Table des matières

[Introduction 2](#_Toc42795698)

[L’effet cigogne dans l’apprentissage automatique par renforcement 2](#_Toc42795699)

[Réflexion 2](#_Toc42795700)

[Problématique 2](#_Toc42795701)

[Question sur les noms 3](#_Toc42795702)

[I- Revue de littérature 4](#_Toc42795703)

[A- Une question triviale ? 4](#_Toc42795704)

[a) Dans le cas d’une régression 4](#_Toc42795705)

[b) Dans le cas de l’apprentissage par renforcement 4](#_Toc42795706)

[B- Exemples 5](#_Toc42795707)

[a) Théoriques 5](#_Toc42795708)

[b) Imaginaires 5](#_Toc42795709)

[C- Exemples dans l’apprentissage automatique 6](#_Toc42795710)

[a) Biais induits au sein des algorithmes 6](#_Toc42795711)

[b) Coïncidence 6](#_Toc42795712)

[c) Données d’apprentissage non représentatives (dont Biais de sélection) 6](#_Toc42795713)

[d) Biais de confirmation 7](#_Toc42795714)

[e) Tous biais confondus 7](#_Toc42795715)

[II- Expérimentation 7](#_Toc42795716)

[A- L’agent 8](#_Toc42795717)

[a) Les algorithmes existants 8](#_Toc42795718)

[b) Ajouter une étape de randomisation 9](#_Toc42795719)

[c) Configuration et enregistrement des résultats 9](#_Toc42795720)

[B- L’environnement 9](#_Toc42795721)

[a) L’implémentation 9](#_Toc42795722)

[b) Paramètres à ajouter 9](#_Toc42795723)

[C- Les biais à implémenter 10](#_Toc42795724)

[Données non représentatives 10](#_Toc42795725)

[Inciter au biais de confirmation 10](#_Toc42795726)

[III- Analyse des résultats 10](#_Toc42795727)

[A- Si la réponse est explicite 11](#_Toc42795728)

[B- Si les résultats nécessitent de comparer les chiffres 11](#_Toc42795729)

[C- Biais de la démarche 11](#_Toc42795730)

[a) Données fictives, donc conditionnées à mon imagination 11](#_Toc42795731)

[b) Cas impossibles à tester ou dont les résultats sont difficiles à analyser 11](#_Toc42795732)

[c) “Effet cigogne” : une notion floue 11](#_Toc42795733)

[D- Ouverture 11](#_Toc42795734)

[a) Comparaisons impossibles 11](#_Toc42795735)

[b) Ce qui n’est pas traité 12](#_Toc42795736)

[Conclusion 12](#_Toc42795737)

[Sources 12](#_Toc42795738)

[VRAC 12](#_Toc42795739)

[Résumés de publications 12](#_Toc42795740)

# Introduction

## L’effet cigogne dans l’apprentissage automatique par renforcement

### Réflexion

Tout vient d’une réflexion sur la différence entre corrélation et causalité. C’est un problème qui survient systématiquement lorsqu’on fait des statistiques descriptives à partir de données dont on dispose.

L’apprentissage automatique peut lui aussi subir ce biais, ne serait-ce que parce que le biais vient des données en entrées (malgré une validation croisée).

Comment peut-on répondre à ce biais ?

Avec une randomisation : couper aléatoirement un échantillon en deux, agir sur seulement l’une des moitiés, comparer les résultats.

Comment peut-on répondre à ce biais automatiquement ?

On peut imaginer un algorithme qui va apprendre non pas à partir de données en entrée, mais face à une situation, en lui permettant d’agir dessus.

On peut donc ajouter d’une randomisation au sein d’un algorithme d’apprentissage par renforcement pour améliorer ses résultats, ou du moins pour voir si ses résultats évoluent.

### Problématique

Par “biais”, on considèrera la définition suivante : “une démarche ou un procédé qui engendre des erreurs”.

