Master Thesis

Deep reinforcement learning appliqué à la gestion d'actif-passif d'un portefeuille de retraite collective

EYROLLE NICOLAS

20 janvier 2025

Reviewer :
P.BRUGIERE
C.Vincent

Table des matières

1 Introduction

1.1 Motivations et contexte

Les compagnies d'assurance ont pour objectif d'assurer une protection financière en cas de survenance d'un sinistre. Pour bénéficier de cette protection les assurés versent au préalable des primes (i.e cotisations). Le conditionnement du versement de prestations futures par le paiement de cotisations en amont, est au cœur de l'activité de l'assureur. Cette inversion du cycle de production, requiert de la part de ce dernier d'investir le montant des primes perçues sur les marchés pour i) mettre en réserve et honorer ses engagements contractuels, ii) dégager une marge et ainsi assurer son développement. Pour parvenir à ses objectifs la compagnie est dans l'obligation de définir une stratégie d'investissement qui prend en compte les risques inhérents à son portefeuille, que ce soit sur les actifs financiers, les engagements présents au passif ainsi que ceux qui sont issus de l'interaction entre les deux. Par ailleurs, la stratégie doit respecter le cadre prudentiel en vigueur et ainsi répondre aux contraintes de solvabilité édictées par la norme européenne Solvabilité II qui impose aux entreprises de faire face à un risque de banqueroute avec probabilité 99%. Dans ce contexte le département de gestion Actif-Passif (ou ALM pour Asset Liabilities Management) réalise des études approfondies, afin de déterminer la stratégie d'investissement qui permet d'atteindre le meilleur couple rendement-risque. Ceci est d'autant plus vrai pour les contrats de retraite dont la longue durée accroit l'exposition aux différents risques portant sur l'actif et le passif.
Pour parvenir à un pilotage adéquat de sa gestion ALM, AXA dispose d'un modèle interne qui permet de projeter sur 60 ans les flux de trésorerie de la compagnie d'assurance. Ces projections économiques, dont chaque trajectoire est générée par modèles de scenarios économiques, permettent d'étudier la robustesse des différentes stratégies d'allocations d'actifs (ou SAA) ainsi que leur exposition aux risques. Cependant, l'analyse financière dont résultent les SAA ainsi que leur évaluation consomme beaucoup de temps ce qui rend difficile la production d'indicateurs de solvabilité étant donnée la cadence imposée par Solvabilité II. Néanmoins les récents développements de l'IA, en particulier ceux réalisés dans l'apprentissage par renforcement, laissent envisager de nouvelles perspectives afin d'améliorer les performances et ainsi mieux satisfaire les contraintes de solvabilité.
L'apprentissage par renforcement est le fruit de la combinaison de champs disciplinaires tels que la programmation dynamique, le controle optimal, la théorie de la mesure et les statistiques orientés pour résolution de problèmes d'optimisation. Les récents progrès réalisés en deep learning ont donné naissance à de nouveau algorithmes de deep reinforcement learning capables de surperformer l'intelligence humaine dans des domaines d'une compléxité inédite. En 2014, Deepmind implémente un algorithme nommé AlphaGo en mesure de jouer au jeu de go. En 2015 ce dernier parvient à battre [10],[11],[3],[9] le champion du monde en titre. Bien que les champs de recherche en soient à leurs balbutiemments, la versatilité de la modélisation permise par le reinforcement learning laisse présager de nombreuses applications sur des sujets variés comme le séquencage de l'adn [4] ou le développement des voitures autonomes voire de l'asset management [1]. La gestion actif-passif ne fait pas exception d'après [2] et semble modéliser de

manière adéquat les enjeux de gestion ALM. En effet, l'apprentissage par renforcement met en scène un agent (ici la compagnie d'assurance), dont l'objectif est de prendre des décisions optimales (les Strategic Asset Allocation), dans un environnement donné (scénarios économiques). C'est pourquoi dans ce mémoire nous nous intéressons à l'implémentation d'un modèle de deep reinforcement learning pour trouver les SAA garantissant le meilleur équilibre entre rentabilité et limitation des risques pour la compagnie. Cette étude porte sur un portefeuille fermé de retraites collectives d'AXA France.

1.2 Structure de l'étude

Afin de mener cette étude nous procéderons pas étapes :

1. On commencera par présenter les enjeux liés à la gestion actif-passif pour un portefeuille de retraite. Cette partie sera l'occasion de présenter le contexte, les caractéristiques et les risques auxquels sont soumis les contrats de retraite. Enfin, on détaillera le procédé adopté par AXA pour mener une étude ALM actif-passif, ainsi que l'impact de la norme en vigueur sur cette dernière.
2. On présentera le modèle ALM qui servira pour notre étude. En particulier, on justifiera le choix de notre protefeuille d'étude et présentera le déroulement de notre modèle ALM.
3. On présentera le cadre du deep reinforcement learning avant de l'intégrer à notre modèle ALM pour analyser les performances et essayer de les interpréter.

.

2 Généralités sur la gestion actif-passif d' un portefeuille de retraite

2.1 Generalités sur l'assurance-vie et les contrats de retraite

2.1.1 Principes de l'assurance vie

L'assurance vie se caractérise par l'engagement de l'assureur à verser une prestation lorsque survient un événement lié à la durée de vie de l'assuré. Il est possible de distinguer deux événements pouvant déclencher la prestation.

- L'assurance en cas de décès : la compagnie s'engage à verser un capital ou une rente à un ou plusieurs bénéficiaires lorsque le décès de l'assuré survient avant la fin du contrat.
- L'assurance en cas de vie : à la différence de la précédente, un capital ou une rente est versé à l'assuré si ce dernier est toujours en vie.

Les contrats d'assurance vie peuvent être assimilés à des contrats d'investissement dont l'échéance est aléatoire. Dans le cas des contrats de retraite, l'assureur s'engage à verser une rente (viagère, différée ou temporaire) au bénéficiaire jusqu'au jour de son décès. Concrètement, le cycle de vie de ces contrats est constitué de deux périodes : une phase de constitution, pendant laquelle les primes des assurés sont capitalisées, puis une phase de restitution, à la date de la retraite de l'assuré, au cours de laquelle le capital accumulé jusqu'à cette date est reversé à l'assuré progressivement sous forme d'une rente dont le montant est fixé en début de restitution.

Considérés comme le "placement financier préféré des français", les contrats d'assurance-vie constituent un moteur essentiel à l'économie française. Ils suscitent l'intérêt de nombreux souscripteurs dont les objectifs sont multiples :

- investir de façon peu risquée ;
- transmettre un patrimoine financier à des tiers du fait des avantages fiscaux en matière de succession ;
- défiscaliser les revenus du capital.

Bien que la baisse des taux d'intérêts observée depuis plusieurs années a impacté l'assurance vie, cette dernière continue à occuper une place importante dans l'économie française.

Au troisième trimestre 2020, les ménages détenaient environ 2 103 milliards d'euros en assurance-vie (en euros ou en unités de compte), soit un peu plus de 38% du total de leur encours d'épargne financière.

2.1.2 Le Système des retraites

Le système des retraites peut être schématisé par une fusée à trois étages :

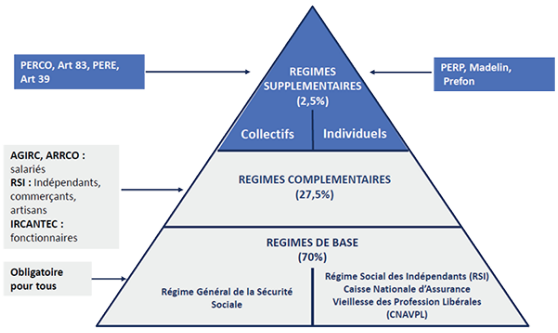

FIGURE 2.1 – Fusée à trois étages representative du système de retraite actuel (données 2020)

1er étage : Les régimes de base

Le régime de base constitue le premier niveau du système de retraite français, 70% de français sont concernés par ce régime. Géré par la sécurité sociale et obligatoire, il est fondé sur le principe de solidarité intergénérationnelle : Les actifs cotisent pour les pensions des retraités actuels. En fonction de leur catégorie socio-professionnelle et de leur statut, les travailleurs sont affectés à un régime de base. Il en existe plus de 42 bien que 3 d'entre eux regroupent près de 80% des pensions de retraites :

- Le régime général qui regroupe les fonds de pensions des salariés du privé. Il s'agit du régime le plus important.
- Le régime d'Etat qui prend en charge les pensions des agents d'Etat.
- La Mutualité Sociale Agricole (MSA qui verse les pensions des salariés et non-salariés agricoles).

Les caisses de régimes spéciaux, gèrent les retraites des salariés et indépendants appartenant à une catégorie particulière tels que les salariées de la SNCF, de la RATP, des avocats etc. Certaines caisses prennent en compte à la fois les régimes de bases et complémentaires.
Chaque régime a ses propres règles de calcul prises en charge par leurs caisses de retraites respectives en fonction du nombre d'années effectuées et de la rémunération obtenue.

2ème étage : Le régime complémentaire

Les régimes complémentaires de retraite ont été créés dès 1947 en raison de l'insuffisance des pensions servies par le régime général. Comme pour les régimes de base, l'affiliation et le versement de cotisations sont obligatoires depuis 1972. Pour les salariés et cadres relevant pour leur retraite de base de la Caisse nationale d'assurance vieillesse (CNAV) ou de la Mutualité sociale agricole (MSA), la retraite complémentaire est gérée par deux entités :

- Pour les cadres, il s'agit de l'Association Générale des Institutions de Retraite Complémentaire des cadres (AGIRC), créée en mars 1947 ;
- Pour l'ensemble des salariés, c'est l'Association des Régimes de Retraite Complémentaire (ARRCO) créée en décembre 1961. Ces deux associations ont été réunies, le 1er juillet 2002, au sein d'un groupement d'intérêt économique : le GIE Agirc-Arrco. Conformément à un accord signé par les partenaires sociaux en octobre 2015, ces deux organismes ont fusionné au 1er janvier 2019 pour constituer un régime unifié qui reprend l'ensemble des droits et obligations de l'Agirc et de l'Arrco

à l'égard de leurs ressortissants.

En 2020, l'Agirc-Arrco comptait 15,2 millions de pensionnés. Les régimes de retraite complémentaire sont gérés et pilotés exclusivement par les partenaires sociaux (représentants des salariés et des employeurs), représentés à égalité dans chacune de leurs instances, au sein de la nouvelle entité Agirc-Arrco. Leur particularité est qu'ils sont gérés en points de retraite et négociés avec les partenaires sociaux des conventions collectives de chaque secteur. Il existe par ailleurs d'autres régimes complémentaires en fonction de la branche d'activité du cotisant présent dans la figure (joindre)

3ème etage : Régime de retraites supplémentaire ou surcomplémentaire

- 3ème étage : Régime de retraites supplémentaire ou surcomplémentaire : Facultatif, le régime supplémentaire ne dépend d'aucun syndicat ni institution étatique, on y souscrit via le recours à des assurances, des instituts de prévoyance ou d'établissements bancaires. Ce troisième niveau, est un régime par capitalisation. En plus de permettre aux travailleurs de se constituer une épargne supplémentaire afin de compléter la pension issue des régimes de base et complémentaires ces contrats bénéficient d'avantages fiscaux.

Comme tout contrat d'assurance vie, leur durée de vie est marquée par une phase de constitution durant laquelle l'assuré cotise afin de se constituer une rente en attendant de pouvoir en bénéficier en phase de restitution lors de l'âge légal de départ en retraite. Lors de la phase de constitution, les primes des assurés sont placées dans des fonds dont les performances et l'exposition au risque dépendent de l'appétence au risque des assurés. Il existe trois fonds majeurs :

1. Les **fonds euros** : Gérés par les compagnies d'assurance, il s'agit de supports financiers constitués majoritairement d'obligations souveraines. Pour les assurés, les fonds euros sont des placements sécurisés qui présentent des garanties de rendement minimal sous forme d'un taux minimum garanti (TMG) ou taux minimum garanti annuel (TMGA).
2. Les **fonds en unité de compte (UC)** sont composés de plusieurs supports financiers tels que les actions, obligations et parts d'organismes de placements collectifs en valeurs mobilières ou immobilière, les fonds indiciels (ETFS)... Les UC constituent pour l'assuré un placement à risque avec un profit espéré supérieur à celui des fonds euros. Ainsi, contrairement à ces derniers, les UC ne présentent aucune garantie de la part de l'assureur d'où l'existence d'options de gestion permettant à l'assuré de composer un panier d'investissement en fonction de son aversion au risque.
3. Les **fonds euro-croissance** : Apparus plus récemment dans la réglementation française, à partir de 2011, ces supports ont vocation à garantir un pourcentage du capital investi par l'assuré avec un horizon de temps donné, généralement 10 ans ou plus. Ainsi, l'assureur qui porte cet engagement peut élargir son allocation d'actif à des placements un peu plus risqués que sur le fonds euros.

Les contrats récents sont le plus souvent des contrats dits multi-support, ce qui permet aux assurés d'arbitrer entre les fonds euros et UC en fonction de l'évolution de leur cible d'investissement. Par exemple, il est fréquent dans les contrats de sécuriser le capital vers le fonds en euros à l'approche de la retraite. En vue de l'essouflement du système des retraites par répartition marqué par un déséquilibre croissant du ratio $\frac{cotisants}{retraitès}$ (cf.Figure 2.1.2), le recours à ces contrats pour les individus en capacité de se constituer une épargne ne devraient cesser de croître. Dans ce contexte, ils représentent une potentielle manne financière importante : les assurances ne cesseront d'adapter leur offre pour entrer en adéquation avec les caractéristiques des souscripteurs.

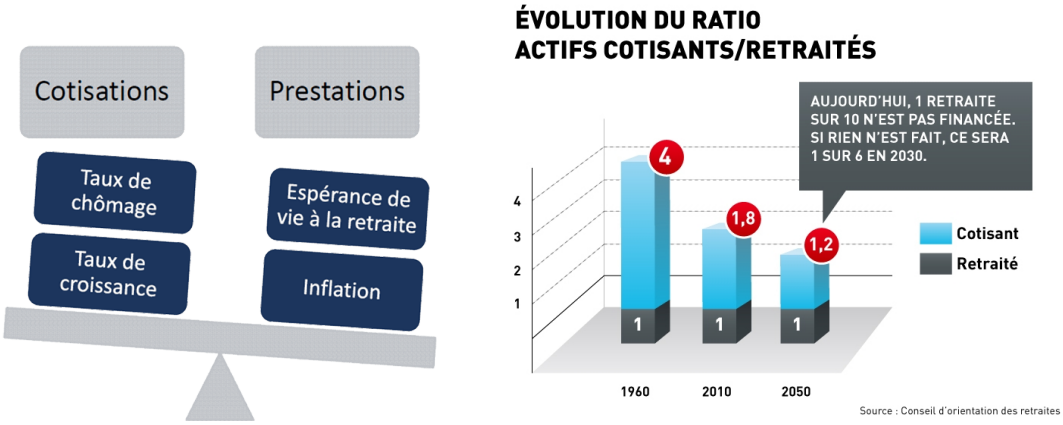

FIGURE 2.2 – Essouflement du système des retraites par répartition

2.1.3 Caractéristiques et Typologie des contrats de retraites collectives supplémentaires

Les contrats de retraites supplémentaires peuvent être souscrits à titre individuel ou collectif. La souscription et la mise en place des seconds sont à l'initiative des entreprises dont les options et garanties financière varient en fonction des contrats. Etant donné que notre étude concerne le portefeuille de retraites collectives d'AXA France, nous nous limiterons au détail de ces contrats. Les contrats de retraite collective peuvent différer par le contenu de leurs options et garanties financières :

- - Les Taux Minimum Garantis (TMG) : Il s'agit d'un taux défini contractuellement qui garantit un rendement minimum sur le capital investi pendant la phase de constitution. Il existe également des TMGA (Taux minimum Garantis Annuels), i.e. des taux qui ne sont valides qu'un an, et redéfinis chaque année, généralement en fonction des taux servis sur les années précédentes, ou en fonction du taux moyen d'emprunt d'Etat. Ces taux sont à définir de manière stratégique en fonction de la concurrence, des taux courants du marché et des performances de l'entreprise. En effet, si le taux est très bas, il risque de pousser les assurés à souscrire auprès d'autres organismes proposant de meilleurs taux ou d'autres contrat d'épargne plus avantageux. tandis que si le taux proposé est trop haut l'assureur sera dans l'obligation d'assurer la compensation financière qui résulte du différentiel entre la performance tirée de on activité et les taux servis. La réglementation a encadré progressivement de manière stricte le niveau de ces taux afin d'éviter que certains assureurs prennent des risques long terme jugés trop importants.
- Les types de sorties du contrat : selon les types de produits (définis par la réglementation), l'assuré peut recevoir de manière lissée dans le temps le capital accumulé lors de la phase de constitution via une rente viagère ou bien choisir de recevoir son capital constitué sous la forme d'une unique indemnité. Il est possible de prévoir au moment de la mise en place de la rente une réversion d'une quote-part de la rente à un bénéficiaire.
- Les contrats de droits individualisés ou non individualisés
- Les options de rachat : Le rachat est l'opération à travers laquelle l'assuré récupère la totalité ou une partie de son épargne. Plusieurs évènements peuvent provoquer une sortie de capital en phase de constitution tels que le décès, l'achat d'un bien immobilier par l'assuré, ou un transfert du contrat

Type de contrat	Descriptif
IFC (Indemnité fin de carrière) (Euros/UC)	TMG/TMGA, pas de phase de rentes
Article 83 (droits individualisés) • Rente Viagère Différée (RVD) (uniquement Euros) • Capitalisation Viagère (uniquement Euros) • Capitalisation Financière (Transformation immédiate) (Euros/UC) • Capitalisation Financière (RPD : Régime à prestations définies), (Transformation au terme) (Euros/UC)	• Garantie du montant de la rente future, aucun engagement en cas de décès • Garantie du montant de la rente future, PM versées en cas de décès • Garantie du taux/table de conversion, PM versées en cas de décès, TMG/TMGA • Pas de garantie de taux/table, PM versées en cas de décès, TMG/TMGA
Article 39 (droits non-individualisés) (Euros/UC)	TMG/TMGA, pas de garantie de taux/table
Rentes en service	Versement d'arrérages avec participation.
Préfon	Valeur de service d'un point ne peut pas diminuer Partification aux Bénéfices minimale à doter à la Provision Technique Spéciale

FIGURE 2.3 – Typologie des contrats de retraites collectives

vers une autre compagnie d'assurance etc.

