A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem

Travail présenté à M. Fermanian et M. Pham dans le cadre du cours *: Machine Learning for finance*

Travail réalisé par WORICK Steven, LANDRY Donovan et WISNIEWSKI Romain
MS FINANCE RISK MANAGEMENT
Année 2021 - 2022

Avril 2022

Table des matières

1 Introduction

La gestion de portefeuille est le processus décisionnel qui réaffecte en permanence une somme d'argent à un certain nombre de produits d'investissement financiers différents, afin de maximiser les rendements tout en minimisant les risques. Les méthodes traditionnelles de gestion de portefeuille peuvent être classées en quatre catégories : « Follow-the-Winner», « Follow-the-Looser », « Pattern-Matching » et « Meta-Learning ».

Il existe également des approches de deep machine-learning aux thématique de trading sur les marchés financiers. Cependant, bon nombre d'entre eux essaient de prédire les mouvements ou les tendances des prix. Avec les prix historiques de tous les actifs comme données, un réseau de neurones peut produire un vecteur qui prédit des prix des actifs pour la période suivante. Cette idée est simple à mettre en œuvre, car il s'agit d'un apprentissage supervisé, ou plus précisément d'un problème de régression. La performance de ces algorithmes basés sur la prédiction des prix dépend du degré de précision de la prédiction.Les précédentes tentatives réussies de modèles d'apprentissage sur le trading algorithmique, sans prédire les prix futurs, traitent du problème de Reinforcement Learning (RL). Ces algorithmes produisent des signaux de trading discrets sur un actif. Étant limités à la négociation d'un seul actif, ils ne s'appliquent pas aux problèmes généraux de gestion de portefeuille. Le Deep Reinforcement learning attire beaucoup l'attention ces derniers temps en raison de ses performances remarquables dans les jeux vidéo (Mnih et al., 2015) et les jeux de société. Ce sont des problèmes avec des espaces d'action discrets, et ne peuvent pas être directement appliqués aux problèmes de sélection de portefeuille, où les actions sont continues. Afin de tirer pleinement parti, les algorithmes de trading doivent être échelonnable. Dans le cadre général continu du deep RL, les algorithmes, ont été introduits en 2014. La production continue dans ces acteurs-critiques les algorithmes sont obtenus par une fonction de politique d'action approximative de réseau de neurones, et un second réseau neurones est formé comme estimateur de la fonction de récompense. Cependant, l'entraînement des deux réseaux de neurones s'avère difficile, voire instable.

A travers ce papier de recherche, nous allons nous pencher sur l'application du reinforcement learning à la gestion de portefeuille. Le coeur de ce papier traite de « Ensemble of Identical Independent Evaluators » (EIIE). Un Identical Independent Evaluators (IIE) est un réseau de neurones dont la principale tâche est d'examiner l'histoire d'un actif et d'évaluer sa croissance potentielle pour l'avenir. C'est à l'aide du Reinforcement learning que les pondérations du portefeuille sont définies et déterminent les actions menées sur le marché. Par exemple, un actif ayant une pondération plus importante sera acheté, et celui ayant une pondération réduite sera vendu. Outre l'historique du marché, les pondérations du portefeuille de la période de négociation précédente sont également utilisées dans l'EIIE. Notre algorithme de RL doit tenir compte de l'effet du coût de la transaction sur son portefeuille. A cet effet, les pondérations de portefeuille de chaque période sont enregistrées dans le Portfolio Vector Memory (PVM). L'EIIE est entraîné à l'aide d'un système d'apprentissage stochastique par lots en ligne (OSBL), qui est compatible avec la formation pré-métier et la formation en ligne pendant les back-tests ou le trading en ligne. Ayant une fonction de récompense explicite, l'EIIE évolue, en formation, le long du gradient ascendant de la fonction. Trois types différents d'IIE sont testés dans ce travail, un Convolutional neural network (CNN), un recurrent neural network (RNN) et un modèle Long short-term memory (LSTM). Étant une approche entièrement machine-learning, le cadre n'est pas limité à un marché particulier. Pour examiner la validité et la rentabilité du projet, ce dernier est appliqué à un exchange de crypto-monnaie, Poloniex.com. Un ensemble de crypto-monnaie sont présélectionnées par leur classement dans le volume d'échange sur un intervalle de temps juste avant une expérience. De plus, afin de valider ce travail de recherche nous étudierons trois expériences de back-test sur des durées bien séparées sont effectuées dans un intervalle d'échange de 30 minutes. Lors trois phases de back-testing les algorithmes EIIE ont nettement dépassé toutes les autres stratégies parues auparavant des divers papiers scientifiques.

Nous nous intéresserons dans un premier temps à la définition de la problématique que cherche à résoudre ce papier ainsi qu'au travail porté sur les données. Dans un deuxième temps, nous étudierons les méthodes de reinforcement learning et les algorithmes de réseaux de neurones. Pour finir, nous nous pencherons sur les résultats obtenus et nous proposerons une critique constructive sur le papier de recherche.

2 Le modèle

Le portefeuille se compose de m actifs. Le cours de clôture de tous les actifs constituent le vecteur de prix pour la période t, v_t. En d'autres termes, le i-ème élément de v_t, $v_{i,t}$, est le prix de clôture du i-ème asset sur la periode t. De même, $v_t^{(hi)}$ et $v_t^{(lo)}$ indiquent les prix les plus élevés et les plus bas de la période. Le premier actif du portefeuille est spécial, c'est-à-dire la devise cotée, appelée cash pour le reste de l'article. Étant donné que les prix de tous les actifs sont cotés en cash, les premiers éléments de v_t, $v_{0,t}^{(hi)}$ et $v_{0,t}^{(lo)}$) sont toujours égaux à 1.