La définition formelle en statistique de biais, “différence entre la valeur de l’espérance d’un estimateur et la valeur qu’il est censé estimer” correspondrait à un problème de régression. Mais pour un algorithme d’apprentissage par renforcement, l’objectif n’est pas tant d’estimer au plus proche un résultat que de maximiser ce dernier.

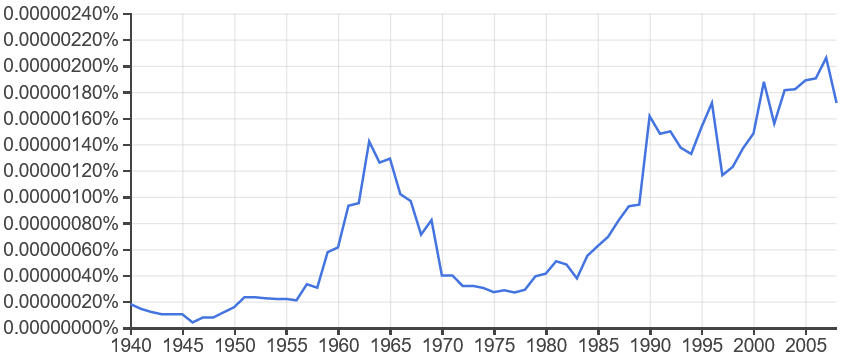
* Peut-on utiliser l’apprentissage par renforcement pour différencier ces corrélations de causalités ?
* Dans quelle mesure ce biais influence-t-il le résultat de prédictions basées sur l’apprentissage automatique ?

### Question sur les noms

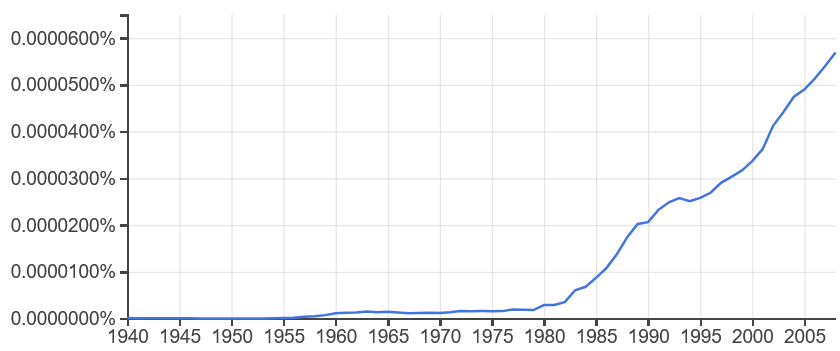
Le français “apprentissage automatique” est-il vraiment synonyme de “machine learning” en anglais ?

Cela n’est pas certain quand on compare les fréquences de ces expressions au cours du temps.

Soit l’expression “apprentissage automatique” était porteuse d’un tout autre sens dans les années 1960, soit il s’agit de deux notions différentes, l’apprentissage automatique englobant le machine learning.



Fréquences de “apprentissage automatique” en français



Fréquences de “machine learning” en anglais

# I- Revue de littérature

## A- Une question triviale ?

### a) Dans le cas d’une régression

Dans le cas d’une régression, la question peut sembler triviale.

Puisqu’une régression consiste à mesurer les corrélations entre toutes les variables afin d’en estimer une à partir des autres, il suffit que certaines variables soient corrélées avec la celle à estimer pour biaiser les résultats. Pour combattre ce biais, il faut qu’un être humain analyse le contexte pour déterminer s’il y a causalité entre les variables.

http://www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/tr-bias.pdf

Cependant, cette étude montre que même dans le cas d’une régression il existe des méthodes permettant de diminuer ce biais

### b) Dans le cas de l’apprentissage par renforcement

Si le cas de la régression semblait trivial, c’est peut-être parce que l’algorithme n’a pas l’occasion d’interagir avec son environnement pour tester ce qui est une corrélation et ce qui est une causalité.