- Le type de fonds d'investissement : Euros, UC
- Le mode de cotisations : cotisations/prestations définies...
- : Le taux de participation aux bénéfices (PB) qui est la part du résultat financier et technique réalisé par l'assureur et dû aux assurés. Le Code des Assurances impose aux assureurs de reverser au minimum 90% des résultats techniques et 85% des résultats financiers aux assurés. Le taux de participation aux bénéfices reversé aux assurés est également un élément important permettant aux assureurs de se différencier de leurs concurrents.

L'ensemble de ces critères nous permettent de définir une typologie de contrats présente dans la figure 2.3

2.2 Solvabilité II

Pour piloter la stratégie de la compagnie, la gestion Actif-Passif représente une activité indispensable à l'évaluation des risques tant sur le plan qualitatif que quantitatif. Elle permet de satisfaire les exigences de la norme SII qui a pour objectif principal de garantir la stabilité de l'activité d'assurance ainsi qu'une protection tant pour les investisseurs que les assurés. Ce cadre prudentiel se décline en 3 piliers dont nous présentons les grands principes.

2.2.1 Pilier 1 : Quantitatif

Ce premier pilier définit de nouvelle mesures de solvabilité et méthodes de valorisation de l'actif et du passif. L'introduction de ces dernières au bilan permettent d'adopter une vision économique et prudentielle de l'activité de la compagnie.

Mesures de risques S2 définit deux mesures pour évaluer la solvabilité de la compagnie :

- Le MCR (Minimum Capital Requirement) qui représente le montant minimum de fonds propres nécessaire à la compagnie pour avoir le droit d'exercer.
- Le SCR (Solvency Capital Requirement) qui est le montant nécessaire de fonds propres pour faire face à une ruine économique à horizon un an avec une probabilité de 99%

Ces indicateurs peuvent être estimés par deux méthodes. La première consiste en l'application d'une formule standard, présente dans les réglements délégués de S2, pour tous les organismes d'assurance. La seconde, appelée modèle interne, implique l'utilisation de scénarios de chocs propres à chaque entité permettant de prendre en compte les spécificités du portefeuille d'assurance. Ces chocs sont calibrés par l'entreprise et son modèle interne doit être validé par le régulateur. Dans ce contexte le SCR est la $VAR_{99\%}$ des différences entre les marges futures (ou VIF : value in force) réalisées dans le contexte d'un scenario économique non choqué avec celles de scénarios où un module de risque est choqué. Le SCR s'interprete comme les réserves en capital dont doit disposer la compagnie d'assurance pour faire face au niveau de risque considéré. On a :

$$SCR_{risque\ X} = VIF_{BC} - VIF_{choquee}$$

En pratique le SCR est appliqué à chaque module de risque présent dans la figure 2.4 puis agrégé via une matrice de corrélation dont les coefficients sont déterminés par .. On a :

$$SCR_{global} = \sqrt{\sum_{i=0}^{n} SCR_{rique_i} \cdot SCR_{rique_j} \cdot \rho_{i,j}}$$

où $\rho_{i,j} = Corr(i,j)$

Vision *Best estimate* En plus d'intégrer les marges de solvabilité, le bilan prudentiel adopte une nouvelle méthode d'évaluation, dite *Best estimate*, de ses actifs et passifs.
Les actifs sont dorénavant comptabilisés en valeur de marché coté actif et en *Fair Value* pour le passif. La Fair value est le montant estimé pour lequel un assureur serait prêt à racheter le portefeuille de contrats. L'évaluation *Best estimate* du passif consiste en l'actualisation des flux de trésorerie futurs des engagements de la compagnie avec la courbe de taux sans risque.

$$BEL = \mathbb{E}^{\mathbb{P}} \otimes^{\mathbb{Q}} \left[\sum_{i \geq 1} CF_i \cdot D_i \right] \tag{2.1}$$

avec CF_i et D_i les cashflows et le facteur d'actualisation à la date $t = i$ et $\mathbb{P}$ et $\mathbb{Q}$ les lois de durée de vie des assurés et la probabilité risque-neutre. Ce passage d'une vision comptable à une vision économique prudentielle est schématisé par la figure 2.5.

FIGURE 2.4 – Décomposition du SCR par module de risque

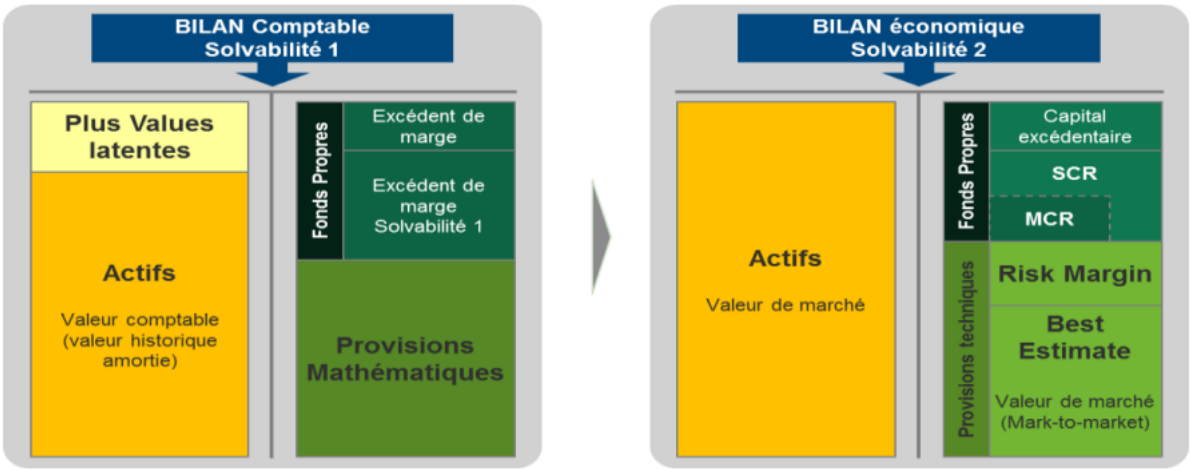

FIGURE 2.5 – Vision comptable vs vision prudentielle

2.2.2 Pilier 2 : Qualitatif

Les mesures de risques et les méthodes de calcul édictées dans ce pilier ont l'avantage de définir un cadre prudentiel propice à la comparaison entre plusieurs entités. Néanmoins, elles requièrent d'être adaptées pour correspondre au mieux à l'activité de l'entreprise. Ce pilier de la Directive vise à décrire les exigences en matière de gouvernance et de gestion des risques au sein de l'entreprise. Cette dernière doit en particulier identifier, cartographier, suivre les risques inhérents et définir une politique interne concrète en vue de les gérer compte tenu de le niveau de risque que s'est fixé la société.

2.2.3 Pilier 3 : Communication

SII impose aux entreprises une cadence de publication de leurs rapports financiers dans lesquels figurent leur performance ainsi que leur niveau de solvabilité. Cette cadence soutenue peut s'avérer contraignante lorsque le processus d'évaluation prend du temps. C'est une des raisons qui nous pousse à changer de paradigme.

2.3 Principes généraux de la Gestion Actif-Passif

Le cadre prudentiel instauré par S2 implique de conditionner chaque prise de décision par une étude des risques. Dans notre étude, les décisions sont matérialisées par une stratégie d'allocation suite au placement des primes des contrats de retraites sur les marché financiers. Afin que la compagnie soit toujours en mesure de remplir ses engagements, les allocataires doivent conditionner la stratégie d'investissement par une étude appronfondie du profil de risque du passif mais aussi des actifs financiers sur lesquels ils investissent. L'enjeu est de trouver une adéquation entre la performance et l'exposition aux risques des produits financiers compte tenu des obligations contractuelles et des risques auxquels sont soumis les engagements. Pour y parvenir, les sociétés ont recourt à une étude ALM.

2.3.1 Risques liés à la gestion actif-passif

L'identification des risques ALM est une étape importante de l'étude. Elle consiste en l'analyse des caractéristiques des engagements de l'assureur dans le but d'investir sur les actifs qui génèrent des flux financiers en mesure de satisfaire les besoins de trésorerie.

Risques à l'actif

En assurance vie, les risques à l'actif sont essentiellement des riques financiers qui résultent de la structure du bilan et de l'évolution des marchés financiers. Il s'agit en particulier des risques de taux d'interêt, du risque de contrepartie et de liquidité.

1. Le risque de taux Le risque de taux d'intérêt est engendré par les supports financiers à taux variables détenus en portfeuilles ainsi que des caracteristiques en termes de garantie de taux des contrats au passif. Le risque n'est pas le même en fonction du mouvement des taux.

Le risque de taux à la hausse : En cas d'augmentation des taux, les assurés peuvent demander le transfert de leur contrat vers une compagnie d'assurance concurrente s'ils estiment que le taux servi n'est pas suffisant afin de réaliser un arbitrage. De plus, si la société est contrainte de vendre des obligations

pour honorer ses engagements, elle risque de réaliser des moins-values latentes sur les obligations à taux fixes en cas de hausse de taux d'intérêt.

Risque de taux à la baisse : En cas de baisse des taux, le réinvestissement des coupons rapporte moins ce qui affecte le rendement. De plus, les obligations achetées pour remplacer celles arrivées à terme ont des taux moins intéressants. En cas de baisse sévère, l'assureur n'est plus en mesure de garantir le TMG et puise dans ses fonds propres.

2. Risque de liquidité Le risque de liquidité intervient lorsque les maturités des créances et des dettes ne coïncident pas. Cette différence conduit à une inadéquation entre la liquidité des actifs et du capital exigé pour honorer les engagements. Par conséquent, l'assureur doit investir sur des actifs ayant une liquidité suffisante pour faire face à tous les flux passifs sortants.

2.3.2 Les risques au passif

Les risques du passifs résident dans les options et garanties financières des contrats qui composent le portefeuille ainsi que dans le comportement des assurés. Les principaux risques proviennent des sorties de capital inopinées (rachat, décès).

Sorties en capital

On distingue majoritairement deux raisons provocant une sortie de contrat précoce :

1.Le risque de rachat L'encours de l'assuré peut être racheté à n'importe quelle date selon les règles définies lors de la souscription du contrat. Nous distinguons deux types de rachats en assurance-vie :

- Les rachats structurels : Ce type de rachats dépend des caractéristiques du contrat d'assurance. En retraite collective, les rachats dépendent de la nature du contrat souscrit, et peuvent également dépendre des conditions techniques du contrat (TMG, etc...). La décision de rachat peut être prise directement par l'entreprise souscriptrice, ou bien par une tête assurée si la nature du contrat le permet et s'il respecte les contraintes réglementaires pour un rachat (par exemple achat de sa résidence principale, naissance d'un 3ème enfant,...)
- Les rachats conjoncturels (ou dynamiques) sont les rachats déclenchés par la conjoncture économique ainsi que par la performance de l'assureur. Ils sont généralement estimés à partir du spread entre le taux servi par l'assureur et le taux concurrent. Ces derniers sont plus difficiles à modéliser car il n'existe pas d'historique observable dans la plupart des compagnies.

2.Le risque de mortalité ou longévité Le risque de longévité en retraite collective correspond au risque de constater un nombre de décès significativement moins élevé que ce qui est prévu par les tables de mortalité réglementaires utilisées pour tarifer les contrats de retraite. Il peut se produire par exemple suite à une pandémie, guerre ou catastrophe naturelle, ou bien être observé comme une dérive progressive liée à l'amélioration des conditions de vie.

La gestion de ces sorties en capitaux précoces nécessite de la part de l'assureur d'investir dans des actifs majoritairement liquides.

Exemple risque actif-passif : adossement en duration Un élément majeur de la gestion actif-passif réside dans l'adéquation de la sensibilité aux taux entre l'actif et le passif. Cette sensibilité est mesurée par la

duration qui est une approximation d'ordre 1 de la sensibilité de flux financiers au taux sans risque utilisé pour leur actualisation.

La duration La duration mesure l'élasticité des flux actualisés par rapport à une variation des taux. Mathématiquement elle s'exprime comme la dérivé de la valeur actuelle probable (VAP) de flux financiers en fonction du taux d'actualisation. On a donc :

$$Duration = \frac{\frac{\partial VAP}{\partial i}}{VAP} = \frac{1}{VAP} \cdot \frac{\partial \sum_{t=0}^{n} \frac{Flux\_probable_t}{(1+i)^t}}{\partial i} \tag{2.2}$$

Si le taux baisse,toutes choses égales par ailleurs, la VAP augmente. Elle diminue si le taux augmente.

Le gap de duration Le gap de duration s'exprime de la manière suivante :

$$\Delta := Gap\_de\_duration = duration\_passif \; - \; duration\_actif \tag{2.3}$$

Il permet de mesurer la différence de sensibilité aux taux entre l'actif et le passif. Il existe trois cas de figure :

1. $\Delta > 0$: cela sigifie que le passif est davantage sensible à une variation des taux que l'actif. Ainsi la compagnie est soumise au risque de taux en cas de **baisse** de celui-ci. En effet, une **baisse** de taux implique une augmentation de la VAP du passif supérieure à celle de la VAP de l'actif.
2. $\Delta < 0$: cela sigifie que l'actif est davantage sensible à une variation des taux que le passif. Ainsi la compagnie est soumise au risque de taux en cas de **hausse** de celui-ci. En effet, une **hausse** de taux implique une diminution de la VAP de l'actif supérieure à celle de la VAP de l'actif.
3. $\Delta = 0$: cela sigifie que le passif et l'actif ont la même elasticité face à une variation des taux. Toute variation des taux au passif est compensée par une variation de même ordre à l'actif. [1]

Contrairement au dernier cas où les mouvements de l'actif et du passif se compensent en cas de variation des taux, lorsque le gap de duration est non nul, l'assureur doit compenser la différence observée entre la VAP de l'actif et du passif. En effet, nous rappelons que l'assureur est tenu réglementairement de garantir ses engagements à tout moment. Pour y parvenir, il doit s'assurer de disposer d'un montant d'actif équivalent à celui de son passif. En cas de défaut, il est contraint de puiser dans ses réserves.

On déduit du raisonnement précédent qu'il est primordial pour l'assureur de combler le gap de taux afin de se couvrir de la variation des taux.

2.3.3 L'allocation d'actifs gestion ALM

Après avoir identifié le profil de risque de l'actif et du passif, il est nécessaire d'évaluer leur impact en fonction des décisions d'allocations prises. Pour quantifier leur influence sur le bilan de la compagnie, AXA dispose d'un modèle interne qui permet de projeter les flux de trésorerie sur le bilan de l'entreprise.

1. Nous présentons ici un raisonnement simplifié. En réalité, le calcul de la duration n'est qu'une approximation d'ordre 1 de la sensibilité de la VAP par rapport à un mouvement de taux. En pratique la direction des investissements étudie la convexité des flux par rapport au taux. De plus, en général l'actif est intrinséquement moins sensible au taux que le passif du fait que la valorisation de certains actifs dépend peu voire pas du niveau de taux.

Fort de cet outil,la direction des investissements est en mesure de tester différentes allocations afin de trouver celle qui garantit les meilleurs performances pour un niveau de risque donné. Dans cette partie nous déroulons le protocole suivi par les équipes d'AXA afin de trouver l'allocation stratégique optimale.

2.3.4 Choix de l'allocation d'actif stratégique

La détermination de l'allocation via une étude ALM est le fruit d'un processus en 3 grandes étapes :

1. Définition d'une plage d'allocations tests
2. Evaluation des allocations
3. Choix de l'allocation stratégique

Détermination des plages tests d'allocations

Compte tenu des caractéristiques de l'allocation courante, du contexte économique, des objectifs de performances souhaités et du profil de risques des contrats, la direction des investissements définit une plage d'allocations test. A partir de cette plage, différentes allocations vont être testées afin de trouver l'allocation cible souhaitée.

Evaluation des allocations test

Pour parvenir à determiner la SAA AXA dispose d'un modèle interne équipé de scenarios économiques générés (ESG). Ces scénarios stochastiques ont pour objectif de simuler différents environnements économiques qui permettent de projeter les flux de trésorerie à l'actif et au passif. A l'issue de ce procédé chacune des allocations testées est évaluée selon une metrique de performance et une metrique de risque :

$$Metrique\_perf = VAN(rendements futurs) \tag{2.4}$$

$$Metrique\_risque = SCR \tag{2.5}$$

Choix de l'allocation Optimale

Le choix de l'allocation optimale se fait d'après le paradigme de Markovitz. Il consiste à trouver un portefeuille efficient, celui qui optimise le couple rendement-risque. Pour trouver la décision optimale, chaque allocation est projetée dans un plan à deux dimensions dont les axes caractérisent la métrique de risque et de performances choisie. Une fois projetées la direction des investissements identifie une courbe de portefeuilles efficients : en fonction des candidats l'équipe ALM analyse la faisabilité et la pertinence des allocations correspondantes pour séléctionner le portefeuille optimal. (cf.figure 2.6)

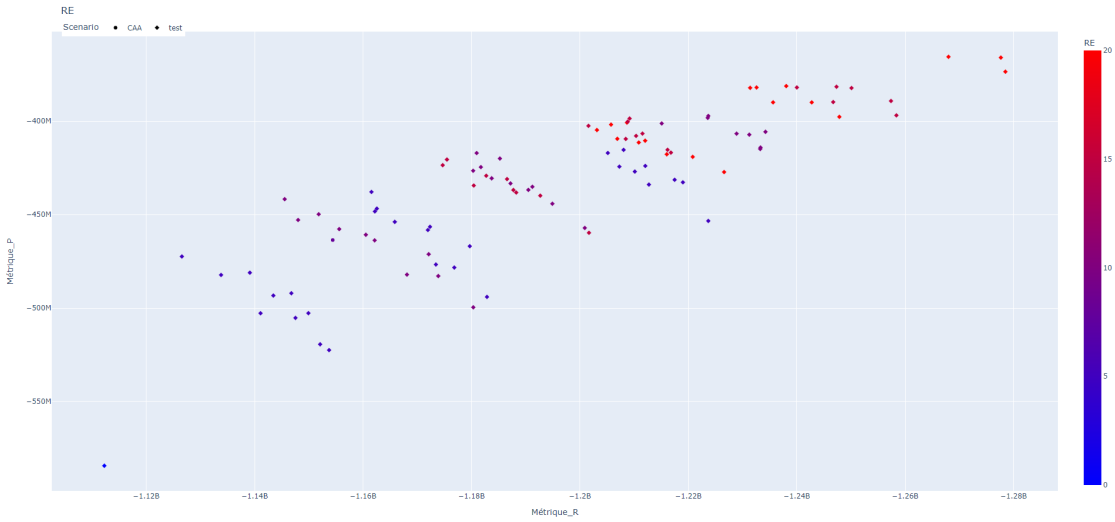

FIGURE 2.6 – Projections des portefeuilles testés afin de determiner la frontière efficiente

3 Modèle ALM

Dans cette partie nous allons définir en quoi consiste un modèle ALM. Nous rappelons que le modèle ALM a pour objectif de projeter les flux de trésorerie. Il permet entre autre de modéliser l'écoulement des flux du passif, de l'actif, mais aussi les interactions entre les deux au fil du temps. En particulier le modèle ALM permet la projection des bilans comptables et prudentiels définis dans le chapitre 2. Afin de comprendre le déroulement d'un modèle ALM nous procéderons en trois étapes.