Pour les marchés continus, les éléments de v_t sont les prix d'ouverture pour la période $t+1$ aussi bien que les prix de clôture pour la période t. Le vecteur de prix relatif de la période d'échange t, y_t, est défini comme la division par élément par élément de v_t par $v_t - 1$:

$$y_t := v_t \oslash v_{t-1} = \left(1, \frac{v_{1,t}}{v_{1,t-1}}, = v_{2,t-1}, ..., \frac{v_{m,t}}{v_{m,t-1}}\right)^\intercal$$

Le vecteur de prix relatif peut être utilisé pour calculer la variation de la valeur totale du portefeuille au cours d'une période. Si p_{t-1} est la valeur du portefeuille au début de la période t, sans tenir compte du coût de transaction,

$$p_t := p_{t-1} y_t \cdot w_{t-1},$$

où w_{t-1} est le vecteur de pondération du portefeuille. Le taux de rendement pour la période t est alors :

$$\rho_t := \frac{p_t}{p_{t-1}} - 1 = y_t \cdot w_{t-1} - 1,$$

Le taux de rendement logarithmique correspondant est :

$$r_t := \ln \frac{p_t}{p_{t-1}} = \ln y_t \cdot w_{t-1},$$

S'il n'y a pas de coût de transaction, la valeur finale du portefeuille sera :

$$p_f := p_0 \exp(\sum_{t=1}^{t_f+1} r_t) = p_0 \prod_{t=1}^{t_f+1} y_t \cdot w_{t-1},$$

où p_0 est le montant de l'investissement initial. Le travail d'un gestionnaire de portefeuille est de maximiser p_f pour une période donnée.

Dans un scénario réel, l'achat ou la vente d'actifs dans un marché n'est pas gratuit. Le coût est provient normalement de frais de commission.

Le vecteur de portefeuille au début de la période t est w_t. En raison de l'évolution des prix sur le marché, à la fin de la même période, les pondérations évoluent vers :

$$w_t^{'} = \frac{y_t \odot w_{t-1}}{y_{t-1} \cdot w_{t-1}}$$

où $\odot$ est la multiplication elément par élément. La mission du gestionnaire de portefeuille à la fin de la période t consiste à réallouer le vecteur de portefeuille de $w_t^{'}$ à w_t en vendant et en achetant les actifs appropriés. En payant tous les frais de commission, cette opération de réallocation réduit la valeur du portefeuille d'un facteur μ_t, $\mu_t \in (0, 1]$, et sera appelé le facteur résiduel de la transaction :

On note $p_t - 1$ comme valeur du portefeuille au début de la période t et $p_t^{'}$ à la fin, $p_t = \mu_t p_t^{'}$.

Le taux de rendement et le taux de rendement logarithmique sont maintenant

$$\rho_t = \frac{p_t}{p_{t-1}} - 1 = \frac{\mu_t p_t^{'}}{p_{t-1}} = \mu_t y_t \cdot w_{t-1} - 1,$$

$$r_t = \ln \frac{p_t}{p_{t-1}} = \ln(\mu_t y_t \cdot w_{t-1}),$$

et t la valeur finale du portefeuille devient

$$p_t = p_0 \exp(\sum_{t=1}^{t_f+1} r_t) = p_0 \prod_{t=1}^{t_f+1} \mu_t y_t \cdot w_{t-1}, ,$$

Le reste du problème est de déterminer le facteur de transaction résiduel μ_t. Pendant la réaffectation du portefeuille du poids à poids, une partie ou la totalité de l'actif i doit être vendue, si $p_t^{'} w_{t,i}^{'} > p_t w_{t,i}$ or $w_t^{'} > \mu_t w_{t,i}$. Le montant total de cash est obtenu par toutes les ventes :

$$(1 - c_s) p_t^{'} \sum_{i=1}^{m} (w_{t,i}^{'} - \mu_t w_{t,i})^+$$

Finalement on obtient un théorème qui permet d'obtenir le montant de réserve de cash nécessaire à l'achat de nouveaux assets tout en prenant en compte le taux de comission à l'achat :

Théorème 1. Notons

$$f(\mu) := \frac{1}{1 - c_p w_{t,0}} \left[1 - c_p w_{t,0}^{'} - (c_s + c_p - c_s c_p) \sum_{i=1}^{m} (w_{t,i}^{'} - \mu_t w_{t,i})^+ \right],$$

la sequence $\left\{ \tilde{\mu}_t^{(k)} \right\}$, défini par :

$$\left\{ \tilde{\mu}_t^{(k)} \middle| \tilde{\mu}_t^{(0)} = \mu_\odot \text{ and } \tilde{\mu}_t^{(k)} = f(\tilde{\mu}^{k-1}), k \in \mathbb{N}_0 \right\}$$

pour tout $\mu\odot \in [0, 1]$

Ce théorème fournit un moyen d'estimer le solde de la transaction facteur μ_t à une précision arbitraire. Tout au long de ce travail, un seul taux de commission constant pour la vente et l'achat de tous les actifs non monétaires est utilisé, cs = cp = 0,25 %, le taux maximum chez Poloniex.

Le but de l'agent algorithmique est de générer une séquence temporelle de vecteurs de portefeuille $w_1, w_2, \ldots, w_t, \ldots$ afin de maximiser le capital accumulé en tenant compte du coût de transaction.

Dans ce travail, les transactions de back-test ne sont prises en compte que lorsque l'agent fait semblant d'être de retour dans le temps à un moment donne de l'histoire du marché, ne connaissant aucune information de marché "future". Comme exigence pour les expériences de back-test, les deux hypothèses suivantes sont imposées :

1. Zéro glissement : la liquidité de tous les actifs du marché est suffisamment élevée pour que chaque transaction puisse être effectuée immédiatement au dernier prix lorsqu'un ordre est passé.
2. Aucun impact sur le marché : le capital investi par l'agent commercial de logiciels est si insignifiant qu'il n'a aucune influence sur le marché.

Dans un environnement de trading réel, si le volume des transactions sur un marché est suffisamment élevé, ces deux hypothèses sont proches de la réalité.

