affectations possibles des variables parentales de Y étant donné a, puis ils estiment la distribution du reward étant donné une configuration parentale particulière; les deux distributions sont mises à jour après chaque essai, à mesure que de nouveaux échantillons sont observés. Leur algorithme, que nous avons repris pour notre comparaison s'appelle *OC-TS*.

Nous combinons ainsi les études de ces deux papiers en les confrontant, et en les comparant à deux multi-armed bandits classiques, et en ajoutant un modèle d'agent explorant le graphe d'inférence de façon epsilon-greedy.

3.2 Agents implémentés

- Comme base de référence à laquelle les performances de tous les agents suivants seront comparées, nous implémentons ces algorithmes conventionnels de multi armed bandits pour notre cadre de bandit causal. Ces agents ignorent simplement les dépendances du graphe causal entre les variables, et les traitent comme étant indépendantes (ceci est réalisé en ignorant les valeurs échantillonnées de tous les X et en utilisant celles de tous les X). Nous avons implémenté un UCB et un Thomson Sampling.

- Nous reprenons une implémentation du bandit OC-TS de [10]

- Nous implémentons également un agent se basant sur un multi armed bandit de type epsilon-greedy dont nous avons tiré inspiration des recherches de Yash Shah et. al [12]. Il s'agit d'une stratégie basée sur un modèle pour trouver l'intervention optimale, qui estime la distribution du graphe en utilisant l'historique des valeurs échantillonnées de l'ensemble des variables pendant l'exploration, puis utilise cette estimation pour évaluer la récompense attendue pour chaque action pendant l'exploitation. L'algorithme epsilon-greedy est alors combiné à un graphe d'inférence de type Bernoulli. À chaque itération, l'objectif est de choisir l'action a^* telle qu'elle maximise sur l'ensemble des actions A la formule $E[Y = 1|a]$, qui permet ensuite de calculer des probabilités conditionnelles $P(Y = 1|a)$ par parcours du graphe (haute complexité), ou par exacte inférence comme dans Shah et al. [12]. Bien que cet algorithme soit sous-optimal en terme de complexité temporelle et de mémoire (complexité $2^{|G|}$, la confrontation de [12] avec les algorithmes de base offraient des résultats intéressants.

3.3 Graphes Bayésiens

Les graphes d'inférence peuvent être implémentés selon divers types de topologies. Ces structures régissent ainsi les différents liens possibles entre les variables du graphe, à considérer en fonction du cas de la vie courante étudié. Les papiers [11] et [12] proposent des structures de graphes diverses et variées (linéaires, parallèles, aléatoires, indépendantes...). Nous allons focaliser la comparaison sur 3 structures de graphes différentes (figure 3).

Les structures sont :

-Linéaire. Chaque variable, à l'exception d'une variable racine, a exactement un parent. L'idée ici est que l'agent doit apprendre que la meilleure action est celle dans laquelle l'ancêtre le plus proche possible intervient (à une valeur de 0 ou 1 selon la distribution).

- Aléatoire. Il s'agit d'un graphe bayésien complètement aléatoire obtenu en ajoutant itérativement des variables et en choisissant au hasard si chaque variable précédemment ajoutée est parente de la précédente.

- Disjointe ou indépendante. Ici, toutes les variables (à l'exception de la variable de récompense) sont indépendantes par paire et sont les parents de la variable de récompense. Dans ce cas, la

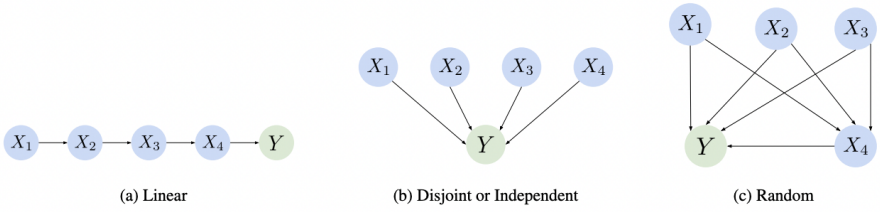


Figure 3: Structures de graphes retenues [12]

meilleure action est celle pour laquelle la probabilité marginalisée sur les variables non-intervenues est maximale.

Nous allons ainsi comparer les résultats des modèles précédemment énoncés pour ces trois structures de graphes.