Permièrement nous tenterons d'expliquer les mécanismes comptables à l'origine des modèles ALM.

Par la suite nous verrons la modélisation de l'actif, du passif, et interactions ainsi que leur écoulement au cours des projections.

Enfin nous mettrons en avant les caractéristiques de notre portfeuille ainsi que les hypothèses de modélisation retenues pour notre étude.

3.1 Modélisation du bilan comptable

Dans un bilan comptable l'actif représente l'ensemble des possessions d'une entreprise évaluées en valeur d'achat. Il s'agit de la face émergée de l'iceberg. Le passif quant à lui représente l'origine des capitaux qui ont permis d'acquérir ces actifs. On distingue deux origines différentes : les fonds propres et les dettes.

Dans le cas des contrats de retraites, la majeure partie des actifs est acquise à l'aide des primes des assurés. Par conséquent ils détiennent une dette sur la compagnie dont le montant est évalué par une provision mathématique. D'autres provisions techniques peuvent êtres mises en representations d'actifs au bilan, telles que les provisions pour dépréciations durables (PDD), ou risque d'éligibilité (PRE) dont l'objectif est de compenser les moins value latentes des actifs. Cependant dans notre le modèle, seules les provisions mathématiques avec les actifs qui leur correspondent sont représentées au bilan. De plus, tout résultat net issue de l'inadéquation entre la valeur d'actifs et la PM est intégré au compte cash. Lorsque le montant de cash dépasse un certain seuil, il est réalloué de manière proportionnelle à chaque classe d'actif. On peut trouver une illustration du bilan comptable du périmètre des retraites collectives d'AXA tel qu'il est représenté dans notre modèle 3.2.

3.1.1 Projection du bilan

Projeter le bilan équivaut à dérouler les flux de trésorerie du passif et de l'actif à chaque pas de temps. Pour y parvenir, le modèle dispose de scénarios stochastiques qui permettent de simuler les conditions économiques (évolution du niveau de l'inflation, des taux etc.) ainsi que des lois biométriques permettant de simuler le comportement des assurés (lois de rachat et de décès). Historiquement les modèles de projections ne disposaient que d'un scénario central. Depuis l'entrée en vigueur de S2 en 2016, les assureurs sont tenus de compléter la vision comptable de leur activité par une vision économique dite *market*

consistent. Elle consiste à générer une multitude de scénarios dans un univers risque neutre [1] afin d'évaluer les engagements et actifs en valeur de marché en appliquant les méthodes de Monte-Carlo. Cette méthode a l'avantage de valoriser les quantités d'intérêts en prenant en compte l'incertitude qui pèse sur les actifs et le comportement des assurés.

3.1.2 Estimation best Estimate

L'estimation *best estimate* décrite par l'équation 2.1 de la section 2.2.1 peut être estimée à l'aide des méthodes de Monte-Carlo décrites par la formule 3.1 et l'implémentation de l'algorithme suivant à partir des scénarios générés par le modèle.

$$Best\_estimate = \frac{1}{N_s} \sum_{s}^{N_s} \sum_{t=0}^{N_t} Flux_{t,s} \cdot D_{t,s} \tag{3.1}$$

avec :

- N_s (resp. N_t) le nombres de scénarios économiques (resp. de pas de temps par scénario)
- $Flux_{t,s}$ les flux de trésorerie de l'année t pour le scénario s.
- $D_{t,s} = \frac{1}{\left(1+r_{ZC(t,N_t)^s}\right)^t}$ le déflateur de l'annéé t pour le scénarios s avec $r_{ZC(t,N_t)^s}$ le taux donné par la courbe des taux de maturitéN_t en au temps t du scénario s.

Algorithme 1 Calcul d'un *Best Estimate*

```
nb_scenarios ← 1000
T ← 50
BE_scenarios ← liste[nb_scenarios]
pour n = 1 à nb_scenarios faire
    flux ← 0
    pour t = 1 à T faire
        flux ← flux + Flux_Actualise(num_scen = n, annee = t)
    fin pour
    BE_scenarios[n] ← flux
fin pour
BE ← MOYENNE(BE_scenarios)
```

FIGURE 3.1 – Caption

3.2 Modélisation du passif

Le passif est le montant de capital que l'assureur doit détenir à chaque instant pour être en mesure de répondre à ses engagements. La valorisation de ce dernier consiste en l'actualisation de l'écoulement des provisions techniques. Afin d'estimer correctement la PM, le modèle doit être en mesure d'accéder aux information informations de chaque *model points* du portefeuille de contrats.

1. L'existence d'un probabilité risque neutre est équivalente à vérifier la condition d'absence d'opportunités d'arbitrage, qui est une condition necessaire pour l'évaluation de portefeuille répliquant les flux des variables d'intérêts étudiés.

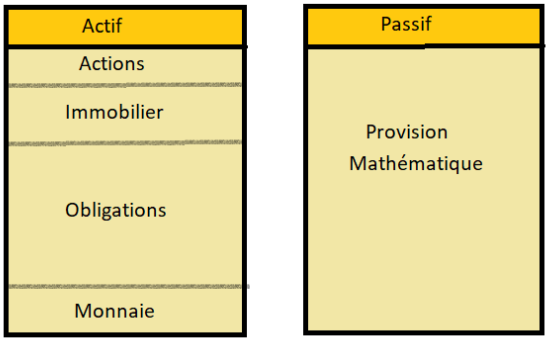

FIGURE 3.2 – Bilan comptable dans le modèle ALM implémenté

3.2.1 Modélisation d'un portfeuille de retraite

Le portefeuille d'assurés d'une compagnie d'assurance est constitué de plusieurs contrats. Chaque contrat est défini selon des caracteristiques qui lui sont propres tel que l'age du propriétaire, le taux garanti, les frais associés au contrat etc. Afin d'optimiser le temps d'optimiser le temps de calcul des projections, les contrats du portfeuille sont regroupés en *model points* (MP). Un model point est une agrégation de contrats avec des caractéristiques similaires. La modélisation des models points est particulièrement délicate puisqu'il s'agit d'anticiper le comportement des détenteurs des contrats qui le constituent. Il convient de définir correctement les caractéristiques qui gouvernent l'évolution des flux au sein d'un modèle point. Ces caracteristiques peuvent être assimilés à des hypothèses qui conditionnent le comportement de l'individu et donc la valorisation des contrats. Au sein du modèle ALM développé, le *model point* se caractérise par les éléments suivants :

- La Provision Mathématique (PM)
- le taux de produits financiers (PFI)
- le Taux Minimum Garanti (TMG)
- la marge sur encours (ENC)
- L'état du contrat, s'il est en phase de constitution ou de restitution
- L'age des assurés

Toutes ces informations sont primordiales dans le calcul de valeur des engagements dont nous allons voir le déroulé. D'autres informations sont à prendre en compte afin de pouvoir calculer la PM correctement telles que les lois biométrique des assurés membre de chaque model points. Elles permettent de calculer la Valeur Actuelle Probable (VAP) de nos engagements qui est un outil essentiel dans le calcul des différents postes du bilan comptable.

3.2.2 La Provision mathématique (PM)

Au passif le poste principal à calculer est la Provision mathématique. Elle représente la dette durable des assureurs envers les assurés. En cours de période elle est revalorisée de l'intégration des produits financiers via l'intégration de la participation aux bénéfices (PB) en cas de performance affichée des soldes techniques et financiers.

Le calcul de la provision mathématique diffère selon la phase du contrat.

Calcul de la PM en phase de constitution

Durant la phase de constitution l'assureur perçoit les primes payées par l'assuré qu'il capitalise. A date de versement de la première prime l'assureur a une provision mathématique d'un montant équivalent à la prime déduite des chargements :

$$PM(t) = Prime\_brute(t) - chargement\_sur\_prime = Prime\_nette(t)$$

Au cours du temps, le niveau de PM varie en fonction des détails du contrat, du comportement des assurées et de leur durée de vie. En effet à chaque pas de temps les primes versées sont investies sur un fonds euros. En contrepartie l'assureur doit verser un taux minimum garanti. Il doit donc conserver le montant de la prime capitalisée dans ses provisions. Or en cours d'années certains assurés rachètent leur contrat ou décèdent qu'il faut donc déduire de la prime capitalisée. On a donc :

$$PM(t+1) = PM(t) \cdot (1 + TMG) \cdot (1 - q_x(t) - r(t))$$

avec , $q_x(t)$ et (t) les probabilités de décès et rachats.

Phase de restitution En phase de restitution, (i.e lors du départ à la retraite de l'assuré) l'assureur restitue le capital constitué sous forme d'arrérages jusqu'à la mort de l'assuré. On a :

$$PM\_restitution(t, x) = \sum_{t=0}^{\omega - x} \frac{Arrerage}{(1 + i_{technique})^t (1 + g)} * p_{x|t} \tag{3.2}$$

avec :

- x l'age de l'assuré, ω l'age maximal de la table de mortalité
- $i\_technique$ Le taux d'intérêt technique qui correspond à une valeur entre 60 et 75 % du Taux Moyen d'Etats. Ce taux prudent est inférieur à ceux donné par la courbe des taux pour anticiper afin de surévaluée la PM en cas de choc sur les taux.
- $p_{x|t}$ la probabilité de survie d'un individu agé de x années pendant t années.
- g le taux de chargement dû au versement de la rente.

Le montant d'arrérage est constant et determiné lors du passage en rente de la PM de l'assuré. Il vérifie :

$$PM\_restitution(0, x = age\_retraite) = CC. \tag{3.3}$$

où CC est le capital accumulé pendant la phase de constitution.

3.2.3 Calcul de la participation aux benefices (PB)

Le Code des assurances exige de la part des assureurs de verser une participation aux bénéfices (PB). La participation aux bénéfices est primordiale dans un modèle ALM. Elle est la passerelle entre l'actif et le passif. En effet, la PB sur les produits financiers tirés des actifs impactent le montant des provisition techniques (passif). A chaque pas de temps la PM est revalorisée puis gonflée de la valeur de PB selon un calcul que nous précisons.

Il convient de distinguer la PB contractuelle de la PB réglementaire :

- **PB réglementaire :** Le calcul de la PB réglementaire se fait au niveau de l'entreprise d'assurance et non par contrat ou fond. La rémunération globale de la compagnie à partir des primes ne peut être supérieur à 10 % du résultat technique (i.e résultat lié à la mortalité et à la gestion) et 15% du résultat financier des produits financiers sur engagements. Autrement dit, le montant minimum du taux de PB sur le résultat technique est de 90% (resp.85% sur le résultat financier).
- **PB contractuelle** : Elle peut être définie au niveau du contrat ou d'un ensemble de contrats. Le montant total des PB contractuelles d'une compagnie est souvent supérieur à celui des participations réglementaires.

Calcul de PB sur le solde financier

On note ENC, le niveau de marge du MP considéré et α un pourcentage de produits financiers distribués pour encadrer le niveau de participation aux bénéfices financiers minimal contractuel. En notant PFI le montant total des produits financiers réalisés, en général nets des frais de placement, et les intérêts techniques (IT) issus du taux minimum garanti (TMG) ou du taux technique de la rente, on obtient une participation aux bénéfices complémentaire aux intérêts techniques :

$$PB\_financiere = max(0, \alpha \cdot PFI - IT - ENC)$$

Les intérêts techniques sont calculés différement selon la phase du cycle de vie des contrats. On a donc :
En phase de constitution,

$$IT\_contit(t) = PM\_constit(t) \cdot TMG + (Primes(t) - Prestations(t)) \cdot ((1 + TMG)^{\frac{1}{2}} - 1))$$

En phase de restitution,

$$IT\_restit(t) = PM\_restit(t) \cdot TMG - Prestations(t) \cdot ((1 + TMG)^{\frac{1}{2}} - 1))$$

PB sur le solde technique Le résultat technique du contrat est défini comme suit :

$$Resultat\_technique(t) = Primes\_nettes(t) + PM\_ouverture(t) - PM\_cloture(t+1) + IT(t) - Prestations\_brutes(t)$$

La PB technique se calcule d'après la formule suivante :

$$PB\_technique = max(0, \beta \cdot resultat\_technique)$$

où β désigne le taux de distribution des produits technique d'un montant minimum de 90%.

Mutualisation des resultats technique et financiers : Pour les contrats qui comprennent une mutualisation technico-financière, l'assureur se laisse la liberté de compenser les pertes du compte de resultat technique par la PB financière. On a donc :

$$PB_{technico\_financi\%re} = max(0, PB_{financiere} + \cdot Resultat_{technique})$$

Calcul des provisions techniques

Dans notre modèle les provisions techniques se réduisent à la PM augmentée de la PB. [2] L'incorporation de la PB à la PM diffère en fonction de la phase du contrat.

En phase de constitution on a :

$$PM\_constit(t+1) = PM\_constit(t) \cdot (1 - q_x(t) - r(t)) + IT\_constit(t) + PB\_financiere(t)$$

Soit $PM\_restit\_avantPB(t+1)$ la PM de début de période diminuée des arrérages versés pendant l'année en cas de survie de l'assurée. En phase de restitution l'integration de la PB à la PM est calculée comme suit :

$$PM\_restit\_apresPB(t+1) = PM\_restit\_avantPB(t+1) + IT\_constit(t) + PB\_financiere(t)$$

Cette incorporation donne lieu à une revalorisation de l'arrérage afin de prendre en compte le produit financier dans la rente.

3.2.4 Projection du passif

En vertu des calculs de revalorisation de la PM décrits ci-dessus on peut décrire la chronologie des flux de passif comme suit dans la figure 3.3

3.2.5 Carcateristiques du portefeuille étudié

Notre étude porte sur un portefeuille de retraites collective "Article 83". Ces contrats "individualisés" bénéficient d'une phase de constitution qui peut donner lieu à des rachats (à titre individuel ou de l'entreprise), puis d'une phase de restitution avec rente. Les assurés peuvent opter pour une sortie de capital sous forme d'indemnité ou de rente. Le portefeuille est constitué de 916 models points dont 608 en phase de constitution et 308 en phase de restitution. La provision mathématique de l'ensemble du portefeuille est de 13,1 milliards d'euros.

TMG

Plusieurs niveaux de taux garantis (TMG ou taux technique) sont pris en compte (cf.figure 3.5) afin de de prendre en compte la diversité du portefeuille qui s'est construit sur plusieurs générations de taux.

Partage des produits financiers et marge sur encours

En ce qui concerne la marge sur encours et le niveau de partage des produits financiers, on peut considérer que ces niveaux sont très homogènes dans le portefeuille, avec toutefois deux niveaux pour chaque paramètre. Cela correspond à la fois à une réalité du portefeuille réel, tout en limitant le nombre de model points modélisés et en simulant des niveaux de marge différents pour les contrats. Les tables de la figure 3.2.5 présentent les paramètres choisis.

2. En règle générale les, en plus de la PM et de la PB, provisions techniques regroupent d'autres compte de provision tels que la Provision pour dépreciation durable , la provision pour risque d'exigilibilité ... Dans notre modèle nous nous contentons de la PM et la PB. On a donc $PT = PM + PB$

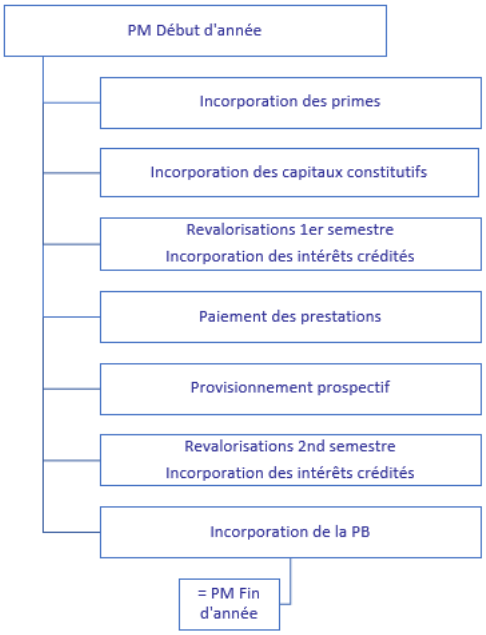

FIGURE 3.3 – Etapes de projection de la PM

	Provision Mathématique à t=0 (en M€)	taux garanti	âge moyen	marge sur encours moyenne	% partage des produits financiers
Constitution	8 816	0,87%	47,5	0,50%	97,50%
Restitution	4 312	1,50%	72	0,50%	97,50%
TOTAL	13 128	1,08%	55,5	0,50%	97,50%

FIGURE 3.4 – caracteristiques Moyenne du portefeuilles

TMG/ taux technique	Constit	Restit
0,00%	25%	14%
0,50%	25%	14%
1,00%	25%	14%
1,50%	0%	14%
2,00%	25%	14%
2,50%	0%	14%
3,00%	0%	14%

FIGURE 3.5 – Repartition des taux minimum garanti

% partage des produits financiers	Constit	Restit
95%	50%	50%
100%	50%	50%

Marge sur encours	Constit	Restit
0,30%	50%	50%
0,70%	50%	50%

FIGURE 3.6 – Caption

Lois biométriques

Nous rappelons que l'écoulement de la PM est impacté par les caractéristiques des assurés qui sont modélisés par les lois biométriques :

- Loi de mortalité : Les lois de mortalité sont calibrées, elles indiquent les probabilités de survie des assurés par classe d'age. Le stables utilisées pour notre portefeuille sont les tables TG-TH-02.
- Lois de rachats : Le taux de rachat partiel annuel est fixé à 2% . De plus une loi de rachat dynamique est simulée pour modéliséer au mieux la compétition sur les marchés financiers.