3 Traitement des données

Les back-tests de trading se sont fait sur le site Poloniex. C'est une plateforme d'échange de crypto-monnaies où il y avait environ 80 crypto-monnaies échangeables en mai 2017. Cependant, seul un sous-ensemble d'actifs est pris en compte par l'algorithme. Les auteurs ont décidé de sélectionner les onze plus grosses valorisations sur le marché des crypto-monnaies. De ce fait, le portefeuille est constitué de 11 crypto-monnaies et d'une réserve d'argent dans une devise internationale. Ce choix se justifie par le fait qu'un volume plus important implique une meilleure liquidité du marché d'un actif. Ce qui signifie que la situation du marché est plus proche de l'hypothèse 1. Des volumes plus élevés suggèrent également que l'investissement peut avoir moins d'influence sur le marché, établissant un environnement plus proche de l'hypothèse 2. Considérant la fréquence de trading relativement élevée (30 minutes) par rapport à certains algorithmes de trading journalier, la liquidité et la taille du marché sont des caractéristiques particulièrement importantes. En effet, le marché des crypto-monnaies n'est pas stable. Certains actifs peuvent avoir une augmentation soudaine ou une baisse de volume dans un laps de temps. Par conséquent, le volume pour la présélection des actifs est d'une durée plus longue, par rapport à la période d'échange. Dans ces expériences, des volumes de 30 jours sont utilisés.

Une fois la sélection des actifs effectué, les données historiques sur les prix sont utilisés comme input dans le réseau de neurones, ce qui permet d'obtenir un vecteur de portefeuille. Avant d'être utilisées comme input dans le réseau de neurones à la fin de la période t, les données sont stockées dans un tensor de rang 3 avec une dimension (f, n, m), où m est le nombre de crypto-monnaies pré-sélectionnées, n est le nombre de période avant la date t, et f = 3 est le nombre de feature. Etant donné que les prix les plus éloignés dans le temps ont beaucoup moins de corrélation et d'impact sur le cours actuel, il est donc décidé que n = 50 (une journée et une heure). Le nombre d'actif est choisit par rapport à la capitalisation des actifs et les features correspondent au prix de clôture, le prix le plus élevé et le prix le plus bas de chaque actif dans l'intervalle de temps. Étant donné que seuls les variations de prix détermineront la performance de la gestion du portefeuille, tous les prix de clôture présent au sein du tensor seront normalisés.

Certaines observations sont manquantes dans les historiques des actifs sélectionnés. Cette absence est due au fait que ces crypto-monnaies venait d'apparaître rien 2017. Les données manquantes sont marqués comme Not A Numbers (NANs). Pour répondre à cette problématique, les chercheurs ont décidés de générer de fausses séries de prix avec un faible taux de décroissance, approximativement 0.01, afin que les réseaux de neurones évitent de

sélectionner les données manquantes lors de la phase d'entraînement. Cependant, il s'est avéré que les algorithmes « Neural Network » été biaisé par cette génération de données. Pour cette raison, les données générées auront la même valeur que la première observation historique.

4 Reinforcement Learning

Afin de représenter l'agent et l'environnement, nous allons représenter un état d'environnement par son prix du fait que l'on a accès à tout l'historique du marché.Par conséquent, des schémas de sous-échantillonnage pour l'information sur l'historique des ordres sont utilisés pour simplifier à l'avenir la représentation de l'état de l'environnement de marché.

Ces schémas comprennent la présélection des actifs décrite à la section 3, l'extraction de caractéristiques périodiques et la coupure de l'historique. L'extraction de caractéristiques périodiques discrétise le temps en périodes, et ensuite extrait les prix les plus élevés, les plus bas et les prix de clôture de chaque période. Dans le cadre actuel, cette influence est encapsulée en considérant w_{t-1} comme une partie de l'environnement et en l'introduisant dans la politique d'action de l'agent. Ainsi, l'état à t est représenté par la paire X_t et w_{t-1}. L'état s_t se compose de deux parties, l'état externe représenté par le tenseur des prix, X_t , et l'état interne représenté par le vecteur de portefeuille de la dernière période, w_{t-1} . Nous pouvons donc représenter le vecteur de portefeuille de la dernière période : $s_t = (X_t, w_{t-1})$. Parce qu'en vertu de l'hypothèse 2 de la section 2, le montant du portefeuille est négligeable par rapport au volume total de transactions du marché, p_t n'est pas inclus dans l'état interne.

$$R(s_1, a_1, ..., s_{t_f}; a_{t_f}; s_{t_{f+1}}) := \frac{1}{t_f} ln \frac{p_f}{p_0} = \frac{1}{t_f} \sum_{t=1}^{t_f+1} ln(\mu_t y_t \cdot w_{t-1}) \frac{1}{t_f} \sum_{t=1}^{t_f+1} r_t$$

Avec cette fonction de récompense, le cadre actuel présente deux distinctions importantes par rapport à de nombreux autres problèmes de RL. La première est que les récompenses épisodiques et cumulées sont exactement exprimées. En d'autres termes, la connaissance du domaine de l'environnement est bien maîtrisée et peut être pleinement exploitée par l'agent. Le fait d'avoir une fonction action-valeur définie justifie encore plus l'approche de l'exploitation complète, puisque l'exploration dans d'autres problèmes de RL sert principalement à essayer différentes classes de fonctions action-valeur. D'autre part, sans exploration, les optima locaux peuvent être évités par une initialisation aléatoire des paramètres de la politique, ce qui sera discuté ci-dessous.