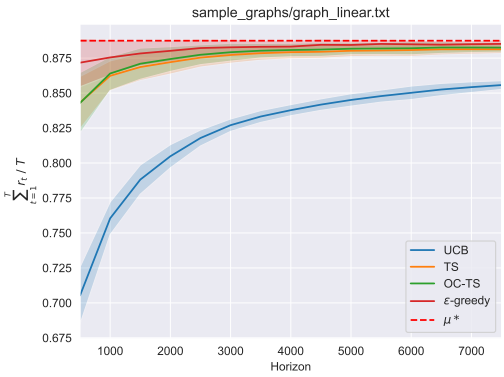
3.4 Applications possibles

Nous générons dans ce problème un graphe aléatoire à données flottantes, n'ayant pas de signification intrinsèque. Cependant, il est tout à fait possible de transposer cette structure à un problème courant sous réserve que celui-ci se modélise par l'une des structures de graphes que nous considérons. La bibliothèque Python 'Why Not' [13] permet en effet de simuler des structures causales via des graphes, mais dont la nature des graphes demeure tout de même à convertir avant implémentation. La bibliothèque propose des simulateurs de base, comme des traitements potentiels du virus Zika ou la propagation du VIH. Il est possible en dépassement de ce projet de générer un nouvel environnement, dont les états et les paramètres sont à définir au préalable, à l'instar d'une étude de causalité simplifiée de l'effet de la distanciation sociale sur la propagation du COVID-19 [14].

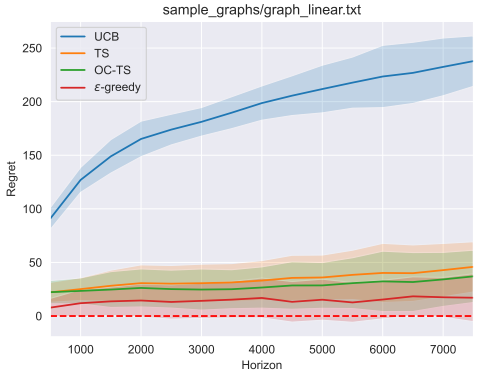
3.5 Résultats des expériences

Pour chacune des topologies, nous comparons les évolutions du regret des différents agents, ainsi que la convergence vers la valeur optimale $P(Y = 1|a^*)$. Les expériences sont calculées à horizon fixe de 7500 et sur 15 expériences (random seed). Ces paramètres sont modifiables dans les premières cellules du notebook. Les graphes sont générés notamment grâce aux fonctions inspirées de [11] et [12] dans les fichiers python "Graph Generator". Ils sont générés aléatoirement, selon une classe Graph, fournie par Shah et al. [12]. Pour plus de significativité, les écarts types des calculs à chaque étape sont également implémentés. Nous avons pris la même valeur d'epsilon que dans [12], qui a été estimée de façon empirique.

Le classement des algorithmes, en terme d'allure du regret ou de la convergence vers la valeur optimale de probabilité est toujours le même quelle que soit la topologie choisie. En revanche, les écarts peuvent être plus significatifs pour une combinaison linéaire plutôt qu'aléatoire ou disjointe. Conformément à nos attentes, ce sont les algorithmes pour lesquels nous prenons en compte l'inférence causale qui performent le mieux dans ce cas d'application. La méthode epsilon greedy est celle qui permet la convergence la plus rapide en terme d'horizon et d'itérations vers la valeur optimale de $E[Y = 1|a^*]$, bien que toujours très proche de ce qui est proposé dans Sachidananda et al. [10].