Taux de chargement

- taux de chargement sur rentre : 2%
- taux de chargement sur la pm : 0.25%
- taux de chargement sur clms :1%

3.3 Modélisation de l'Actif

La valeur de l'actif de notre modèle ALM est défini comme la somme des valeurs de marché des actifs financiers qui le compose. Notre modèle modélise 3 classes d'actifs : les actions, les obligations ainsi que les titres immobiliers.

D'un point de modélisation les obligations sont classés dans les actifs amortissables tandis, que les titres immobiliers et actions, dans la catégorie des actifs non amortissables.

- actifs amortissables : Selon l'Article R343-9 du Code des Assurances, il s'agit de titres donc la valeur comptable du bien est modifiée chaque année. Dans notre modèle seule les obligations à taux fixes en font parties. La différence entre la valeur d'achat et le nominal est amorti au cours du temps via le recours à un coefficient déterminé par une interpolation,par exemple.
- les actifs non amortissables : Les actifs non amortissables sont définis dans l'Article R343-10 du Code des Assurances. Il s'agit de l'ensemble des actifs n'étant pas énumérés à l'Article R343-9. Ces derniers sont comptabilisés sur la base du prix d'achat, ou de revient.

Cette distinction justifie le traitement différencié des classes d'actifs avec d'une part les obligations, de l'autre les actions et titres immobiliers.

3.3.1 Les obligations

Une obligation est un titre de dette. Le détenteur de l'obligation achète un titre de dette auprès d'un agent en besoin de financement. Pour caracteriser une obligation nous avons besoin de 3 élément :

- La maturité (T)
- Le nominal (N)
- Le taux de coupon (c)
- Une corube de taux Zéro Coupons $R_{ZC(t,T)}$ pour $t \in [\![0, T]\!]$

En contrepartie d'avoir prété un montant égal au nominal le détenteur d'une obligation se voit verser un coupons à chaque période de temps jusqu'à maturité. A maturité, en plus du coupon le débiteur rembourse le montant du nominal. On déduit la valeur boursière (VB) comme suit :

$$VB\_oblig(t_0) = \sum_{t=0}^{T-t0-1} \frac{N \cdot c}{(1 + r_{ZC(t,T)})^t} + \frac{N \cdot (1 + c)}{(1 + r_{ZC(t,T)}))^T}$$

3.3.2 Les actions et titres immobiliers

Les actions sont des titres de propriétés émis par une entreprise. En fonction du contrat, ils peuvent délivrer ou non des dividendes. Les titres immobiliers peuvent être modélisés d'une manière semblable. A la place de recevoir des dividendes le propriétaire reçoit des loyers. La différence entre les deux réside dans la liquidité des actifs. En effet, il est plus difficile d'effectuer des transactions avec des titres à valeurs immobiliers que des titres en valeurs mobilières. Il n'y a pas de règle concernant la valorisation de ces actifs,ils dépendent exclusivement du scénario économique en vigueur. Ces classes d'actifs se caractérisent par :

- leur valeur nette comptable (VNC) ou book value (BV) : Elle correspond à la valeur de l'actif à la date d'achat. Elle ne change pas en fonction des conditions économiques
- leur valeur boursière (VB) ou market value (MV) : La VB correspond à la valeur de marché des titres.
- l'indice de marché dont ils dépendent : La valeur de marché des actifs non amortissables est indexée au cours boursier boursier dont ils dépendent. Les fluctuatations de l'indice et par conséquent de leur valeur boursière est décrite par les scénarios stochastiques générés par le modèle.

3.3.3 La Monnaie

La monnaie enregistre toutes les plus-moins values réalisés au cours du temps. Dans notre modèle elle sert de variable d'ajustement de sorte à égaliser le montant d'actif et de passif sur chaque période.

3.3.4 Flux financiers liés à la réallocation d'actifs

En début de projection la VB ainsi que la VNC sont initialisées par la valeur du MP en input. A chaque pas de temps la VNC et la VB sont initialisées par la valeur en fin de période. Les flux de trésorerie qui provoquent une modification de la VB, de la VNC et du cash du portfeuille sont de 3 ordres :

1. Ils correspondent à une revalorisation des actifs financiers dûe à l'évolution des conditions de marché(ne concerne que la VB).
2. Ils proviennent de la distribution de revenus inhérents à la possession de ces actifs (obtention de dividendes) (ne concerne que le cash).
3. Ils sont la résultantes d'opération des opérations de réallocations (concerne la VNC et la VB).

Impact de la performance des marchés sur la valeur de l'actif En début de période le montant de VB est hérité de celui de la période précédente. Il est revalorisé en fonction des performances de l'indice boursier décrits par les scénarios économiques. Pour isoler l'évolution net de l'actifs avec celle des dividendes obtenus,on définit le taux net comme la différence entre le taux de performance brute et de dividendes :

$$VB_{avant\_perf}(t) = VB(t-1) \cdot (1 + taux_{net})$$

La VNC est un montant comptable qui n'est pas impacté par la conjoncture économique. A l'issue de l'application des resultats de performance on a donc $VNC(t) = VNC(t-1)$

Impact du calcul des revenus sur l'actif La distribution de revenu est inhérent à la possession des actifs considérés. Ils génèrent un taux de dividende propre à la classe d'actif considéré donné par le modèle. La VB et la VNC ne sont pas concerné par les flux issus de la distribution de revenu. Ces revenus sont donc intégrés à la monnaie. On a donc

$$Cash_{apres\_perf(t)} = cash(t-1) + Revenu(t)$$

avec $Revenu(t) = VB(t-1) \cdot taux_{dividende}(t)$

Impact de la réallocation sur la valorisation de l'actif La reallocation d'actif consiste à vendre/acheter des actifs financiers sur les marchés afin d'atteindre une allocation cible. Cette modification du volume et de la valeur des actifs financiers impactent à la fois la VNC et la VB. Ces opérations dégagent des plus values en cash intégrées au compte de l'entreprise.

Impact d'une opération de vente d'actifs sur la VB et la VNC Nous rappelons que la VNC est la valeur comptable lors de l'achat d'actif en to. On a donc $VB_{apres\_vente}(t) = VB_{avant\_vente}(t) - montant\_vendu(t)$. La vente d'actif a aussi un impact sur la VNC qui voit son montant réduire proportionnellement à la variation du montant de VB d'où :

$$VNC_{apres\_vente}(t) = VNC(t) \cdot \frac{VB_{apres\_vente}}{VB_{avant\_vente}}$$

De plus, les ventes réalisent une plus ou moins values PMVR sur le compte de cash. En effet,$\Delta VNC = VNC_{apres\_vente} - VNC_{avant\_vente}$ correspond à la valeur de l'actif au prix de vente tandis que $\Delta VNC = VB_{apres\_vente} - VB_{avant\_vente}$ correspond à la valeur de l'actif au prix d'achat. La différence entre les deux est soit une plus value ou une moins value d'où :

$$PMVL = \Delta VB - \Delta VNC = montant\_vendu - montant\_achete$$

Impact d'unie operation d'achats d'actifs sur la VB et la VNC L'achat d'actif en t entraine la création d'un nouveau MP : $VB(t) = VNC(t) = Prix\_actif$ avec l'indice boursier qui lui correspond.

3.3.5 Projection de l'actif

Les étapes de la projection l'actif sont détaillées chronologiquement dans la figure 3.7

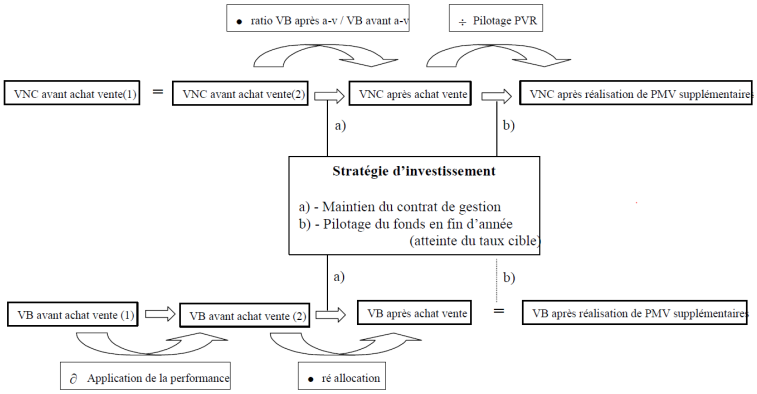

FIGURE 3.7 – Projection de la VB et VNC dans le cadre d'une stratégie d'investissement

	Valeur Comptable à t=0 (en M€)	Valeur Boursière à t=0 (en M€)	Plus-value Latente à t=0	Taux de PVL	Allocation (hors cash)
GOVIES	10 412	16 077	5 665	54%	83%
ACTIONS	1 167	1 637	470	40%	9%
IMMOBILIER	940	1 281	341	36%	8%
CASH	609	609	-	0%	
TOTAL	13 128	19 605	6 477	49%	100%

FIGURE 3.8 – Volumes et caractéristiques moyennes principales de l'actif

	Inflation	Equity TR	Equity div	Real Estate TR	Real Estate div
Mean	1,94%	5,15%	2,39%	4,71%	2,35%
Std Variation	3,18%	18,00%	1,76%	9,52%	0,33%
Minimum	-9,06%	-51,92%	0,00%	-31,94%	1,23%
Q1 - 25%	-0,10%	-7,57%	1,28%	-1,92%	2,12%
Q2 - 50%	1,70%	3,67%	2,01%	4,25%	2,33%
Q3 - 75%	3,77%	16,20%	2,91%	10,82%	2,56%
Maximum	20,20%	122,04%	35,58%	64,10%	4,35%

FIGURE 3.9 – Statistiqques descriptives des scénarios utilisés

	Inflation	Equity TR	Equity div	Real Estate TR	Real Estate div
Inflation	100%	13%	-19%	21%	-4%
Equity TR		100%	-8%	40%	-4%
Equity div			100%	-8%	21%
Real Estate TR				100%	-10%
Real Estate div					100%

FIGURE 3.10 – Corrélation entre mes scénarios

3.3.6 Caractéristiques du portefeuille d'actifs considéré

Les caractéristiques de l'allocation initiale de notre portefeuille d'actifs sont détaillées dans la figure 3.8. Les obligations sont sur-représentées par rapport aux autres classe d'actifs. En effet, elles disposent d'une plus grande sécurité et limite les risques d'insolvabilité de la part de l'assureur. A l'image de l'actif, le grand volume de plus values latentes s'explique par l'ancienneté de certains titres dans le portefeuille AXA. En effet, plus le titre détenu est ancien plus la différence entre la valeur boursière et la valeur nette comptable (i.e la plus value latente) est marquée.

L'écoulement des performances de chaque actif dépend des conditions économiques simulées par les scénarios. l'évolution de leur valeur ainsi que de leurs performances sont indexées aux variations d'indices boursiers. Dans notre étude nous disposons de 2000 scénarios économiques générés par les équipes du risk-management du groupe AXA, et utilisés par les équipes ALM dans leurs modèles. Il s'agit de scénarios dits *real world*. Cela signifie que contrairement aux scénarios risques neutres utilisés pour calculer le SCR, ils modélisent une prime de risque pour les atifs risqué afin de modéliser la marge financière généré par l'activité de l'assureur. Les hypothèses retenues sont détaillées dans la figure 3.3.6

3.4 Déroulement du modèle ALM

Après avoir vu la modélisation ainsi que les mécanismes de vieillissement des MP d'actif et de passif nous sommes à présent en mesure de décrire le modèle ALM implémenté dans son ensemble. Le modèle suit les étapes suivantes :

1. Simulation des nouvelles conditions de marché : modifie le niveau des taux, de l'inflation

(en M€)	Brut IS	Net IS
VIF	630	428
SCR	858	742
RM	268	182
EOF hors Fonds propres		246
Besoin minimal en Fonds Propres		496

FIGURE 3.11 – Etat de solvabilité du portefeuille

2. Réevaluation des actifs en fonction des nouvelles conditions économiques (niveau des indices boursiers,courbe de taux).
3. Ajustement des écarts entre l'actif et le passif. Les pertes ou excedents sont tranférés sur le compte de cash l'actif avec le passif la NAV est transférée sur le compte de cash (si l'allocation du niveau de cash dépasse 15% de l'allocation d'actif, la partie excedentaire est réallouée de manière proportionnelle aux différentes classes d'actifs)
4. Réallocation des actifs
5. Détermination des produits technico-financiers après la réalloc
6. Revalorisation des engagement : intégration des pfi dans le calcul de la PM
7. Calcul des métriques de solvabilité (SCR et BEL).

Métriques de solvabilité du portefeuille

Les métriques de solvabilité en vision économique à $t = 0$ sont les suivantes (cf. 3.11)

4 Deep reinforcement learning pour la recherche d'allocation stratégique

Les récents développement du reinforcement learning ouvrent de nouvelles perspectives d'application dans l'industrie. D'après [13], la gestion d'actif passif ne fait pas exception. Ainsi dans ce chapitre, nous nous attachons à implémenter un modèle ALM dans lequel les décisions d'investissement sont définies par un agent entrainé par renforcement.

Tout d'abord nous présenterons le paradigme et de l'apprentissage par renforcement et étudierons sa compatibilité avec notre problème d'optimisation de la stratégie d'investissement

Puis nous implémenterons le modèle deep ALM qui consiste à substituer les décisions définies par la direction des investissements par celles d'un agent entrainé par apprentissage par renforcement. En particulier nous implémenterons deux agents différents dont nous verrons les détails

4.1 Le reinforcement learning

4.1.1 Généralites

Le reinforcement learning, met en scène un agent qui interagit dans un environnement dynamique avec pour objectif d'optimiser son utilité suite à la réalisation d'une tache donnée.

A chaque pas de temps, l'agent reçoit un signal qui synthétise sa perception (imparfaite) de l'environnement. Fort de cette information et de la connaissance accumulée au fil du temps, il élabore une stratégie qui optimise son bien-être de long terme. Lors de chaque prise de décision, l'agent provoque une interaction avec son environnement. Cette dernière modifie l'état du monde dans lequel il se trouve et lui transmet un *feedback* qui lui permet d'ajuster sa stratégie.

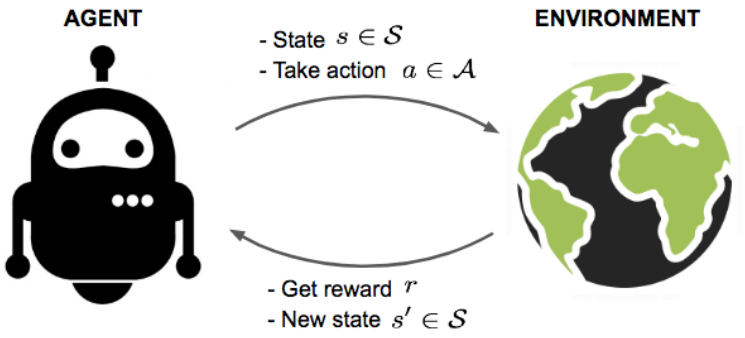

FIGURE 4.1 – Interaction between agent and environment

Etant donné que l'objectif de l'agent est d'optimiser son bien-être sur le long terme, il ne peut se contenter de prendre les décisions qui lui procurent la meilleure récompense à l'instant t. Il doit au préalable évaluer l'impact que sa décision aura dans le futur. La difficulté de cette tache réside dans le caractère dynamique de l'environnement dont il n'a qu'une connaissance partielle.

Pour parvenir à son objectif, l'agent doit donc explorer son environnement et accumuler suffisamment de connaissances qui lui permettent de définir une stratégie robuste qui prend en compte les évolutions du système avec lequel il interagit. En définitive pour définir un problème de reinforcement learning, quatre composantes essentielles sont requises. Deux qui sont relatives à l'environnement et deux à l'agent :

- **Une espace d'états (environnement)** : La variable d'état permet de synthétiser l'information du système perçue par l'agent.
- **Une fonction de reward (environnement) :** le reward est le signal reçu de l'environnement par l'agent suite à une prise de décision. Il peut être positif ou négatif, il permet à ce dernier d'ajuster sa stratégie.
- **Une politique (agent)** : la politique correspond à la stratégie établie par l'agent. Il s'agit d'une séquence d'actions définie à chaque état.
- **Une fonction de valeur (agent) :** Il s'agit de la fonction de gains espérés qui permet d'évaluer la politique de l'agent à chaque état.

Un peu d'histoire...

Politique et *fonction de valeur* sont des concepts empruntés au *contrôle optimal* dont le reinforcement learning se fait héritier. En effet, depuis les années 50, Bellman s'est interessé à résoudre des problèmes d'optimisation de contrôle sur des systèmes (i.e environnement) dynamiques sur lesquels on peut agir au moyen d'une politique (ou contrôle). Le but est alors d'amener le système d'un état initial à un certain état final, en respectant éventuellement certains critères à optimiser. Afin d'y parvenir Bellman développe des processus iteratifs qu'il définit de programmation dynamique.