Une politique est un mapping de l'espace d'état vers l'espace d'action, $\pi : S \to A$. Avec une exploitation complète dans le cadre actuel, une action est produite de manière déterministe par la politique à partir d'un état. La politique optimale est obtenue en utilisant un algorithme d'ascension de gradient. Pour ce faire, une politique est spécifiée par un ensemble de paramètres θ, et $a_t = \pi_\theta(s_t)$. La métrique de performance de π_θ pour l'intervalle de temps $[0, t_f]$ est définie comme la fonction de récompense correspondante (21) de l'intervalle, $J_{[0,t_f]}(\pi_\theta) = R(s_1, \pi_\theta(s_1), ..., s_{t_f}; \pi_\theta s_{t_f}; s_{t_{f+1}})$. Avec la formule du gradient descent sous cette forme : $\theta \to \theta + \lambda \nabla_\theta J_{[0,t_f]}(\pi_\theta)$

Cette approche par mini-batch de l'ascension de gradient permet également l'apprentissage en ligne, ce qui est important dans le trading en ligne où de nouveaux historiques de marché arrivent constamment à l'agent. L'apprentissage automatique en ligne est une méthode d'apprentissage automatique dans laquelle les données deviennent disponibles dans un ordre séquentiel et sont utilisées pour mettre à jour le meilleur prédicteur pour les données futures à chaque étape, par opposition aux techniques d'apprentissage par lots qui génèrent le meilleur prédicteur en apprenant sur l'ensemble des données d'apprentissage en une seule fois.

5 Policy Networks

Les fonctions de politique $\pi\theta$ seront construites à l'aide de trois réseaux neuronaux profonds différents. Avec trois innovations importantes, la topologie de mini-machine inventée pour cibler le problème de gestion de portefeuille ; la mémoire de vecteur de portefeuille, et un schéma d'apprentissage en ligne stochastique en mini-batch.

Dans les deux figures, un exemple hypothétique de vecteur de portefeuille de sortie est utilisé, tandis que la dimension du tenseur des prix et donc le nombre d'actifs sont des valeurs réelles déployées dans les expériences. et donc le nombre d'actifs sont des valeurs réelles déployées dans les expériences. Le site dernières couches cachées sont les scores de vote pour tous les actifs non monétaires. Les résultats softmax de ces scores et d'un biais d'argent deviennent les poids réels du portefeuille correspondant. Pour que le le réseau neuronal prenne en compte le coût de

transaction, le vecteur de portefeuille de la dernière période, w_{t-1}, est inséré dans les réseaux juste avant la couche de vote.

Une caractéristique commune essentielle de ces trois réseaux est que les réseaux circulent indépendamment pour les m actifs alors que les paramètres du réseau sont partagés entre ces flux. Ces flux sont comme des réseaux indépendants mais identiques de plus petite envergure, qui observent et évaluent séparément les différents actifs non monétaires. Ils ne s'interconnectent qu'au niveau de la fonction softmax, juste pour s'assurer que leurs poids de sortie sont non négatifs et que leur somme est égale à l'unité. Nous appelons ces flux des mini-machines ou plus formellement des Evaluateurs Indépendants Identiques (IIE), et cette caractéristique topologique Ensemble d'IIE (EIIE) surnommée approche mini-machine. Dans la figure 1, un EIIE est simplement une chaîne de convolution avec des noyaux de hauteur 1, tandis que dans la figure 2, il s'agit d'un LSTM ou d'un RNN de base prenant en entrée l'historique des prix d'un seul actif comme entrée. L'EIIE améliore considérablement les performances de la gestion de portefeuille. Se souvenant de la performance historique des actifs individuels, un réseau intégré de la version précédente est plus réticent à investir de l'argent dans un actif historiquement défavorable, même si cet actif a un avenir beaucoup plus prometteur. D'autre part, sans être conçu pour révéler l'identité de l'actif assigné, un réseau intégré est capable de juger de sa hausse et de sa baisse potentielles en se basant simplement sur la base d'événements plus récents.

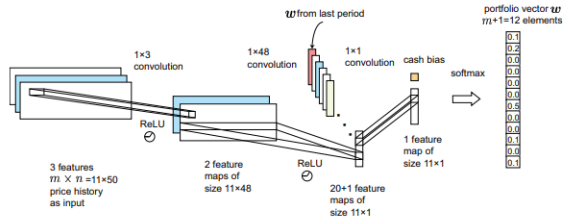


FIGURE 1 – *CNN dans l'EIIE*

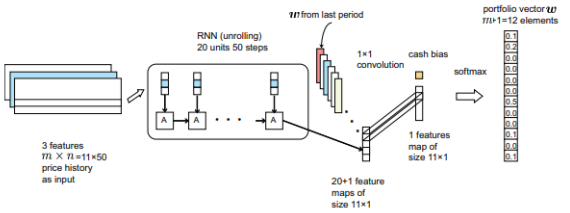


FIGURE 2 – *RNN dans l'EIIE*

D'un point de vue pratique, l'EIIE présente trois autres avantages cruciaux par rapport à un réseau intégré.

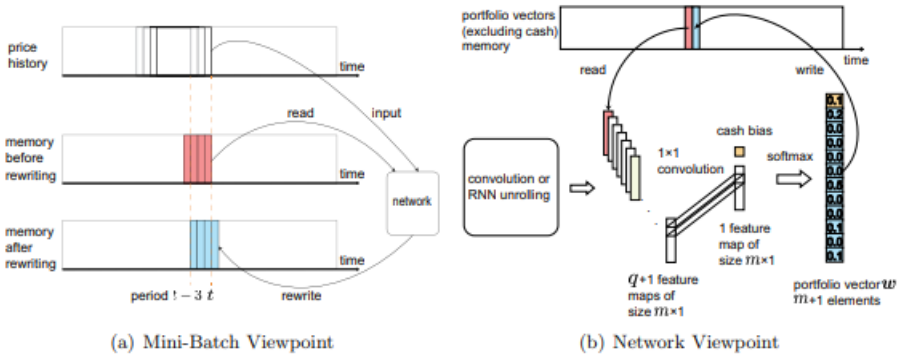
Le premier est l'évolutivité en nombre d'actifs. Les mini-machines étant toutes identiques avec des paramètres partagés, le temps d'apprentissage d'un ensemble évolue de façon à peu près linéaire avec m. Le deuxième avantage est l'efficacité de l'utilisation des données. Pour un intervalle d'historique de prix, une mini-machine peut être entraînée m fois sur différents actifs. Le dernier avantage est la plasticité de la collecte des actifs. Puisque la capacité d'évaluation des actifs d'un IIE est universelle sans être limitée à un actif particulier, une EIIE peut mettre à jour son choix d'actifs et/ou la taille du portefeuille en temps réel, sans avoir à réentraîner le réseau à partir de zéro.