On peut d’ailleurs diiférencier trois types de variables qui construisent la réalité avec laquelle interagit un agent :

* les données intrinsèques à l’environnement (une partie des observations)
* les entrées (les actions de l’agent sur l’environnement)
* Les sorties issues des actions sur l’environnement (la récompense et une partie des observations)

Au sein des algorithmes, ces variables sont découpées de la sorte :

* Les observations (qu’elles dépendent des actions de l’agent ou non)
* Les actions (les décisions prises par l’agent)
* La récompense (ce que l’agent doit maximiser)

Ainsi, contrairement au cas d’une régression, il est déterminé dès le départ sur quoi l’agent peut agir. Il ne peut pas vérifier l’existence de causalité entre deux variables qui ne dépendent pas de lui

## B- Exemples

### a) Théoriques

Sachant que A est corrélé à B, il y a plusieurs explications possibles :

* A cause B
* B cause A
* A cause B et B cause A
* C connu cause A et B
* C inconnu cause A et B
* Coincidence

Si on ne peut observer C, peut-on différencier le cas 5 du cas 6 ? Peut-on estimer que nos données semblent répondre à une variable supplémentaire inconnue à partir d’une certaine quantité de données permettant d’écarter l’idée d’une coincidence ? Si oui, peut-on mesurer cette variable ? (sachant qu’on risque de la confondre avec le vrai bruit statistique)

### b) Imaginaires

Dans le cas où A, la quantité vendue, est corrélé à une variable B

* La quantité vendue cause le résultat net
* Le prix cause la quantité vendue
* Le nombre de ventes cause la fréquentation à venir, qui cause le nombre de ventes à venir
* Les jours d’affluence causent la quantité vendue d’un produit et celle d’un autre produit
* La complémentarité entre la farine et la levure cause une corrélation entre les quantités vendues de l’un et de l’autre
* Une variable n’ayant aucun lien de causalité avec quoi que ce soit (ex: l’horoscope des sagittaires) peut néanmoins se retrouver corrélé avec d’autres variables si on ne dispose pas d’un échantillon suffisament grand. Peut alors exister un biais de sur-apprentissage.

## C- Exemples dans l’apprentissage automatique

### a) Biais induits au sein des algorithmes

https://arxiv.org/pdf/1907.02908.pdf

Quand un agent est développé pour répondre à un besoin spécifique, on peut être tenté de le paramétrer via des connaissances préexistantes afin d’améliorer ses résultats. Cela peut causer des erreurs supplémentaires, en plus de rendre l’algorithme moins généralisable.

#### Exemple de ce problème dans notre expérimentation

Si notre agent doit acheter des marchandises puis les vendre, sa récompense est sa marge sur coûts variables.

Pour lui éviter d’essayer des cas triviaux et a priori contreproductifs, on serait tenté de le paramétrer de telle sorte qu’il ne fixe jamais de prix de vente inférieur au prix d’achat. Ce qui serait une bonne idée sans compter que :

* Un produit d’appel peut être vendu à perte et pourtant améliorer le revenu global
* Si l’agent doit gérer ses stocks, il peut arriver qu’il doive vendre à perte pour déstocker (ex: péremption)
* Vendre ou acheter à perte peut parfois être une obligation légale
  + Exemple : EDF qui achète à un prix plancher l’électricité issue d’énergies renouvelables sur le marché de l’électricité

### b) Coïncidence

Il s’agit d’une question purement statistique. Il suffit d’avoir assez de données.

Peut-on considérer les erreurs qui y sont dues comme dues à un sur-apprentissage ?

### c) Données d’apprentissage non représentatives (dont Biais de sélection)

#### Peut-on tromper l’agent s’il ne peut déterminer l’importance de ses actions dans la récompense finale ?

Si 80% de sa récompense est basée sur 20% de ses actions, l’algorithme mettra plus de temps à estimer l’importance respective de chaque variable.