La formalisation des problèmes de controle optimal et programmation dynamique sont très similaires à celles du reinforcement learning. Il en devient difficile de les dissocier. Cependant d'après Sutton et Barto dans [12], l'apprentissage par renforcement se distingue des deux premiers par le caractère incertain du système dans lequel l'agent évolue. Il doit l'explorer pour "apprendre" et ainsi "renforcer" ses décisions :

" In this book, we consider all of the work in optimal control also to be, in a sense, work in reinforcement learning. [...] Accordingly, we must consider the solution methods of optimal control, such as dynamic programming, also to be reinforcement learning methods. Of course, almost all of ***these methods require complete knowledge of the system to be controlled****, and for this reason it feels a little unnatural to say that they are part of reinforcement* ***learning****."*

En plus d'hériter de ses concepts, l'apprentissage par renforcement hérite de la grande applicabilité du contrôle optimal. En effet, les systèmes étudiés par le controle optimal sont divers : aéronotique, automobile, mécanique, biologie, etc. Récemment le reinforcement learning a connu d'énorme progrès dans le domaine des jeux vidéos ou de société à l'image d'Alpha Zero, le premier algorithme à battre le champion du monde de go. De nombreuses pistes sont actuellement explorées à l'instar des recherches menées par DeepMind en génétique ou encore pour la voiture autonome.

4.1.2 Formalisation d'un problème de reinforcement learning

Dans cette partie nous allons étudier les principes du reinforcement learning sous un angle plus formel. Nous nous inspirons de la modélisation développée par Sutton et Barto dans [12]. Les auteurs présentent les bases de la modélisation, qui s'appuient sur des processus de décision markovien (MDPs). Ces derniers donnent un cadre cohérent aux briques de bases définies précédemment (état du système, politique, fonction de récompense, fonction de valeur). Un processus de décision markovien est un quadruplet S, A, T,R définissant :

- un ensemble d'états $\mathcal{S}$, qui peut être fini, dénombrable ou continu ; cet ensemble définit l'environnement tel que perçu par l'agent.
- un ensemble d'actions $\mathcal{A}$, qui peut être fini, dénombrable ou continu et dans lequel l'agent choisit les interactions qu'il effectue avec l'environnement.
- une fonction de transition $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longrightarrow [0;1]$. cette fonction définit l'effet des actions de l'agent sur l'environnement. En particulier, $\mathcal{P}^a_{ss'} = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$ et caractérise la probabilité de se retrouver dans l'état s′ après avoir effectué l'action a dans l'état s.
- une fonction de récompense $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longrightarrow \mathbb{R}$. Elle définit la récompense (positive ou négative) reçue par l'agent. En particulier, $R(s, a, s')$ est la récompense obtenue en t+1 pour être passé de l'état s à s′ en ayant effectué l'action a.

Les processus de décision markovien caractérisent parfaitement les modèles à informations parfaites pour lesquelles l'information de l'état actuel résume parfaitement l'historique des actions antérieures. Ces derniers vérifient la propriété de Markov :

$$\mathbb{P}(R_{t+1} | S_1, A_1, S_2, ..., A_{t-1}, S_t) = \mathbb{P}(R_{t+1} | S_t) .$$

Sutton et Barto affirment que si cette propriété n'est pas parfaitement vérifiée pour le problème étudié, le modèle peut toutefois partir de l'information contenue dans l'état courant et en tirer son apprentissage. Un problème se rapprochant de cette hypothèse obtiendra donc de meilleures performances. Dans le cas présent, il est possible de penser qu'on peut appliquer la théorie markovienne à ce problème dans la mesure où cette hypothèse est couramment utilisée pour la valorisation financière.

Déroulement du jeu adapté à la gestion ALM

Formellement on peut définir le cadre général de la manière suivante : à chaque étape de décision $t \in [\![1, N]\!]$, l'agent se trouve à un état précis $S_t \in \mathcal{S}$ où lui est permis un ensemble d'actions $A_t \in \mathcal{A}(S_t)$. En fonction de l'action choisie en t, l'agent perçoit à l'étape d'après (en t + 1), une récompense $R_t \in \mathbb{R}$ et se retrouve dans un nouvel état du jeu S_{t+1}. Une politique décrit les choix des actions à jouer par l'agent dans chaque état. Formellement, la politique est modélisée par une fonction $\pi : \mathcal{S} \longrightarrow \mathcal{A}$ dans le cas d'une politique déterministe, et $\pi : \mathcal{S} \times \mathcal{A} \longrightarrow [0;1]$ dans le cas stochastique. On note $\pi(a|s)$ la probabilité de jouer a dans l'état s à l'instant t, i.e. $\mathbb{P}[A_t = a | S_t = s]$. L'agent choisit une politique à l'aide de la fonction de récompense R. Notons $R_t = R(S_t, A_t, S_{t+1})$ la récompense effective obtenue après avoir effectué l'action A_t par l'agent qui suit la politique π. Nous rappelons que l'agent cherche à maximiser son gain sur le long terme. Ce qu'il cherche à maximiser c'est son reward cumulé sur toute la période. A titre d'exemple, voici plusieurs critères d'intérêts que l'agent peut chercher à maximiser :

- $\mathbb{E}_\pi \left[\sum_{t=0}^{h} r_t \right]$ espérance de la somme des récompenses à un horizon fini fixé h.

- $\liminf_{h\longrightarrow\infty} \mathbb{E}_\pi \left[\frac{1}{h}\sum_{t=0}^{h} r_t\right]$ ou $\limsup_{h\longrightarrow\infty} \mathbb{E}_\pi \left[\frac{1}{h}\sum_{t=0}^{h} r_t\right]$: récompense moyenne à long terme.
- $\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$: récompense escomptée (ou amortie) à horizon infini où $0 \leq \gamma < 1$

Dans la suite nous maximiserons la récompense escomptée à horizon infini où $0 < \gamma < 1$. γ traduit la préférence de l'agent pour le présent. Ainsi si $\gamma \longrightarrow 0$ cela signifie que l'agent manifeste une préférence pour le présent. Inversement si $\gamma \longrightarrow 1$, l'agent accorde presque autant d'importance au futur qu'au présent.

Fonctions de valeurs

Lorsqu'une politique et un critère sont déterminés, deux fonctions centrales peuvent être définies :

- $V_\pi : \mathcal{S} \rightarrow \mathbb{R}$: c'est la fonction de valeur des états ; $V_\pi(s)$ représente le gain (selon le critère adopté) engrangé par l'agent s'il démarre à l'état s et applique ensuite la politique π ad infinitum. Dans le cas de gains escomptés on a $V_\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} R_{t+k+1} | S_t = s\right]$.
- $Q_\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: c'est la fonction de valeur des états-actions ; $Q_\pi(s,a)$ représente le gain engrangé par l'agent s'il démarre à l'état s et commence par effectuer l'action a, avant d'appliquer ensuite la politique π ad infinitum. Dans le cas de gains escomptés on a :
  $Q_\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} R_{t+k+1} | S_t = s, A_t = a\right]$

Ces fonctions sont essentielles dans la modélisation. C'est à partir d'elles que l'agent évalue la qualité de la politique mise en oeuvre. On remarquera la relation suivante :

$$V_\pi(s) = \sum_{a\in\mathcal{A}(s)} \pi(a|s) Q_\pi(s,a) \tag{4.1}$$

Problème d'optimisation associé

L'objectif d'un problème d'apprentissage par renforcement consiste à trouver la politique qui maximise la fonction de valeur de l'agent. Autrement dit il s'agit de trouver $\pi_*, \forall s \in \mathcal{S}\ V_{\pi_*}(s) = \max_\pi V_\pi(s)$.

4.1.3 Apprentissage de l'agent

La résolution du problème consiste en l'optimisation de la stratégie de l'agent. Afin d'y arriver l'agent doit être en mesure de :

- Evaluer correctement l'impact de ses décisions sur son bien-être
- Apprendre de ses experiences pour maximiser ses actions au fil des épisodes.

Afin de répondre à ce de double objectif, les équations de Bellman consituent un résultat fondamental à la base de la majorité des algorithmes d'apprentissage par renforcement.

Les équations de Bellman

Théoreme 4.1.1 (Equations de Bellman) $\forall a \in \mathcal{A},\ \forall s, s' \in \mathcal{S}$,

$$Q_\pi(s,a) = R(s,a,s') + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} \sum_{a'\in\mathcal{A}} \pi(a'|s') Q_\pi(s',a') \tag{4.2}$$

$$V_\pi(s) = \sum_{a\in(A)} \pi(a|s) \left(R(s,a,s') + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} V_\pi(s') \right) \tag{4.3}$$

Corollaire 4.1.1.1 (Equations de Bellman optimales) *Soit π une politique, et π_* la politique optimale on a $\pi = \pi_*$ si et seulement si, une des condition ci-dessous est satisfaite :*

$$V_{\pi_*}(s) = \max_{a \in \mathcal{A}} Q_{\pi_*}(s,a) \tag{4.4}$$

$$Q_{\pi_*}(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} V_{\pi_*}(s') \tag{4.5}$$

$$V_{\pi_*}(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} V_{\pi_*}(s')\right) \tag{4.6}$$

$$Q_{\pi_*}(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} \max_{a' \in \mathcal{A}} Q_{\pi_*}(s',a') \tag{4.7}$$

Les équations du théorème 4.1.1 introduisent une relation de récurrence sur les fonctions entre $Q_\pi(S_{t+1}, A_{t+1})$, $Q_\pi(S_t, A_t)$ et $V_\pi(S_{t+1}), V_\pi(S_t)$. En particulier elles définissent un processus itératif qui permet d'actualiser les connaissances de l'agent en fonction du reward obtenu lors de chaque décision prise par ce dernier. Par ce processus, l'agent est en mesure d'évaluer au mieux sa stratégie.

De plus, les équation optimales du corollaire 4.1.1.1 décrivent un second processus itératif qui permet d'orienter les choix de l'agent lors de chaque décision prise vers la politique optimale. En effet, elles décomposent le problème d'optimisation de politique initiale (global), en une succession d'optimisation sur l'espace des actions à chaque intervalle de temps.

On a :

$$\pi' \geq \pi \Leftrightarrow \forall s \in \mathcal{S},\ V_{\pi'}(s) \geq V_\pi(s) \Leftrightarrow \forall s \in \mathcal{S},\ Q_{\pi'}(s, \pi'(s)) \geq Q_\pi(s, \pi(s)) \tag{4.8}$$

On vérifie bien qu'en appliquant l'algorithme décrit dans la figure 4.2, l'agent est en mesure d'améliorer sa stratégie à chaque prise de décision. En effet : Soit π on définit π' sa version améliorée qui vérifie $\pi'(s) = \arg\max_{a \in \mathcal{A}} Q_{\pi'}(s,a)$. On a :

$$\begin{aligned} Q_\pi(s, \pi'(s)) &= Q_\pi(s, \arg\max_{a \in \mathcal{A}} Q_\pi(s,a)) \\ &= \max_{a \in \mathcal{A}} Q_\pi(s,a) \geq Q_\pi(s, \pi(s)) = V_\pi(s) \end{aligned}$$

Le processus s'arrête lorsque le corollaire 4.1.1.1 est vérifié (toutes les équations sont équivalentes).

4.1.4 Deep Reinforcement learning

Dans la phase d'amélioration de l'algorithme 4.2 à chaque iteration, l'agent doit effectuer le calcul d'optimisation suivant :

$$a* = \arg\max_a Q_\pi(s,a) \tag{4.9}$$

Quatre cas de figure se présentent :

1. $Card(\mathcal{A}) < +\infty, Card(\mathcal{S}) < +\infty$
2. $Card(\mathcal{A}) = +\infty, Card(\mathcal{S}) < +\infty$
3. $Card(\mathcal{A}) < +\infty, Card(\mathcal{S}) = +\infty$

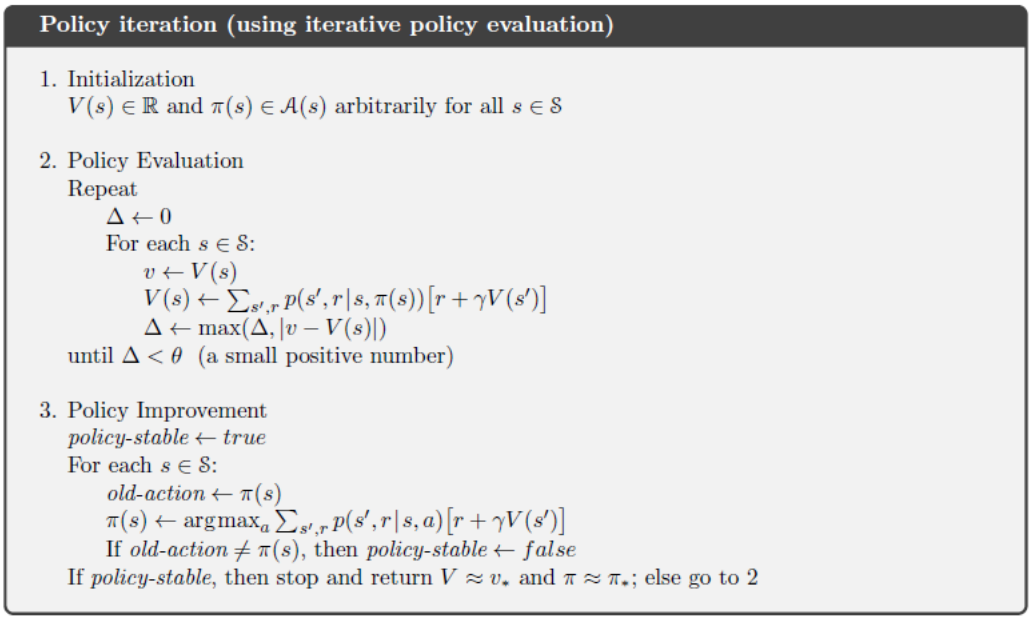

FIGURE 4.2 – Algorithme d'optimisation de la politique par iteration sur la fonction de valeur

4. $Card(\mathcal{A}) = +\infty, Card(\mathcal{S}) = +\infty$

Lorsque le nombre d'états et d'actions est fini (cas 1), la fonction de valeur peut être représentée par une matrice en dimensions finie dans laquelle chaque ligne correspond à une action $a \in \mathcal{A}$ et chaque colonne un état $s \in (S)$. Chaque scalaire de la matrice est la fonction de valeur $Q_\pi(s,a)$. Dans ces conditions, en fonction du cardinal de $\mathcal{A}$ et $\mathcal{S}$, le problème d'optimisation 4.9 est soluble assez rapidement. Si le nombre d'actions n'est plus fini (cas 2,3,4), il devient impossible de représenter le problème sous forme de tableau. $Q_\pi(s,a)$ est alors modélisée par une fonction. La fonction choisie doit être en mesure de capter au mieux les caracteristiques de l'environnement afin d'estimer correctement les valeurs des actions de l'agent, lui permettant de prendre des décisions optimales. En pratique la fonction de valeur est modélisée par un réseau de neurones en vertu du théorème d'approximation universel [1]. Ce dernier stipule que toute fonction continue sur un espace compact peut être approchée par un réseau de neurones à une couche et un nombre arbitraire de neurones (Un autre théorème mentionne cette propriété avec des réseaux à plusieurs couches profondes et un nombre de neurones fini par couche). Le réseau profitera des conditions des différentes transitions générées par les interactions entre l'environnement et l'agent pour actualiser ses paramètres pour s'adapter au mieux. Nous détaillons le processus d'apprentissage dans la section 4.2.2

4.1.5 Le biais de confirmation : arbitrage exploration/exploitation

Le problème majeur auquel est confronté l'agent lors de son apprentissage est le biais de confirmation. En suivant la procédure itérative de la figure 4.2, l'agent risque de focaliser ses choix uniquement sur les meilleures actions **parmi celles qu'il est déjà en train d'exploiter**. En effet, seules les actions choisies

1. Il existe une multitude de théorème d'approximation

sont réevaluées pendant la phase d'évaluation de l'algorithme. Sachant qu'elles ont été choisies de sorte à améliorer la fonction de valeurs lors la phase d'amélioration (*Policy Improvement*) elles vont être davantage appréciées que celles qui n'ont pas été choisies. Ainsi à l'étape d'après, l'agent a plus de chance de les choisir etc. Le risque est que l'agent n'explore plus les actions qui dès le départ, n'étaient pas optimales dans un contexte. En effet, ces dernières pourraient lui procurer plus d'utilité que celles qu'il exploitent le plus dans d'autres contextes. C'est pourquoi, comme nous le verrons par la suite, il est courant d'introduire une contrainte d'exploration sur les décisions de l'agent afin que sa tâche d'optimisation ne se résume pas à un sous ensemble de l'espace des actions trop restreint.

4.2 Application du reinforcement learning à la gestion ALM

Comme nous l'avons vu dans la section 4.1.2 un modèle de reinforcement learning peut se résumer au quadruplet suivant :

- Espaces d'actions (Agent)
- Espaces d'états (Environnement)
- Matrice de Transition (Environnement)
- Fonction de récompense

4.2.1 Environnement

L'environnement qui modélise à la fois l'espace d'états et la matrice de transition est le modèle ALM décrit dans le chapitre 3.1.2.

Espace d'Etats

Notre espace d'états est caractérisé à chaque pas de temps par des transformations des variables suivantes :

Actif	Passif	Actif-passif
Allocation du portefeuille (en %)	Taux Minimum Garanti moyen	SCR total (normalisé par la PM)
Duration	Duration	SCR marché (normalisé par la PM)
Taux de marge		
Taux de richesse de l'actif ($\frac{VB}{VNC}$)		
Taux de produits financiers		

L'ensemble de ces variables est calculé selon les modalités décrites dans le chapitre précédent. D'un point de vue modélisation, cela signifie que l'agent reçoit une variable d'état de dimension 12, dont tous les scalaires prennent des valeurs continues. Dans ces conditions nous devons avoir recours à l'utilisation de réseaux de neurones (cf. section *deep reinforcement learning* 4.1.4) reinforcement learning dont l'architecture dépendra de l'espace d'action considéré dans notre étude.