Afin que l'agent de gestion de portefeuille minimise les coûts de transaction en se limitant à des changements importants entre des vecteurs de portefeuille consécutifs, la sortie des poids de portefeuille de la période de négociation précédente est entrée dans les réseaux.

Une mémoire de vecteurs de portefeuille (PVM) dédiée, est introduite pour stocker les sorties du réseau. Comme le montre la figure 3, la PVM est une pile de vecteurs de portefeuille dans l'ordre chronologique.

Avant toute formation du réseau, la PVM est initialisée avec des poids uniformes. À chaque étape de l'apprentissage, un réseau de politiques charge le vecteur de portefeuille de la période précédente à partir de l'emplacement

de mémoire à $t-1$, et écrase la mémoire à t avec sa sortie. Le partage d'une seule pile de mémoire permet d'entraîner un réseau de manière simultanée par rapport à des points de données dans un mini-lots, ce qui améliore considérablement l'efficacité de l'entraînement.



(a) Mini-Batch Viewpoint (b) Network Viewpoint

FIGURE 3 – *Portfolio vectory memory*

Avec l'introduction de la mémoire de sortie du réseau, l'apprentissage en mini-batch devient plausible, bien que le cadre d'apprentissage nécessite des entrées séquentielles. Dans cette optique, nous proposons ici un schéma d'apprentissage par lots stochastique en ligne (OSBL). À la fin de la tième période, le mouvement des prix de cette période sera ajouté à l'ensemble d'apprentissage. Une fois que l'agent a terminé ses commandes pour la période $t+1$, le réseau de politiques sera entraîné contre N_b mini-lots choisis aléatoirement dans cet ensemble. Un lot commençant à la période $t_b < t-n_b$ est choisi avec une probabilité géométriquement distribuée : $P_\beta(t_b) = \beta(1-\beta)^{t-t_b-n_b}$.

Où $\beta \in (0,1)$ est le taux de décroissance de la probabilité déterminant la forme de la distribution de probabilité et l'importance des événements récents du marché, et nb est le nombre de périodes dans un mini-batch.

6 Expérience

Les outils développés jusqu'à présent dans l'article sont examinés dans trois expériences de back-test de différentes périodes avec les trois réseaux politiques sur l'échange de crypto-monnaie Poloniex. Les résultats sont comparés à de nombreuses stratégies de sélection de portefeuille bien établie. La principale mesure financière comparée est la valeur du portefeuille ainsi que le maximum drawdown et le ratio de Sharpe

Différentes mesures sont utilisées pour mesurer la performance d'une stratégie de sélection de portefeuille particulière. La mesure la plus directe de succès d'une gestion de portefeuille sur une période donnée est la valeur cumulée du portefeuille (APV), $p_t = p_t/p_0$, où $p_0 = 1$.

Un inconvénient majeur de l'APV est qu'il ne mesure pas les facteurs de risque, puisqu'il se contente de résumer tous les rendements périodiques sans tenir compte de la fluctuation de ces rendements. Une seconde métrique, le ratio de Sharpe, est utilisée pour prendre en compte le risque. Le ratio est un rendement moyen ajusté au risque, défini comme la moyenne du rendement sans risque par son écart-type. Bien que le SR tienne compte de la volatilité des valeurs du portefeuille, elle traite également les mouvements à la hausse et à la baisse. Dans ces experiences, l'actif sans risque est Bitcoin. Comme la devise cotee est egalement le Bitcoin, le rendement sans risque est nul.

En réalité, la volatilité à la hausse contribue a des rendements positifs, mais a la baisse a la perte. Afin de mettre en évidence la déviation vers le bas, le Maximum Drawdown (MDD) est également considéré. MDD est la plus grande perte d'un pic a un creux, et mathématiquement : $D = \max_{\tau>t} \dfrac{p_t - p_\tau}{p_t}$

Les performances des trois réseaux politiques EIIE proposés dans le présent document seront comparées a celles du CNN intégré (iCNN) (Jiang et Liang, 2017), à plusieurs stratégies basées sur des modèles bien connues ou récemment publiées et à trois références.Les trois indices de référence sont le Best Stock, l'actif avec le plus de fAPV sur l'intervalle de back-test, l'Uniform Buy and Hold (UBAH), une approche de gestion de portefeuille répartissant simplement de manière égale le fonds total dans les actifs présélectionnés et les détenant sans faire aucun les achats ou les ventes jusqu'au bout, et portefeuilles à rééquilibrage constant uniforme (UCRP).

	2016-09-07 to 2016-10-28			2016-12-08 to 2017-01-28			2017-03-07 to 2017-04-27		
Algorithm	MDD	fAPV	SR	MDD	fAPV	SR	MDD	fAPV	SR
CNN	0.224	**29.695**	**0.087**	0.216	**8.026**	**0.059**	0.406	31.747	0.076
bRNN	0.241	13.348	0.074	0.262	4.623	0.043	0.393	**47.148**	**0.082**
LSTM	0.280	6.692	0.053	0.319	4.073	0.038	0.487	21.173	0.060
iCNN	**0.221**	4.542	0.053	0.265	1.573	0.022	0.204	3.958	0.044
Best Stock	0.654	1.223	0.012	0.236	1.401	0.018	0.668	4.594	0.033
UCRP	0.265	0.867	-0.014	**0.185**	1.101	0.010	**0.162**	2.412	0.049
UBAH	0.324	0.821	-0.015	0.224	1.029	0.004	0.274	2.230	0.036
Anticor	0.265	0.867	-0.014	0.185	1.101	0.010	0.162	2.412	0.049
OLMAR	0.913	0.142	-0.039	0.897	0.123	-0.038	0.733	4.582	0.034
PAMR	0.997	0.003	-0.137	0.998	0.003	-0.121	0.981	0.021	-0.055
WMAMR	0.682	0.742	-0.0008	0.519	0.895	0.005	0.673	6.692	0.042
CWMR	0.999	0.001	-0.148	0.999	0.002	-0.127	0.987	0.013	-0.061
RMR	0.900	0.127	-0.043	0.929	0.090	-0.045	0.698	7.008	0.041
ONS	0.233	0.923	-0.006	0.295	1.188	0.012	0.170	1.609	0.027
UP	0.269	0.864	-0.014	0.188	1.094	0.009	0.165	2.407	0.049
EG	0.268	0.865	-0.014	0.187	1.097	0.010	0.163	2.412	0.049
B^K	0.436	0.758	-0.013	0.336	0.770	-0.012	0.390	2.070	0.027
CORN	0.999	0.001	-0.129	1.000	0.0001	-0.179	0.999	0.001	-0.125
M0	0.335	0.933	-0.001	0.308	1.106	0.008	0.180	2.729	0.044