On peut maximiser ce biais :

* En fournissant une récompense et/ou des observations aggrégées à une granularité trop épaisse
* Nous arrivons alors à une coïncidence et un biais de surapprentissage

#### https://app.wandb.ai/stacey/aprl/reports/Adversarial-Policies-in-Multi-Agent-Settings–VmlldzoxMDEyNzE

Résumé du protocole de cette publication :

* On prend deux agents, A et B, et un jeu compétitif.
* A apprend à jouer à partir de données de véritables joueurs.
* Puis B apprend à jouer contre A

Il en résulte que la meilleure manière pour B de gagner consiste à ne pas jouer. En effet, A n’a appris à jouer que contre des personnes qui savent jouer. B faisant des choses inattendues, A perd tout seul.

Conclusion : l’apprentissage par renforcement gagne sur le long terme face à un programme exclusivement formé sur des données qui ne recouvrent pas assez de cas.

### d) Biais de confirmation

Les algos y sont-ils sensibles ? Causalité au début qui décroit avec le temps, mais l’algo continue dans le sens initial ?

### e) Tous biais confondus

Même s’ils se corrigent facilement et automatiquement dans les algos déjà existants, on peut toujours en mesurer et comparer leurs interties.

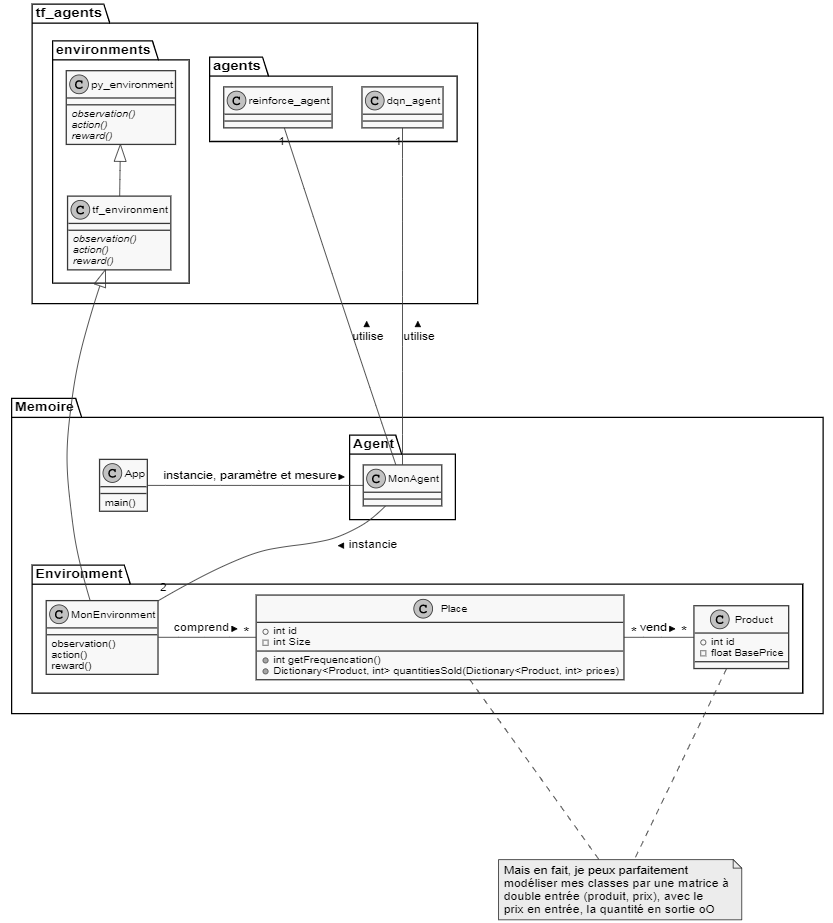
# II- Expérimentation

Faire interagir un agent suivant plusieurs algorithmes d’apprentissage par renforcement pour apprendre face à un environnement biaisé.

Pour créer des situation biaisées, on préfèrera utiliser une librairie permettant de créer un environnement.

J’ai choisi (/ commencé à utiliser) Tensorflow, car la documentation semble claire et bien fournie, et que la librairie implémente plusieurs algorithmes.