Matrice de transition

La matrice de transition est une loi de probabilité sur l'espace des états. Dans notre modèle, la loi de transition d'un état à un autre est définie par les scénarios stochastiques dont les caractéristiques sont détaillées dans la figure 3.3.6 du chapitre 3.1.2

Fonction de récompense

Le choix de la métrique est stratégique. il doit à la fois répondre aux objectifs fixés mais aussi prendre en compte les différentes évolutions de l'actif et du passif afin que l'agent prenne ses décisions en adéquations avec ces dernières. Dans notre étude, nous adoptons le point de vue des actionnaires. Le *reward* est lié au bénéfice généré par le portefeuille pour ce dernier. Nous choisissons d'optimiser la métrique de *Return on Equity* (ROE) au delà du taux sans risque et du coût en capital. Il est calculé comme suit :

$$\rho_{t,i} = \frac{M_{t,i}}{K_{t,i}} - \kappa - \theta_{t,i} \tag{4.10}$$

où :

- $\rho_{t,i}$ est le reward à la date t et du scénario i
- $M_{t,i}$ est la marge nette de l'assureur à la date t du scénario i.
- $K_{t,i}$ est le capital immobilisé par l'assureur à la date t du scénario i. Il s'agit du produit entre le ratio S2 cible et le SCR
- κ est le coût du capital exigé par l'actionnaire.
- $\theta_{t,i}$ est le taux sans risque à la date t du scénario i. On utilise le taux OAT 10 ans en tant que taux sans risque.

La marge nette est une mesure de performance qui prend en compte les resultats financiers et techniques de la compagnie. L'avantage de la normaliser par le capital immobilisé de l'actionnaire qui est calculé en fonction du niveau de SCR est de prendre en compte l'exposition au risque des choix de composition de portefeuille.

4.2.2 Agent

L'espace d'états est synthétisé par les variables continues. Les actions de l'agent caractérisent un changement d'allocation dans chacune des classes d'actifs et mentionne une duration cible. Autrement dit, il s'agit d'un vecteur $(\Delta Action, \Delta Immobilier, \Delta Obligations, Duration\ cible)$. On distingue 2 cas de figure :

1. Espaces d'actions fini.
2. Espaces d'actions continu.

En fonction des caractéristiques de l'espace d'actions nous avons étudié deux modèles qui présentent des hypothèses et méthodes d'apprentissages différentes.

Espace d'actions fini : Deep Q network (DQN)

Dans un premier temps nous nous concentrerons sur l'étude de la stratégie d'un agent dans un espace discret. Kanervisto, Scheller, et Hautamäki ont montré dans leur étude [5] qu'un espace d'actions dit multi-discret (soit une discrétisation selon plusieurs axes différents) peut certes conduire à des performances plus faibles qu'en espace continu, mais il assure pourtant une capacité d'apprentissage de l'agent plus efficace en étant notamment moins gourmand en exploration et plus rapide pour converger. Dans la même logique, ils préconisent une réduction du nombre d'actions possibles, même si un nombre plus élevé d'actions pourrait donner de meilleures performances sur les choix à réaliser en situations extrêmes. Le deep Q network met en scène un agent en mesure de prendre un nombre fini d'actions. La fonction

de valeur Q_π est modélisée par un reseau de neurones qui prend en entrée un vecteur d'état dont la dimension varie en fonction du nombre de variable économiques prises en compte (ici 12) et retourne N vecteurs d'allocations de dimension 4, avec N, le nombre d'actions permise par l'agent. Dans notre cas $N = 25$ le réseau de neurones. Le choix d'une allocation parmi les 25 retournées par le réseau de neurones dépend de la phase exploration-exploitation de ce dernier :

1. Phase d'exploitation : l'agent selectionne l'allocation qui maximise la fonction Q_π,
2. Phase d'exploration : l'agent choisit une allocation aléatoirement parmi les 25 permises.

Action id	0	1	2	3	4	5	6	7	8	9	10	11	12
Δ Actions	0%	8%	-8%	0%	0%	0%	0%	0%	0%	0%	2%	-2%	2 %
Δ Immobilier	0%	0%	0%	8%	-8%	0%	0%	0%	0%	0%	0%	0%	0%
Δ obligation	0%	0%	0%	0%	0%	8%	-8%	8%	-8%	8%	-8%	0%	0%
Duration cible	0	0	0	0	0	0	0	0.5	0.5	-0.5	-0.5	0	0

Action id	13	14	15	16	17	18	19	20	21	22	23	24
Δ Actions	0%	0%	0%	0%	0%	0%	0%	0%	2.67%	-2.67%	0.67%	-0.67%
Δ Immobilier	2%	-2%	0%	0%	0%	0%	0%	0%	2.67%	-2.67%	0.67%	-0.67%
Δ obligation	0%	0%	2%	-2%	2%	-2%	2%	-2%	2.67%	-2.67%	0.67%	-0.67%
Duration cible	0	0	0	0	0.5	0.5	-0.5	-0.5	0	0	0	0

Deep Q-learning Nous rappelons que l'apprentissage de l'agent repose sur l'actualisation et l'optimisation de la fonction de valeurs décrite dans la section 4.1.3. Dans le cas paramétrique, l'apprentissage se fait via l'actualisation des paramètres du réseau de neurones. En effet, lors de chaque changement d'allocation décidé par l'agent, un vecteur de transition

$$(S_t, A_t, R_t, S_{t+1}) = (s, a, r, s') \tag{4.11}$$

est stocké en mémoire. En vertu de l'équation de Bellman optimale 4.7, et en exploitant un sous échantillon de vecteurs de transition 4.11 stocké en mémoire, une *target*

$$y_t = r + \arg\max_a Q_\pi(s', a|\theta_t) \tag{4.12}$$

est générée. En fin d'épisode, les paramètres du réseau de neurones sont actualisés après régression sur les *targets optimisées*.

Soit n, la taille du sous-échantillon considéré :

$$Soit,\ y_t \in \mathbb{R}^n,\ \theta_{t+1} = \arg\min_\theta {}_t\mathbb{E}_\pi \left[(Q_\pi(s, a|\theta_t) - y_t)^2 \,|(s, a, r, s') \right] \tag{4.13}$$

Le processus d'apprentissage complet est décrit dans 4.3

:

Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

FIGURE 4.3 – Algorithme Deep Q Network (DQN)

Experience Replay L'apprentissage du réseau de neurones consiste en une régression sur les targets générées suite aux prises de décisions. Pour que celle-ci soit mise en oeuvre de manière appropriée, il est nécessaire que les targets $(y_t^i),\ i \in \{1, ..., n\}$ soient indépendantes, ce qui n'est vraissemblement pas le cas dans notre étude. En effet, chaque target est le fruit d'un changement d'allocation suite aux prises de décisions successives du modèle ALM. Bien que les scénarios introduisent un bruit qui réduit la corrélation entre les *targets*, elles n'en demeurent pas moins dépendantes par la propriété de Markov [2]. C'est pourquoi Mnih [7], suggère d'introduire un tampon (i.e buffer) qui enregistre les targets au fil du temps. Lors de l'apprentissage, les targets utilisées pour l'optimisation sont tirées uniformément du tampon ce qui limite les relations de corrélation et causalité qui existe entre chacune d'elle. Afin de tirer profit au mieux de l'*experience replay* des recherches [8] suggèrent d'accorder un poids plus important aux observations ayant obtenues de faibles résultats *reward*. L'implémentation de cette stratégie de *boosting* est censée permettre de renforcer l'apprentissage de l'agent sur les mauvaises décisions qu'il a mal effectué plutôt que sur les bonnes.

Deep Q Learning :Implémentation

En pratique notre réseau de neurones prend en input une variable d'état $s \in \mathcal{S}$ et produit un score pour chaque décisions présente dans la plage d'allocations test (ici 25). Dans notre cas le réseau Q_π est une matrice de taille $(13, 25)$ dont les poids varient en fonction de l'architecture totale du réseau. En effet, notre réseau prend en entrée un tenseur de taille $dim(s) = 13$ et retourne un vecteur v de dimension 25, dont chaque scalaire est un score pour chaque allocation. Soit $i \in [\![0 : 24]\!]$, on a $v_i = Q_\pi(s, a_i)$. En phase d'exploitation l'agent choisit l'action a_i qui apporte le meilleure score (i.e $a_* = \arg\max_{i[\![0:24]\!]}$

2. On rappelle que notre espace d'états vérifie la propriété de Markov et assume par conséquent une dépendance directe de S_{t+1} avec le couple (S_t, A_t)

tandis que durant l'exploration il choisit une action aléatoirement afin d'éviter le biais de confirmation. Les phases d'exploitation et exploration sont modélisées par une variable de bernouilli tiré lors de chaque prise de décision. Plus le paramètre de la loi est proche de 1 plus l'agent est susceptible d'explorer son environnement sans réellement appliquer de stratégie. A l'inverse plus le paramètre est proche de 0, plus l'agent est enclin à suivre sa stratégie en prenant l'action qui lui procure le meilleure score quitte à oublier d'autres actions oubliées dans la stratégies de l'agent. Ce paramètre est crucial dans l'apprentissage de l'agent, on peut imaginer le faire varier au cours du temps.

Espace d'actions continu : Deep deterministic Gradient Policy (DDPG)

Dans le cas d'un espace d'actions continu, il est impossible d'avoir recours au Deep Q-Learning. En effet l'optimisation 4.9 nécessite de parcourir l'espace des actions entier pour un état donné ce qui est impossible en pratique. Pour pallier à cette limite, le Deep Deterministic Policy Gradient introduit deux agents :

1. **Un acteur** : L'acteur prend des décisions d'allocation en continu. Chaque action est l'image d'une fonction paramétrée $a = \pi(s|\theta^a),\ \forall s \in \mathcal{S}$.
2. **Un critique** : Le critique (i.e fonction de valeur) évalue chaque action prise par l'acteur, $Q_\pi(s, \pi(s|\theta^a)|\theta^c),\ \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$.

Utiliser une fonction paramétrique déterministe pour définir les décisions d'allocation de l'agent permet de transformer notre problème d'optimisation sur l'espace d'actions entier 4.9 en l'optimisation du problème suivant :

$$\mathcal{J} : \theta \longrightarrow V_{\pi_\theta}(s),\ \forall s \in \mathcal{S} \tag{4.14}$$

où $\theta = (\theta^a, \theta^s)$. Contrairement à l'algorithme DQN où l'optimisation de la stratégie de l'agent repose sur le choix de chaque action pour un état donné, dans le paradigme du DDPG, la stratégie optimale s'obtient progressivement via l'application de la descente de gradient sur $\mathcal{J}(\theta)$. A priori, améliorer éfficacement la politique dans son ensemble n'est pas tache aisée. En effet, la performance de l'agent dépend à la fois du choix de l'action mais aussi de la loi sur l'espace d'états. Or, tous les deux sont influencés par la politique en vigueur. Pour un état donné il est facile d'étudier l'impact d'un changement de politique sur la fonction de valeur. En effet, à état fixé on peut isoler l'effet d'un changement d'action par le reward obtenu. En revanche, notre problème consiste à optimiser la fonction de valeur en fonction de π sur l'espace d'états. Or l'agent n'a aucun contrôle sur la distribution associée à l'espace d'états qui est une donnée de l'environnement. Heureusement, le théorème 4.2.1 (*Policy Gradient theorem*), permet d'exhiber une relation directe entre le gradient de la fonction de valeur et π_θ. En particulier, ce résultat nous permet d'exhiber une fonction objective, dont on connait le gradient sur laquelle on peut appliquer une descente de gradient.

Théoreme 4.2.1 (Policy Gradient Theorem)

$$\mathcal{J}(\theta) = \sum_{s\in\mathcal{S}} d_{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \pi(a|s;\theta) Q_\pi(s,a) \propto \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} \pi(a|s;\theta) Q_\pi(s,a) \tag{4.15}$$

$$\nabla\mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta}[\nabla \ln \pi(a|s,\theta) Q_\pi(s,a)] \tag{4.16}$$

L'algorithme d'apprentissage est détaillé dans la figure 4.4. Les auteurs de l'algorithme [6] essaient de tirer le meilleur des algorithmes de descente de gradient sur la politique et du DQN présenté dans la section précédente. En effet on remarque deux aspects identiques entre DDPG et DQN :

Algorithm 1 DDPG algorithm

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer R
for episode = 1, M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for t = 1, T **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in R
    Sample a random minibatch of N transitions (s_t, a_t, r_t, s_{t+1}) from R
    Set $y_t = r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_t (y_t - Q(s_t, a_t|\theta^Q))^2$
    Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_t \nabla_a Q(s, a|\theta^Q)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_t}$$

    Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

  end for
end for

FIGURE 4.4 – Algorithme Deep Deterministic Policy Gradient (DDPG)

1. DDPG introduit un bruit aléatoire qui modifie l'action de l'agent ce qui contraint l'agent d'explorer son environnement comme c'est le cas dans DQN.
2. L'optimisation des poids de $Q_{\pi_{\theta^c}}$ a lieu suite à une regression sur les *targets* définies dans la partie précédente (cf.4.3)

L'actualisation des paramètres de l'acteur π_{θ^a} a lieu après ceux de Q via l'application de la descente de gradient sur $\mathcal{J}$ propres aux méthodes d'amélioration de la politique par descente de gradient.

Target Q Network : Lors de l'apprentissage de la fonction valeur-action $Q_\pi(\cdot|\theta^c)$, le DDPG introduit des copies $Q_\pi(\cdot|\theta^{c\prime})$ de la fonction de valeur appelées *Target Q Network*. Contrairement à 4.12 où la fonction $Q_\pi(\cdot|\theta^c)$ et $Q_\pi(\cdot|\theta^{c\prime})$ sont confondues, DDPG effectue une distinction entre le réseau principal à optimiser $Q_\pi(\cdot|\theta^c)$ (i.e l'agent) et celui qui génère les targets $Q_\pi(\cdot|\theta^{c\prime})$. D'après [7] cette distinction permet de garantir une meilleure stabilité dans l'apprentissage de l'agent. En effet, l'actualisation des poids du target network est pondérée par ceux du réseau principal [3], ce qui prévient l'agent d'une variation trop brutale des paramètres.

3. Soit $\tau \in 0, 1$, on a $\forall t,\ \theta^c_{t+1}{}' = \tau\theta^c_t + (1-\tau)\theta^{c\prime}_t$

Implémentation pratique du DDPG

Le DDPG présente une modélisation plus complexe que le DQN. Nous pouvons caractériser cette complexité par le nombre d'hyperparamètre à gérer.

Tout d'abord, étant donné que l'acteur et le critique sont tous deux modélisés par des reseaux de neurones les combinaisons d'hyperparamètres à gérer sont au moins deux fois plus importante comparé à celles nécessaires à la calibration du DQN.

Le DDPG modélise l'exploration de l'agent par une perturbation bruit autour du choix d'allocations de l'acteur. Contrairement au papier original DDPG [6] qui mentionne l'utilisation d'un processus d'Orstein Ulhenbeck pour modéliser le bruit, nous avons choisi une loi gaussienne centrée. [4]

Enfin, une difficulté supplémentaire s'ajoute au DQN : contraindre l'espace d'arrivé de sorte à ce qu'il corresponde à la plage d'allocation de tests. Afin de permettre une comparaison avec le DQN dont l'espace d'action est discret avec $\pm 8\%$ pour bornes, nous souhaitons contraindre nos changements d'allocations pour chaque classe d'actif à valeurs continues dans un hypercube ayant les mêmes bornes. Afin d'y parvenir nous avons fait le choix d'utiliser comme fonction d'activation en couche d'output la tangente hyperbolique avec une initialisation appropriée dont les poids sont très proche de 0 afin d'éviter les problèmes de dissolution du gradient (*gradient vanishing*).

4.2.3 Architecture finale du modèle

En définitive, notre modèle DeepALM n'est rien d'autre que le modèle ALM décrit dans le chapitre 4 sur lequel est "branché" un agent qui applique une stratégie d'allocation dynamique. Ce dernier profite des interactions avec l'environnement décrit par les scénarios pour adapter ses choix d'allocation. Néanmoins nous avons vu qu'en fonction de la topologie utilisé pour modéliser les espaces d'états et d'actions, pour un modèle ALM fixé, les méthodes d'apprentissage et donc les modèles DeepALM diffèrent sensiblement.

4.2.4 Ecoulement du modèle

On peut trouver ci-joint l'évolution des variables économiques au fil de la projection sur 40 pas de temps. Il s'agit de montant moyens.(cf. figure 4.6)

4. L'impact d'un changement de paramètre sur une gaussienne est plus facile à appréhender que dans le cadre d'un processus Orstein Ulhenbeck ce qui nous laisse plus de fléxibilité dans le choix des paramètres.