FIGURE 4 – *Performances des trois réseaux neurones EIIE, d'un réseau intégré et de certaines stratégies traditionnelles de sélection de portefeuille dans trois expériences de back-test différentes sur l'échange de crypto-monnaie Poloniex. Les indicateurs de performance sont le retrait maximal (MDD), la valeur cumulée finale du portefeuille (fAPV) dans l'unité du montant initial du portefeuille (pf/p0) et le ratio Sharpe (SR). Les algorithmes audacieux sont les réseaux EIIE introduits dans cet article, nommés d'après les structures soulignées de leurs IIE. Par exemple, bRNN est l'EIIE de la figure 3 utilisant des évaluateurs RNN de base. Trois repères (italiques), le CNN intégré (iCNN) proposé précédemment par les auteurs, et certaines stratégies récemment révisées (Li et coll., 2015a ; Li et Hoi, 2014) sont également testés. Les algorithmes du tableau sont divisés en cinq catégories, le réseau neuronal sans modèle, les repères, les stratégies de suivi des perdants, les stratégies de suivi des gagnants et les stratégies d'appariement des modèles ou d'autres stratégies. La meilleure performance dans chaque colonne est mise en évidence en caractères gras. Les trois EIE ont nettement surpassé tous les autres algorithmes des colonnes fAPV et SR, montrant la rentabilité et la fiabilité de la solution d'apprentissage automatique EIIE au problème de gestion de portefeuille.*

La figure 4 montre les scores de performance fAPV, SR et MDD des réseaux de politique EIIE ainsi que des stratégies comparées pour les trois intervalles de back-test repertoriés dans la figure 4. En termes de fAPV ou SR, l'algorithme le plus performant dans les Back-tests 1 et 2 sont le CNN EIIE dont la richesse finale est plus de deux fois supérieure a celle du deuxième de la première expérience. Les trois premiers gagnants de ces deux mesures dans tous les back-tests sont occupés par les trois réseaux EIIE, ne perdant que la mesure MDD. Ce résultat démontre la forte rentabilité et la cohérence du cadre d'apprentissage automatique EIIE actuel.

En ne considérant que fAPV, les trois EIIE surpassent les meilleurs actifs dans les trois back-tests, tandis que le seul algorithme basé sur un modèle le fait est RMR a la seule occasion du Back-Test 3. En raison du taux de commission élevé de 0,25% et la fréquence de trading demi-horaire relativement élevée, de nombreuses stratégies traditionnelles ont de mauvaises performances. En particulier dans le Back-Test 1, toutes les stratégies basées sur un modèle ont des rendements négatifs, avec un fAPV inférieur a 1 ou des SR négatifs équivalents. D'autre part, les EIIE sont capables d'atteindre des rendements au moins quadruples en 20 jours dans différentes conditions de marché.

Les figures 5, 6 et 7 tracent l'APV en fonction du temps dans les trois back-tests respectivement pour les réseaux CNN et bRNN EIIE, deux benchmarks sélectionnés et deux stratégies basées sur un modèle. Les benchmarks Best Stock et UCRP sont deux bons representants du marché. Dans les trois expériences, les EIIE CNN et bRNN ont battu le marché tout au long des back-tests, tandis que les stratégies traditionnelles ne peuvent y parvenir que dans la seconde moitié du Back-Test 3 et très brièvement ailleurs.

7 Conclusion & Critiques

Cet article propose un cadre extensible d'apprentissage par renforcement résolvant le problème général de gestion de portefeuille financier. Inventé pour faire face aux entrées de marché multicanaux et aux pondérations de portefeuille de sortie directes en tant qu'actions de marché, le cadre peut être integré à différents réseaux de neurones profonds et est évolutif de manière linéaire avec la taille du portefeuille. Pour prendre en compte le coût de transaction lors de la formation des réseaux de politique, le cadre comprend une mémoire de poids de portefeuille, la PVM, permettant à l'agent de gestion de portefeuille à apprendre à limiter les ajustements surdimensionnés entre

des actions consécutives, tout en évitant le problème de disparition du gradient auquel sont confrontés de nombreux réseaux récurrents.