Diagramme de classes :



## A- L’agent

### a) Les algorithmes existants

Algorithmes présents dans TF :

* [DQN](https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf)
* [REINFORCE](http://www-anw.cs.umass.edu/%7Ebarto/courses/cs687/williams92simple.pdf)
* [DDPG](https://arxiv.org/pdf/1509.02971.pdf)
* [TD3](https://arxiv.org/pdf/1802.09477.pdf)
* [PPO](https://arxiv.org/abs/1707.06347)
* [SAC](https://arxiv.org/abs/1801.01290)

### b) Ajouter une étape de randomisation

### c) Configuration et enregistrement des résultats

TODO: réussir à les faire marcher, puis écrire une boucle for qui enregistre les résultats de chaque algos pour une configuration de l’environnement donnée

## B- L’environnement

On peut imaginer des prix mis à jour en temps réel par l’agent, l’objectif de l’agent étant de trouver le prix maximisant le résultat net. L’environnement répondrait, pour chaque offre, une demande (un nombre d’achats).

Cas réels qui correspondraient : prix dans un centre commercial connecté, sur un site d’e-commerce, sur un marché à terme en temps réel (financier, de l’électricité, du blé), etc.

### a) L’implémentation

Dans TF, on peut créer deux types d’environnement : py\_environment.PyEnvironment ou tf\_environment.TFEnvironment. Les deux prennent en compte des paramètres similaires. Dans notre exemples :

* Le temps est linéaire et discret
* Action: pour chaque lieu et/ou produit, un prix de vente
* Observation: pour chaque lieu et/ou produit, une demande
* Récompense: la somme, pour chaque lieu et/ou produit, du prix de vente auquel on soustrait le prix d’achat.

### b) Paramètres à ajouter

Afin d’obtenir des résultats plus intéressants, on peut multiplier les paramètres possibles :

* élasticité demande selon offre
* prix plafond et prix plancher
* un seul lieu d’achat/vente ou bien plusieurs avec des prix différents
* les caractéristiques des lieux d’échange (nombre de clients, intérêt pour le produit, prix plafond et plancher)
* un seul ou plusieurs produits
* produits qui répondent ou non au même besoin
* coûts des produits identiques ou non
* coûts des produits qui changent au fil du temps ou non
* quelle part d’aléatoire
* avoir autant de produits que voulu, ou devoir également préciser une quantité à acheter ?

## C- Les biais à implémenter

### Données non représentatives

Un paramètre inconnu (la taille des magasins) est créé, et influence les résultats. Ensuite, on modifie ce paramètre, ou on ajoute des situations en moyenne différente (plus grands ou plus petits), et on observe combien de temps l’algo se laissera berner (aka on mesure son inertie)

A une étape de l’algorithme, arbitrairement jouter ou supprimer des magasins ou produits avec des caractéristiques non représentatives de la population de départ.

Exemples :

* l’expérience était sur les carrefour city, elle inclut par la suite également les carrefour market, d’une taille en moyenne différente. Toutes les quantités varient.
* la chaîne s’étend sur un territoire avec des habitudes de consommation différentes

### Inciter au biais de confirmation

En utilisant le biais du razoir d’Ockham (privilégier les modèles les plus simples peut conduire à oublier une variable) mentionné dans cette publication :

* https://arxiv.org/pdf/cmp-lg/9612001.pdf
* créer un biais du razoir :
  + Créer deux variables corrélées, l’une expliquant, pas l’autre, pour que l’agent en prenne une au hasard
  + Alors qu’en fait, c’était l’autre qui était liée par une relation de causalité à la récompense (ou les deux)

ça peut marcher ? vérifier les implémentations des algos

Ou bien, on verra bien une fois qu’on a testé

# III- Analyse des résultats

Questions auxquelles on souhaitait répondre :