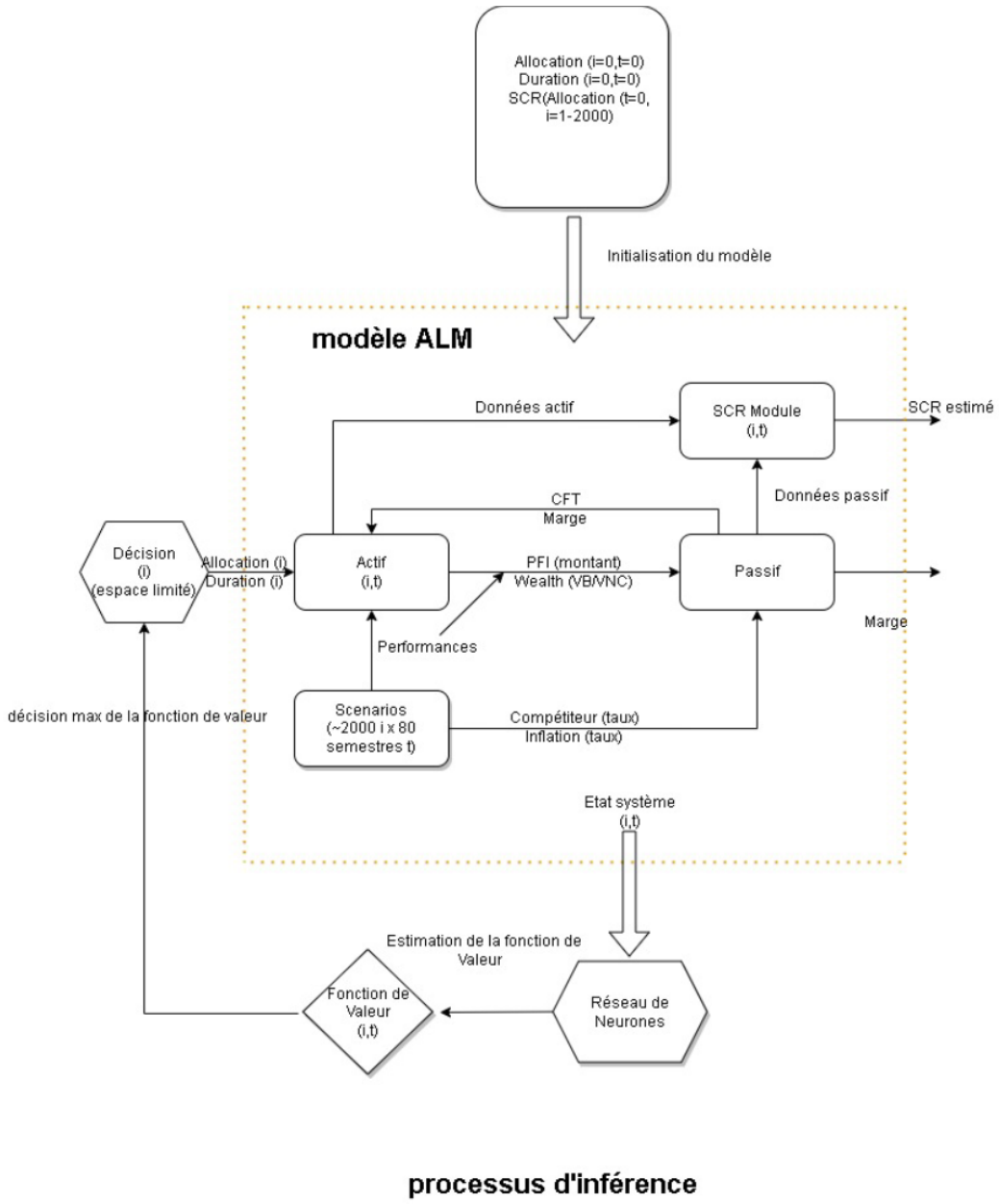

FIGURE 4.5 – DeepALM - Processus d'inférence

FIGURE 4.6 – Ecoulement des variable modèle pour un agent qui reste passif

5 Analyse des resultats

Dans cette partie nous confrontons les résultats obtenus des algorithmes mentionnés dans le chapitre 4. Pour ce faire nous allons adopter le protocole suivant :

1. Définir plusieurs stratégies ”naïves” comme points de comparaison.
2. Parmi les stratégies naïves séléctionner des stratégies de références pour analyser la performance des modèles DQN et DDPG entrainés.
3. Etudier les performances de la meilleure stratégie obtenue suite à la comparaison de l'étape précédente.
4. Interpréter des choix de la meilleure stratégie.

Notre modèle dispose de 2000 scénarios monde réel de 40 pas de temps [1] ce qui représente 80 000 itérations au total.
Les 1800 premiers scénarios servent de base d'apprentissage pour l'agent tandis que les 200 scénarios restants (soit 10% du jeu de donné total) [2] sont utilisés pour l'évaluation des stratégies d'allocation mises en place par ce dernier.

5.1 Choix de la stratégie de référence

Définitions des stratégies naïves

Au cours de notre étude, aucun des modèles testés n'a permis de mettre en valeur une stratégie capable de s'adapter, en moyenne, à l'ensemble des environnements économiques simulés et ainsi atteindre le niveau de performance attendu par les actionnaires (i.e le cout en capital cumulé au taux sans risque cf. métrique d'optimisation 4.10). Toutefois, nous ne pouvons disqualifier l'ensemble de notre étude en partant de ce constat. En effet, il est important de comparer les résultats obtenus par l'agent avec certaines stratégies naïves afin de mettre en exergue la pertinence des choix d'allocation de l'agent. Et ce même s'ils sont sous optimaux. A cet effet nous définissons 4 stratégies ”naïves” :

- **Allocation passive :** L'agent reste neutre et n'effectue aucun choix d'allocation peu importe la conjoncture économique. Autrement dit, l'agent laisse le modèle ALM s'écouler ”naturellement”.
- **Allocation fixe :** L'agent met tout en oeuvre pour conserver l'allocation initiale. En particulier, ce dernier s'affranchit des plages tests (discrètes et continues) définies dans les sections 4.2.2.
- **Allocation cible :** Cette stratégie est celle qui est mise en oeuvre au sein d'AXA. Après une étude ALM approfondie, les allocataires définissent une allocation cible que la compagnie se doit d'essayer

1. Initialement, projection a lieu sur 60 pas de temps. Cependant les 20 dernières années de projections sont sujettes à des instabilités numériques qui biaisent l'apprentissage de l'agent. restreindre
2. L'échantillon de test est moins fourni que les conventions habituelles. ceci est dû au manque de scénarios auquel nous faisons fasse qui ne permet pas à l'agent d'apprendre de manière optimale.

de suivre à chaque pas de temps tout en respectant les plages d'allocations tests. Dans notre cas l'allocation cible est l'allocation initiale (cf.3.8)

— **Allocation aléatoire :** L'agent effectue des changements d'allocation aléatoires en respectant les plages d'allocations tests compte tenu des hypothèses du modèle utilisé. Dans le cas discret (DQN), les changements d'allocations sont restreints aux vingt-cinq actions définies en **??** tandis que dans le cas continu (DDPG), les changements d'allocations sont restreints à l'hypercube $[-0.8, +0.8]^3 \times [-1, +1]$ (les trois premiers intervalles definissent les changements d'allocation de chaque actif et la plage $[-1, 1]$ correspond à la duration cible des actifs).

L'ensemble de ces stratégies "naïves" nous servirons de points de comparaison afin d'analyser plus précisement la performance de l'agent. En effet, obtenir de meilleurs rendements que la stratégie mise en place par AXA (allocation cible) s'avère être une bonne performance. Au contraire, générer moins de profits que les stratégies d'allocation aléatoire et/ou passive révèlent un problème dans l'apprentissage de notre agent.

Bien que les stratégies naïve d'allocation cible et aléaloite soient compatible à une implémentation dans le cas discret (DQN) comme continu (DDPG), nous décidons de ne les modéliser qu'en continu. En effet, la restriction de l'espace d'actions de l'agent DQN à 25 actions peut déja être considéré comme une initialisation assez pertinente de notre algorithme. En ce sens, modéliser la stratégie d'allocation aléatoire dans un espace déjà prédéfini parait moins pertinent que dans le cas continu où aucune restriction *a priori* n'influence la stratégie aléatoire. Par ailleurs, ce choix de modélisation est d'autant plus pertinent qu'il se rapproche des conditions du monde réel.

Performances des stratégies naïves et séléction d'un benchmark

Dans la figure 5.4 sont décrites les principales caractéristiques des stratégies de référence. Sans surprise les stratégies d'allocation fixe et cible surperforment les stratégies aléatoire et passive. Par ailleurs elles affichent des niveaux de performances presque identiques. Cette proximité s'explique par l'objectif commun qu'elles cherchent à atteindre : maintenir l'allocation initiale constante en toute circonstance [3].Malgré cette proximité, l'allocation cible présente deux avantages sur l'allocation fixe : i) Elle affiche un meilleur ratio de sharpe ii) elle se rapproche de la stratégie d'AXA sur les marchés et donc permet d'étudier chaque stratégie à partir d'un benchmark réaliste. En vertu de tous les points mentionnés nous définissons séléctions deux stratégies de référence :

1. La stratégie d'allocation passive : Dans un premier temps cette stratégie nous permet de trier entre les modèles qui ont appris sur les marchés et ceux qui n'ont pas appris. En effet, si un agent n'est pas en mesure de dépasser cette stratégie, cela implique qu'il vaut mieux se retirer des marchés (i.e ne rien faire) plutôt que d'y participer.

2. La stratégie d'allocation cible : proche du comportement d'AXA sur les marchés, cette stratégie nous permet d'analyser avec precision l'impact des décisions prises sur les rendement de l'entreprise. En particulier elle permet d'identifier les stratégies performantes.

3. En revanche ce qui les distingue sont les contraintes qui reposent sur chacune d'elle pour parvenir à conserver l'allocation initiale. L'allocation cible est contrainte par les hypothèses de l'espace d'actions continu tandis que l'allocation fixe ne dispose d'aucune contrainte.

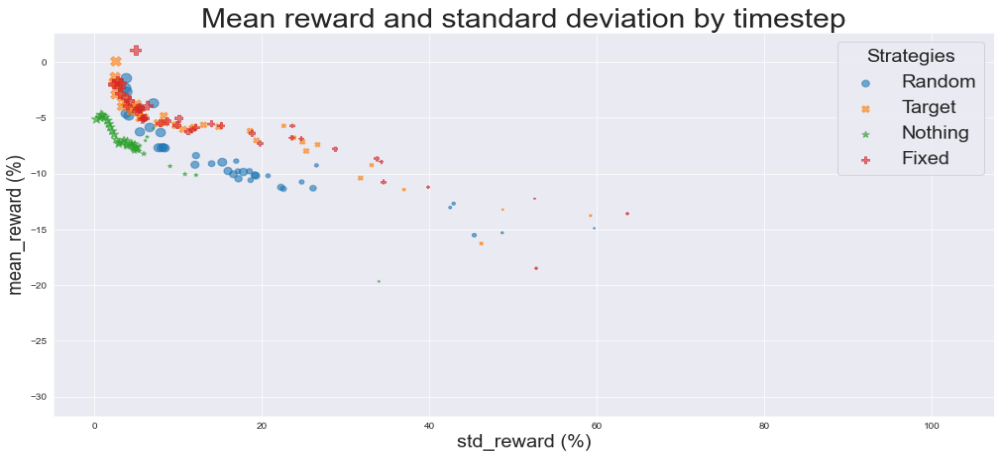

FIGURE 5.1 – Couple (rendement,risque) moyen par année de projection

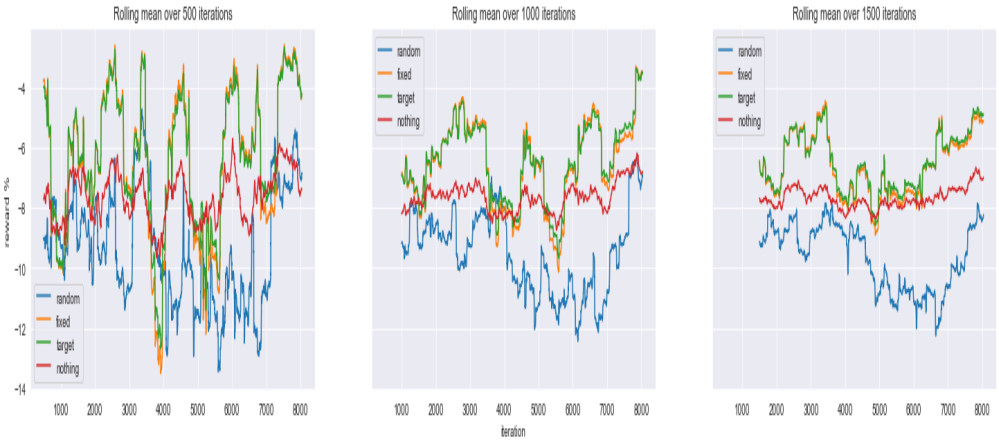

FIGURE 5.2 – Etudes de la performance des stratégies en dynamique

	sharpe_ratio	mean	std	min	25%	50%	75%	max
fixed	-26,01%	-6,35%	24,40%	-596,91%	-5,57%	-3,49%	-1,03%	44,83%
target	-28,56%	-6,27%	21,94%	-465,42%	-5,47%	-3,39%	-0,95%	21,50%
random	-33,74%	-9,35%	27,70%	-1016,49%	-9,06%	-5,71%	-2,80%	62,83%
nothing	-92,36%	-7,65%	8,29%	-173,94%	-9,51%	-5,81%	-4,25%	6,33%

FIGURE 5.3 – Indicateurs de performance des stratégies de référence

FIGURE 5.4 – Caractéristiques des stratégies de référence

5.2 Etude de la performance des modèles implémentés

5.2.1 Propriétés des modèles testés

Dans notre étude, nous avons entrainé plusieurs agents avec des caractéristiques distinctes.

1. **Agents DQN (cas discret) :** Nous avons fait varier principalement deux types de paramètres :
   — Le taux d'exploration/exploitation prenant les valeurs 10, 20 et 30%. Durant nos études préliminaires nous avons trouvé que 20% est le taux qui affiche les meilleures performances moyennes toute architecture confondues.
   — L'architecture du réseau de neurone, en faisant varier le nombre de couches et de neurones par couches. Nous avons testé les architectures suivantes (feed forward) : 64, 128,32x32,64x128 et 16x32, avec un dropout de 20% entre chaque couche cachée.
2. **Agent DDPG (cas continu) :**
   — Distribution du bruit : nous avons posé une loi centrée avec d'écart type 3%. Ainsi, 68% des déviations induites par le bruit ont une amplitude symétrique de 3% et 97.5% ont une amplitude de 6%.
   — Architecture du réseau de neurones : Nous avons testé les architectures 8,8x8,64,64x64,32x32,64x128 avec une Batchnormalization entre les couches cachées ainsi qu'une standardization des variables d'états en amont dont les mesures ont des ordres de grandeur différents. Par exemple
   — Stratégie "dopée" : durant la phase d'apprentissage du réseau 8_DDPG, on remarque une dégradation des performances moyennes sur 1000 scenarios à partir de la 40000 itérations. Cette dégradation survient après une phase de stabilité pouvant provenir du bruit posé. Nous avons donc décidé d'appliquer les poids du reseau de neuronnes au 1075 scénario pour voir s'il constituait une amélioration. Ce qui semble être le cas d'après la table 5.3.

5.2.2 Evaluation des modèles et identification du modèle le plus performant

Les performances des agents testés sont présentées dans la table 5.5 regroupe les caracteristiques de bases concernant les architectures testées. Les valeurs du tableau sont des surperformances relatives à la stratégie d'allocation passive. On remarque :

— Modèle seules les stratégies 32x32_DDPG et 8_doped_DDPG surperforment les stratégies passive et aléatoire parmi les modèles à actions continues.
— Tous les modèles DQN surperforment la stratégie de référence (i.e passive) exceptée 64_DQN.

Si les indicateurs de performances permettent de mettre clairement en exergue les

	sharpe_ratio	mean	std	min	25%	50%	75%	max
8_doped_DDPG	7,87%	1,44%	18,35%	-326,62%	-0,82%	1,84%	6,38%	145,35%
64x128_DQN	7,42%	1,89%	25,47%	-513,63%	-0,24%	2,52%	7,01%	155,26%
target	6,44%	1,39%	21,54%	-372,97%	0,00%	2,27%	7,62%	96,25%
fixed	5,47%	1,31%	23,90%	-434,51%	0,00%	2,23%	7,51%	116,90%
32x32_DQN	4,60%	1,30%	28,14%	-839,98%	-0,45%	3,29%	7,79%	172,31%
32x32_DDPG	1,68%	0,41%	24,68%	-436,85%	-1,42%	2,39%	8,18%	133,02%
16x32_DQN	1,18%	0,38%	32,03%	-871,26%	-0,76%	1,05%	6,11%	177,62%
128_DQN	1,12%	0,41%	36,45%	-1681,68%	-0,66%	2,98%	7,75%	176,72%
64_DQN	-5,19%	-2,38%	45,80%	-857,15%	-0,89%	2,78%	7,72%	173,50%
random	-6,08%	-1,69%	27,85%	-985,23%	-1,72%	0,12%	3,37%	174,71%
64x64_DDPG	-9,07%	-2,54%	27,96%	-397,55%	-2,91%	0,96%	7,23%	175,73%
64_DDPG	-23,38%	-16,04%	68,62%	-1736,59%	-5,26%	-2,04%	2,67%	163,18%
8x8_DDPG	-25,49%	-9,31%	36,54%	-714,55%	-4,87%	-1,90%	1,49%	171,08%
8_DDPG	-45,63%	-5,82%	12,75%	-206,32%	-11,63%	-5,66%	-0,73%	149,91%

FIGURE 5.5 – Performances des modèles testés, classés par ordre ordre décroissant du ratio de sharpe.

A partir du tableau 5.5 trois agents se distinguent : 6_doped_DDPG, 64x128_DQN et 32x32_DQN. Afin d'identifier l'agent atteignant la meilleure performance, nous allons comparer l'éfficacité de leur stratégie

avec celle de l'allocation cible sur un scénario moyen en trois temps. Tout d'abord nous allons étudier les décisions prises par les agents, puis voir la rentabilité économique et enfin étudier l'exposition aux risques.