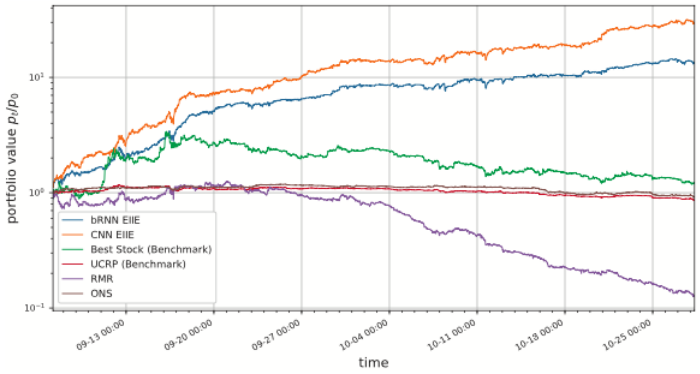


FIGURE 5 – *Back-Test 1 : 2016-09-07-4 :00 to 2016-10-28-8 :00 (UTC). Les valeurs cumulées du portefeuille (APV, p_t/p_0) sur l'intervalle du Back-Test 1 pour le CNN et les EIE RNN de base, le Best Stock, le UCRP, le RMR et l'ONS sont tracées à l'échelle log-10 ici. Les deux EIIL sont en tête tout au long de la période, en croissance constante seulement avec quelques incidents de retrait*

Le système OSBL régit le processus d'apprentissage en ligne, de sorte que l'agent peut digérer en permanence les informations de marché entrantes constantes tout en négociant. Enfin, l'agent a été formé en utilisant une méthode déterministe de gradient de politique, visant à maximiser la richesse accumulée comme fonction de récompense de renforcement.

La rentabilité du cadre dépasse toutes les méthodes traditionnelles de sélection de portefeuille étudiées, comme le démontre l'article par les résultats de trois expériences de back- test sur différentes périodes sur un marché de crypto-monnaie. Dans ces expériences, le cadre a été réalisé en utilisant trois réseaux de soulignement différents, un CNN, un RNN de base et un LSTM. Les trois versions ont obtenu de meilleurs resultats en termes de valeur finale du portefeuille accumulé que les autres algorithmes de trading en comparaison. Les reseaux EIIE ont également monopolisé les trois premieres positions du score ajusté au risque dans les trois tests, indiquant la cohérence du cadre dans ses performances. Une autre solution d'apprentissage par renforcement profond, précédemment introduite par les auteurs, a également été évaluée et comparée dans les mêmes paramètres, perdant également au profit des réseaux EIIE.

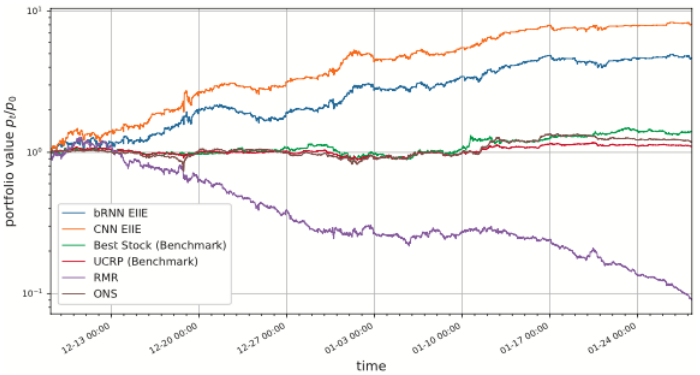


FIGURE 6 – *Back-Test 2 : 2016-12-08-4 :00 to 2017-01-28-8 :00 (UTC), l'échelle de temps logarithmique accumulé. Il s'agit de la pire expérience avec les trois back-tests pour les EIIEs. Cependant, ils sont en mesure de monter régulièrement jusqu'à la fin du test.*

Parmi les trois réseaux EIIE, LSTM a obtenu des scores bien inférieurs à ceux du CNN et du RNN de base. L'écart important de performances entre les deux espèces de RNN dans le même cadre pourrait être un indicateur

du secret bien connu des marchés financiers, que l'histoire se répète. N'étant pas concu pour oublier son historique d'entrée, un RNN vanille est plus capable qu'un LSTM d'exploiter des schémas répétitifs de mouvement des prix pour des rendements plus élevés. L'écart pourrait également être dû à un manque de réglage fin des hyper-paramètres pour le LSTM. Dans les expériences, le même ensemble d'hyper-paramètres structurels a été utilisé pour le RNN de base et le LSTM.

Malgré le succès du cadre EIIE dans les back-tests, il y a place a l'amélioration dans les travaux futurs. La principale faiblesse des travaux actuels réside dans les hypothèses d'impact nul sur le marché et de glissement nul. Afin de tenir compte de l'impact et du glissement du marché, une grande quantité d'exemples de trading réels bien documentés seront nécessaires comme données de formation. Un protocole devra être inventé pour documenter les actions commerciales et les réactions du marché. Si cela est accompli, des expériences de trading en direct de l'agent de trading automatique dans sa version actuelle peuvent être enregistrées, pour que sa future version apprenne les principes derrière les impacts du marché et les dérapages de cet historique enregistré. Une autre lacune du travail est que le cadre n'a été testé que sur un seul marché. Pour tester son adaptabilité, les versions actuelles et ultérieures devront être examinées lors de back-tests et de transactions en direct sur un marché financier plus traditionnel. De plus, la fonction actuelle de récompense devra être modifiée, voire abandonnée, pour que l'agent d'apprentissage par renforcement inclue la prise de conscience des réactions du marché à plus long terme. Cela peut être réalisé par un réseau critique.

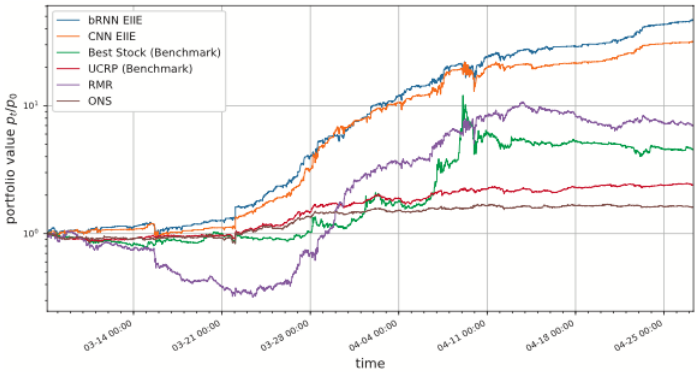


FIGURE 7 – *Back-Test 3 : 2017-03-07-4 :00 à 2017-04-27-8 :00 (UTC), échelle logarithmique de l'humidité accumulée. Tous les algorithmes luttent et se consolident au début de cette expérience, et les deux EIIL connaissent deux baisses majeures le 15 mars et le 9 avril. Cette plongée contribue à leur baisse maximale élevée dans le texte (tableau 2). Néanmoins, c'est le meilleur mois pour les deux EIIL en termes de patrimoine final*