* Quelle est l’efficacité l’agent qui ne peut apprendre qu’avec des données ayant déjà été observées et biaisées (apprentissage supervisé)
* Versus l’efficacité si on pré-entraine l’agent avec des données ayant déjà été observées puis qu’on le laisse se renforcer sans randomisation (apprentissage supervisé et par renforcement)
* Versus l’efficacité si on pré-entraine l’agent avec des données ayant déjà été observées puis qu’on le laisse se renforcer avec randomisation (apprentissage supervisé et par renforcement)

Questions supplémentaires, dans le cas où je développe un algorithme qui effectue explicitement une randomisation pour tester si les corrélations observées sont des causalités :

* Versus l’efficacité de l’apprentissage par renforcement seul avec randomisation
* Versus l’efficacité de l’apprentissage par renforcement seul sans randomisation

## A- Si la réponse est explicite

## B- Si les résultats nécessitent de comparer les chiffres

Pour quels paramètres et quels biais, quels algorithmes obtiennent quel résultat sur un grand nombre d’opérations ?

## C- Biais de la démarche

### a) Données fictives, donc conditionnées à mon imagination

### b) Cas impossibles à tester ou dont les résultats sont difficiles à analyser

### c) “Effet cigogne” : une notion floue

Derrière cette appellation claire en langage naturel, on peut déduire tout un ensemble de biais dans le cas de l’apprentissage automatique. Ce choix de sujet n’est peut-être pas le plus adapté.

## D- Ouverture

### a) Comparaisons impossibles

Seulement des tests d’un agent face à un environnement

#### On ne peut pas extrapoler les comparaisons entre algorithmes en concurrence.

#### Ni comparer avec des résultats obtenus par régression

### b) Ce qui n’est pas traité

#### 1- Le cas de la régression, du clustering et de la classification

#### 2- La loi de goodhart

“Lorsqu’une mesure devient un objectif, elle cesse d’être une bonne mesure”.

Cela semble l’un des principaux biais lors de la mise en oeuvre d’un algorithme d’apprentissage automatique pour répondre à une problématique. Il faut en effet s’assurer qu’il n’y aie pas des manières de maximiser la mesure qui sert d’objectif au détriment de l’objectif réel (effet rebond et effet cobra).

Ce biais et sa réponse, sont cependant moins des questions techniques et intrinsèques aux algorithmes que des questions d’appréciation qualitative de la pertinence des indicateurs et objectifs.

# Conclusion

# Sources

* https://scholar.google.com/scholar?hl=fr&as\_sdt=0%2C5&q=reinforcement+learning+causality
  + En 2009 : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2713351/
  + 2019 : https://arxiv.org/abs/1901.08162
  + 2018 : https://ieeexplore.ieee.org/abstract/document/8115277
  + ? : https://books.google.fr/books?hl=fr&lr=&id=2qt0DgAAQBAJ&oi=fnd&pg=PA295&dq=reinforcement+learning+causality&ots=aypw5lcR00&sig=Buj0QQOXRdF6\_rFoCpeov9HdVYM&redir\_esc=y#v=onepage&q=reinforcement%20learning%20causality&f=false

# VRAC

## Résumés de publications

http://www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/tr-bias.pdf

* **Solution** : réduction de la variance
* **Problème** : coûteux en ressources (pour l’époque ?)
* **Comment l’utiliser** : La question a déjà été traitée, mais il y a 25 ans, et seulement pour les régressions, et ça coûtait trop de ressources.
* **Peut-être toujours d’actualité pour le renforcement ?** De plus, si on combine une réduction de variance avec un algo plus récent et moins coûteux, ça pourrait être pas mal.

“Randomization is paradoxical, because at first glance it seems to increase variance by deliberately introducing variation into the splits in the decision tree.”

https://arxiv.org/pdf/cmp-lg/9612001.pdf

https://people.csail.mit.edu/malte/pub/papers/2019-iclr-variance.pdf

https://dl.acm.org/doi/10.5555/3305381.3305400

à propos de la baisse de la variance grâce à l’algo DQN

https://link.springer.com/article/10.1007/s11219-012-9180-0

sur le biais de confirmation