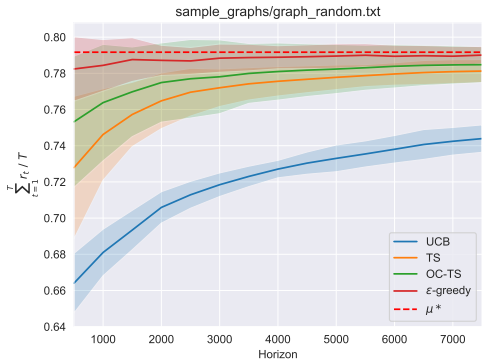


(a) Reward

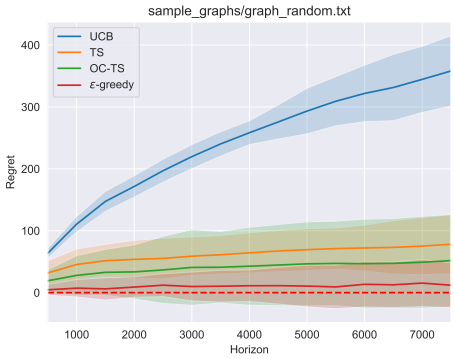


(b) Regret

Figure 4: Résultats pour graphe linéaire

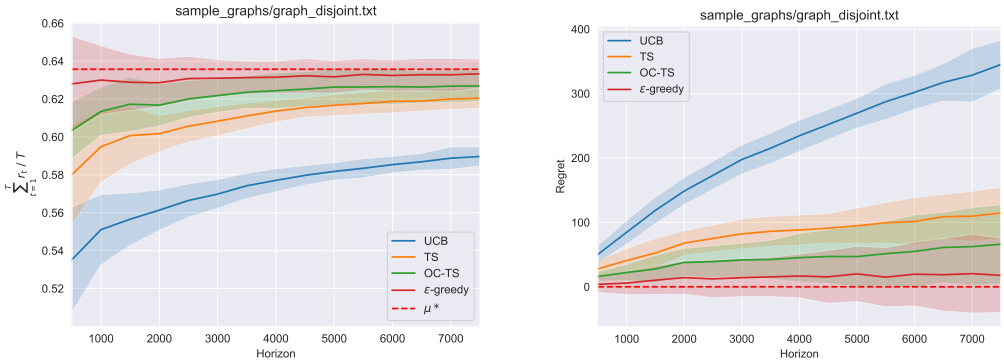


(a) Reward



(b) Regret

Figure 5: Résultats pour graphe aléatoire



(a) Reward

(b) Regret

Figure 6: Résultats pour graphe indépendant

4 Conclusion

Bien qu'il s'agisse encore d'un sujet de recherche actif, qui n'en est qu'à sa phase exploratoire, l'apprentissage par renforcement causal est déjà un domaine qui présente des méthodes implémentables, qui aboutissent à des résultats à même de concurrencer le RL traditionnel, et dont les possibilités d'application sont multiples et faciles à trouver. Cette approche semble donc très prometteuse, d'autant qu'elle contribue aussi à donner du sens à l'apprentissage machine au moment même où les travaux sur l'explicabilité s'intensifient. C'est toutefois une branche de l'apprentissage à la théorie robuste, très abstraite et peu accessible. Certaines des méthodes développées ne sont pas toujours implémentables (conditions d'application précises). Il reste donc encore de la place à la recherche pour démocratiser cette approche qui devrait apporter beaucoup à l'intelligence artificielle.

5 Réferences

[1] CausalAI Laboratory at Columbia University. (2020). Causal Reinforcement Learning (CRL). (https://crl.causalai.net/#)

[2] Linear Program formulation in other causal graphs (nonparametric SCMs): [Balke Pearl, 1996; Zhang and Bareinboim, IJCAI'17]

[3]Incorporating parametric knowledge: [Kallus Zhou, 2018; Namkoong et al., 2020]

[4]Sequential treatments in longitudinal settings: [Zhang Bareinboim, NeurIPS'19; ICML'20]

[5] E.Bareinboim, S.Lee, V.Honavar, and J.Pearl (2013) Transportability from multiple environments with limited experiments. (https://proceedings.neurips.cc/paper/2013/file/02522a2b2726fb0a03bb19f2d8d9524d-Paper.pdf)

[6] E.Bareinboim, and J.Pearl (2014) Transportability from Multiple Environments with Limited Experiments: Completeness Results. (http://ftp.cs.ucla.edu/pub/stat_ser/r443.pdf)

[7] Kocaoglu, Murat, Dimakis, Alexandros G, and Vishwanath, Sriram. Cost-optimal learning of causal graphs. arXiv preprint arXiv:1703.02645, 2017a.

[8] Ghassami, Salehkaleybar, Kiyavash, and Elias Bareinboim (2018) Budgeted Experiment Design for Causal Structure Learning (https://causalai.net/r33.pdf)

[9] Zhang, J., Kumor, D., Bareinboim, E. (2020) Causal Imitation Learning with Unobserved Confounders. (https://causalai.net/r66.pdf)

[10] Sachidananda and Brunskill (2017) Online learning for causal bandits. (https://web.stanford.edu/class/cs234/past_projects/2017/2017_Sachidananda_Brunskill_Causal_Bandits_Paper.pdf)

[11] Lattimore, Reid (2016). Causal bandits: Learning good interventions via causal inference. (https://arxiv.org/pdf/1606.03203.pdf)

[12] Yash Shah, Gaurav Didwania, Rupesh, Kumar Saurav (2019) Exploring Online Learning Algorithms for Causal Bandits

[13] The WhyNot Python package documentation (https://whynot.readthedocs.io/en/latest/examples.html#reinforcement-learning)

[14] Chintan Shah, Smruthi Ramesh, Juan Alfaro. Causal Reinforcement Learning (2021).(https://github.com/chnsh/causal_RL/blob/master/covid19_simulator.ipynb)