Pour la tipologie de réseau, d'un point de vue pratique, EIIE présente trois autres avantages cruciaux par rapport à un réseau intégré. Le premier est l'évolutivité du nombre d'actifs. Ayant les mini-machines toutes identiques avec des paramètres partagés, le temps de formation d'un ensemble évolue à peu près linéairement avec M. Le deuxième avantage est l'efficacité de l'utilisation des données. Pour un intervalle d'historique de prix, une mini-machine peut être entraînée m fois sur différents actifs. L'expérience d'évaluation des actifs des IIE est ensuite partagée et accumulée dans les dimensions temps et actifs. Le dernier avantage est la plasticité de la collecte d'actifs

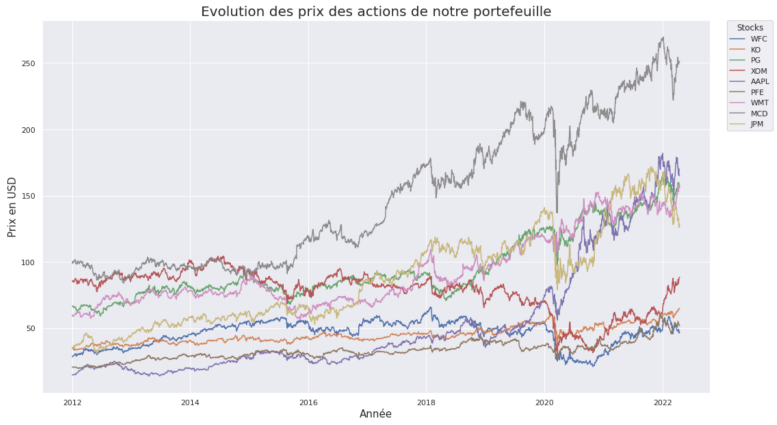
Une facon d'y parvenir est de s'appuyer sur la capacité de mémorisation de RNN, mais avec cette approche, le schéma de normalisation des prix proposé (price tensor) doit être abandonné. Ce schéma de normalisation est empiriquement plus performant que les autres. Une autre solution possible est le renforcement direct (RR) introduit par Moody et Saffell (2001). Cependant, les mémoires RR et RNN souffrent du problème de disparition du gradient. Plus important encore, RR et RNN nécessitent une sérialisation du processus de formation, incapables d'utiliser une formation parallèle dans des mini-lots.

8 Implémentation

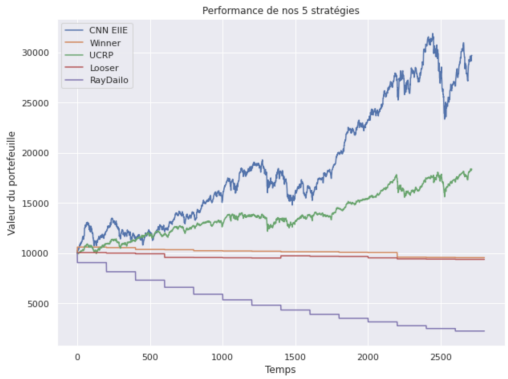
Après la lecture de cet article, nous étions curieux de connaître si nous pouvions implémenter ce travail de recherche. Pour ce faire nous avons décidé de constituer une base de données à partir des données journalières de neufs actions américaines disponibles sur Yahoo Finance. Les actions choisies proviennent toutes d'un secteur économique différent, le but est de minimiser la corrélation entre les actifs. Alors que dans l'article de recherches

les auteurs ont démontré leurs recherches sur des cryptomonnaies, nous souhaitons utiliser un actif plus classique pour deux raisons. La première est le fait que des données sur les cryptomonnaies sont rares et la deuxième raison s'explique par la forte corrélation des actifs crypto entre-eux. Des phénomènes extérieurs peuvent venir perturber les cours des actifs digitaux, notamment comme le milliardaire Elon Musk. Ce dernier affole à chaque tweet les cours du Bitcoin, du Dogecoin et autre. Et étant donné la forte corrélation entre les cryptomonnaies, nous pensons que notre implémentation pourrait être biaisée par des forts mouvements communs à toutes les cryptomonnaies. De plus, nous avons décidé d'implémenter uniquement le réseaux de neurones CNN EIIE, dans l'optique de comparer les résultats obtenus aux autres méthodes de gestion de portefeuilles évoquées dans l'introduction.

Notre base de données débute en le 03 janvier 2012 et se termine le 18 avril 2022.



Nous constatons que notre portefeuille est relativement bien diversifié, nous pouvons maintenant partir à la modélisation de notre algorithme. Dans un premier temps, nous entraînons notre réseaux de neurones et nous l'améliorerions grâce aux fonctions de politique. Par la suite, nous souhaitons mettre en confrontation notre CNN EIIE par rapport à d'autres méthodes de gestion de portefeuille sur notre base de test. L'une des méthodes qui est en concurrence avec notre réseaux de neurones est l'UCRP (Uniform Constant Rebalanced Portfolio). Nous obtenons les performances suivantes :



Nous constatons que notre algorithme et la stratégie UCRP ont de très belles performances et que les autres sont bien plus médiocres. Lorsque nous nous penchons sur les deux performances positives, nous pouvons expliquer l'excellente performance de notre réseaux de neurones par la composition de notre portefeuille.

Nous obtenons de très bons résultats de notre implémentation, ce qui prouve encore une fois que les méthodes de Machine Learning ont tout à fait leur place en finance quantitative. Toutefois, nous notons que notre application du papier de recherche pourrait être améliorée et complétée par d'autres réseaux de neurones mais aussi par une expérience de gestion de portefeuille dynamique comme les auteurs ont pu le réaliser sur l'échange de cryptomonnaies Poliniex.com.