Etude des décisions des agents

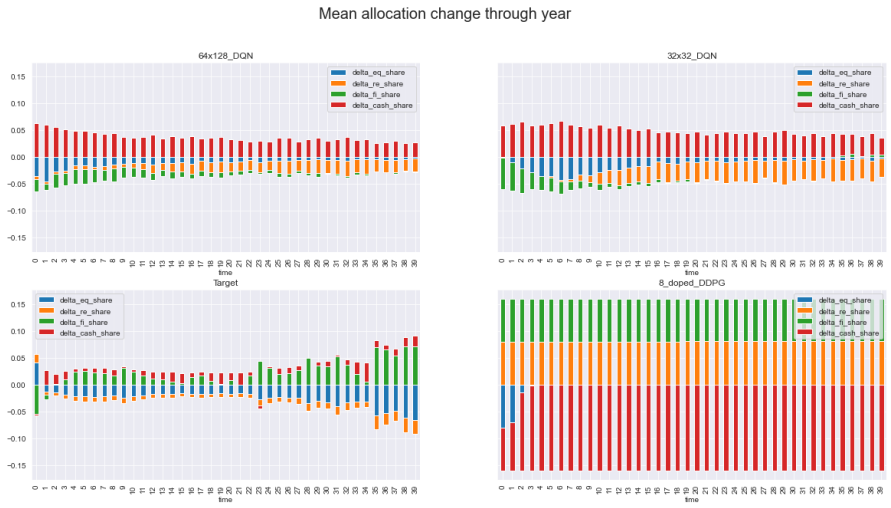

FIGURE 5.6 – Variations moyennes des allocations d'actifs par année de projection

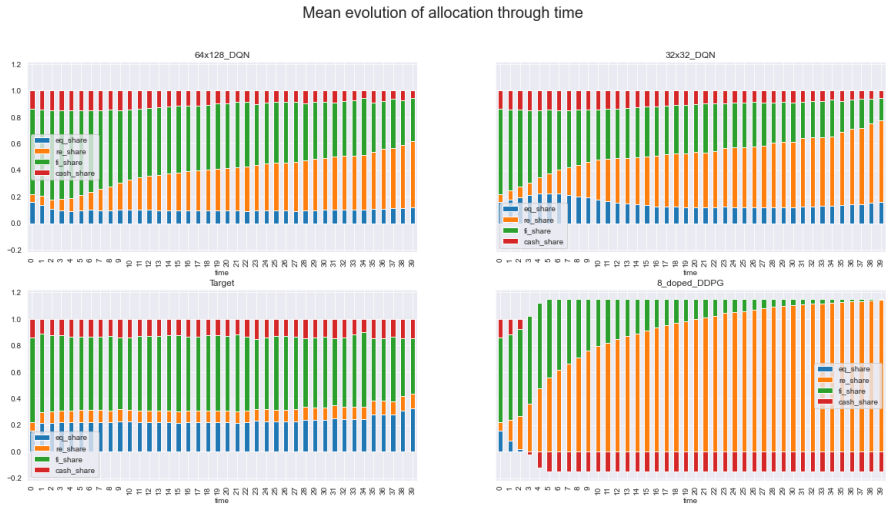

FIGURE 5.7 – Allocation moyenne par pas de temps

On remarque sur les figures 5.6 et 5.7 que les agents DQN n'effectuent que des décaissements dans les trois classes d'actifs. Au fil des projections la désallocation concerne principalement les actifs immobiliers,

bien que leur part relative continue de croitre en moyenne au fil des années de projection au profit des obligations et des actions dont l'allocation se stabilisent au cours du temps. La hausse des parts en immobiliers malgré le choix de désallocation de l'agent peut s'expliquer par un phénomène de compensation en raison du décaissement obligations qui arrivent à expiration.

Contrairement aux agents DQN, une fois après avoir désalloué totalement les actions de son portefeuille d'actifs, l'agent 8_doped_DDPG réalloue continuellement 8% de sa revserve en monnaie en immobilier et obligations de duration 1 an. [4]

Afin de maintenir une allocation constante, la stratégie allocation cible (ou target), doit compenser le décaissement des obligations arrivées à maturité en réinvestissant une partie de ses parts immobilières et actions sur de nouvelles obligations.

Rentabilité économique consécutive au choix de l'agent

Dans cette partie nous allons voir l'impact des changements d'allocation décrits précédemments sur la rentabilité économique de la compagnie. Pour ce faire nous allons étudier les trois ratios suivants :

— Le taux de produits financiers : qui correspond à l'ensemble des profits réalisés suite à la possession ou à l'achat-vente (PMVR, Plus-Moins value réalisées) des actifs de la compagnie.

$$tx\_pfi = \frac{rendements\_financiers + PMVR}{VNC}$$

— Le taux de marge nette : Il s'agit de la somme des produits financiers des résultats techniques et administratifs. Les résultats techniques sont les résultats non financiers liés à la souscription des contrats. Les resultats administratifs prennent en compte les dépenses de gestion liés au traitement des contrats dossiers.

$$\frac{Resultats\_techniques + Produits\_financiers + Resultats\_administratifs}{VNC}$$

— le taux de richesse : $\frac{Valeur\_boursi\&re}{VNC}$

D'après la figure 5.8,il apparait de manière significative que l'agent 8_doped_DDPG obtient de meilleurs résultats que les autres. Cette performance économique est imputée aux resultats financiers qui s'expliquent par l'investissement accru de l'agent sur en obligations et immobiliers comparé aux autres agents.

Les autres agents semblent avoir des niveaux de performances assez similaires à ceux obtenus par la stratégie d'allocation cible.

Exposition au risque : solvabilité

Pour rendre compte de la performance réelle des stratégies étudiées, il est important de voir l'exposition au risque consécutive des choix d'allocation effectués par les agents. Cette étude nous permet entre autre de relativiser les performances de l'agent 8_doped_ddpg. Ce dernier possède un gap de duration au maximum égal à -6 comparé aux autres stratégies pour qui il s'agit de la valeur minimale. Ainsi, l'agent 8_dope_DDPG est sujet à une perte potentielle minimale de 6% en cas de baisse de taux de 1%. Par ailleurs, la figure 5.9 8_doped_DDPG posède un ratio de SCR significativement plus élevé que ses concurrents.

4. Le modèle continu (ddpg) effectue des changement d'allocation discrets aux bornes des plages d'allocations tests de l'hypercube défini en 4.2.2, tandis que les stratégies discrètes (DQN) font preuve d'une plus grande diversité. La focalisation de l'acteur ddpg sur uniquement deux actions peut être interprété de deux manières : i) un mauvais apprentissage de l'agent d'entrainement ii) la détection d'un biais de notre modèle. Nous tachons de répondre à ce problème dans la suite d enotre analyse

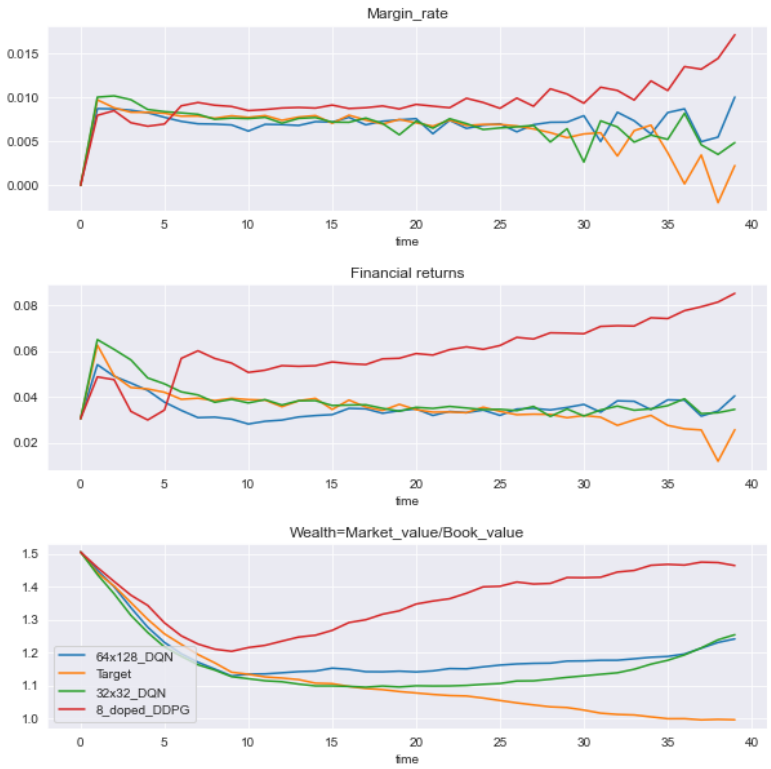

FIGURE 5.8 – Evolution moyenne de la rentabilité économique par année de projection

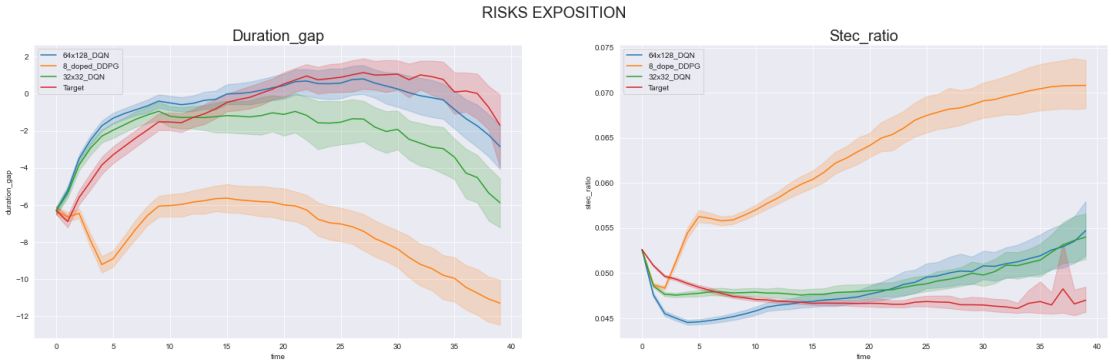

FIGURE 5.9 – distribution moyenne avec écart type de l'exposition au risque des stratégies

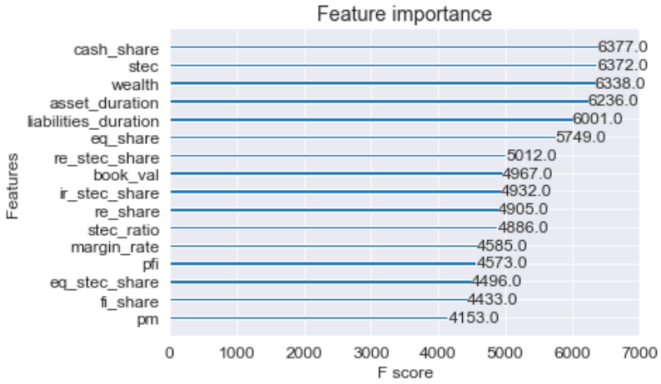

Figure 5.10 – Caption

Conclusion

L'exposition aux risques de 8_dope_DDPG nuancent les performances économiques affichées plus tôt. D'un point de vue prudentiel la compagnie ne peut pas suivre cette stratégie. 64x128_DQN semble atteindre le meilleur équilibre rendement-risque. En effet, il s'agit de la stratégie qui comble le mieux le gap de duration et qui presente un SCR minimal pour des niveaux de performances semblable à ses concurrents.

Interprétabilité

Parmi toutes les variables d'états nous avons chercher à identifier celles qui ont le plus d'impact sur les décisions de l'agent. pour ce faire nous avons appliqué un XGBoost classifier avec en input les variables d'état et en target le choix de l'action de l'agent 62x128_DQN. (Accuracy : 53%,f1-score : 59%, precision :68%,recall :53%) Nous obtenons les résultats de la figure 5.10 Ces résultats mettent en avant que l'agent est limité par les contraintes de cash liées à la simplification de notre modèle en ne modélisant que la classe d'actif correspondant au passif. De plus on remarque une volonté de l'agent de combler le gap de duration avec les scores associées aux duration de l'actif et du passif, ce qui correspond à ce que la direction des investissements effectuerait en temps réels. Par manque de temps nous n'avons pas pu otpimiser le XGBoost de sorte à ce qu'il permette d'obtenir des resultats plus robuste à partir desquels nous pourrions étudier des SHAP values afin de rentrer davantage dans les détails des motifs d'action de l'agent.

5.3 Limites et approfondissements

Si notre étude nous a permis de trouver la meilleure architecture parmis celles testées, elles comportent des limites qui méritent d'être traitées dans le cas d'une étude approfondie.

5.3.1 Améliorations techniques

Optimisation de code et puissance de calcul Le problème majeur auquel nous avons été confronté lors de notre étude est le manque de puissance de calcul. Cette contrainte a eu une incidence non négligeable sur la qualité de nos résultats. En effet, elle a limité le nombre de modèle testés. L'utilisation de GPU voire de TPU avec une parallélisation adéquate des calculs permettrait de tester plus de modèles avec des architectures plus robustes afin de parvenir à de meilleurs résultats. Cela nous permettrait entre autre le recours à une optimisation bayesienne.

Modélisation des plages d'allocations tests Les plages d'allocations tests représentent une hypothèse essentielle du modèle. Elles définissent le support d'actions de l'agent. Or dans les modèles implémentés elles ont été définies avec la même amplitude pour chaque classe d'actif. Dans une future étude On pourra penser à différencier les plages tests en fonction des caracteristiques spécifiques de chaque actifs telle que la contrainte de liquidité.

Autre hypothèse forte de notre agent DDPG, afin de restreindre l'output de notre réseau de neuronnes à la plage d'allocation test, nous avons défini la fonction d'activation suivante en couche de sortie $g(x) = 0.08 \cdot tanh(x)$. Or la fonction $tanh(x)$ a un impact non négligeable sur l'apprentissage de l'agent qui est sujet aux problèmes d'évanouissement de gradient. Dans le cas d'un développement du modèle on pourra à définir une fonction de reward qui comporte une pénalité adaptée hors des bornes de l'hypercube considéré afin d'orienter les actions de l'agent sans entraver son apprentissage. On peut aussi penser à introduire une pénalité inter-temporelle qui limite le nombre de changements d'allocations sur l'ensemble de la période étudiée.

5.3.2 Améliorations "métier"

Ajout de variables Nous utilisons dans notre algorithme un nombre limité de variables pour représenter l'état du système. Avec une puissance de calcul très importante et un générateur de scénario, on pourrait envisager d'alimenter l'agent avec l'intégralité de la connaissance de l'environnement. Sans aller jusque là, plusieurs variables d'état pourraient avoir un impact bénéfique si elles étaient ajoutées au modèle. Par exemple :

- Taux de réinvestissement – le taux OAT 10 que nous utilisons en tant que taux sans risque peut aussi être utilisé pour obtenir une indication sur les revenus futurs générés par un investissement en obligations.
- TMG moyen du passif – il donne une indication sur les produits financiers minimum au-delà duquel l'assureur commence à générer de la marge.

Interpretabilité L'application du machine learning à de nombreux domaines est conditionné par l'interprétabilité des résultats de ces derniers. Le monde financier ne fait pas exception à la règle.En juin 2020, l'ACPR (Autorité de contrôle prudentielle et de réglementation) a publié un rapport sur la gouvernance des algorithmes d'intelligence artificielle dans le secteur financier (Dupont et al. (2020)). Dans ce document elle définit un cadre d'évaluation et de gouvernance pour les modèles de Machine Learning qui s'appliquerait à une éventuelle mise en place du modèle présenté dans ce mémoire.

Les modèles implémentés doivent respecter des contraintes de **stabilité, performance et explicabilité** pour pouvoir être utilisés. La nécessité de transparence des algorithmes déployés a provoqué la

création d'un pan de recherche en deep reinforcement learning dont l'objectif est soit d'optimiser des algorithmes interpretables, soit trouver de nouvelles méthodes danalyse pour rendre compréhensible les décisions des algorithmes.

6 Conclusion

Dans ce mémoire nous avons tenté d'appliquer le paradigme de l'apprentissage par renforcment à l'allocation stratégique d'actifs sur les marchés. Pour la gestion du portefeuille en run-off, l'étude montre qu'un agent choisissant l'action ayant le score maximal prédit par réseau de neurones dense de dimension 64x128 est plus performant que le suivi d'une allocation stratégique cible. Cette surperformance est caractérisée par un rendement sur capital supérieur pour l'actionnaire en espérance avec un risque identique. Il ne s'agit que d'un début d'exploration des applications du deep learning à l'ALM étant donné que de nombreux paramètres restent à optimiser et que l'environnement ainsi que la stratégie d'entraînement peuvent être largement améliorés. Hormis leur éfficacité, les modèle de deep reinforcement learning sont confrontés à la question de l'interprétabilité. En effet, la plupart de ces modèles ne peuvent pas être résumés à des règles de décision simples. Pour les modèles d'apprentissage supervisé, il existe des métriques permettant de donner l'importance des différentes variables et le sens de leur impact, comme les SHAP values. Pour les modèles de reinforcement learning la question est plus complexe. Des approches cherchent à étudier des « circuits » de neurones connectés entre eux par des poids élevés afin d'en déduire des features créés à partir des données d'entrée qui seraient interprétables.

Bibliographie

[1] Miquel Noguer i Alonso and Sonam Srivastava. 2020. Deep Reinforcement Learning for Asset Allocation in US Equities. https://doi.org/10.48550/ARXIV.2010.04404

[2] Alan Fontoura, Diego Haddad, and Eduardo Bezerra. 2019. A Deep Reinforcement Learning Approach to Asset-Liability Management. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. 216–221. https://doi.org/10.1109/BRACIS.2019.00046

[3] Schrittwieser J, I. Antonoglou, T. Hubert, and et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play.. In *Science 588*. t1140–1144. https://doi.org/10.1126/science.aar6404

[4] J. Jumper, R. Evans, and A. et al. Pritzel. 2021. Highly accurate protein structure prediction with AlphaFold.. In *Nature 596*. 583–589. https://doi.org/10.1038/s41586-021-03819-2

[5] Anssi Kanervisto, Christian Scheller, and Ville Hautamäki. 2020. Action Space Shaping in Deep Reinforcement Learning. https://doi.org/10.48550/ARXIV.2004.00980

[6] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. https://doi.org/10.48550/ARXIV.1509.02971

[7] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. https://doi.org/10.48550/ARXIV.1312.5602

[8] Andrew W. Moore and Christopher G. Atkeson. 1993. Prioritized Sweeping : Reinforcement Learning with Less Data and Less Real Time.

[9] J. Schrittwieser, I. Antonoglou, T. Hubert, and et al. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. In *Nature 588*. 604–609. https://doi.org/10.1038/s41586-020-03051-4

[10] D. Silver, A. Huang, C. Maddison, and et al. 2016. Mastering the game of Go with deep neural networks and tree search. In *Nature 529*. 484–489. https://doi.org/10.1038/nature16961

[11] David Silver, J Schrittwieser, I. Antonoglou, K. Simonyan, and et al. 2017. Mastering the game of Go without human knowledge. In *Nature 550*. 354–359. https://doi.org/10.1038/nature24270

[12] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning : An introduction*. MIT press.

[13] Konstantinos Saitas Zarkias, Nikolaos Passalis, Avraam Tsantekidis, and Anastasios Tefas. 2019. Deep Reinforcement Learning for Financial Trading Using Price Trailing. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3067–3071. https://doi.org/10.1109/ICASSP.2019.8683161