Optimiser un portefeuille par apprentissage par renforcement

Concevoir sa fonction objectif, comparer Q-learning et SARSA, mesurer un P&L réel

1 Avertissement

Ce document n'est pas un énoncé au sens classique. L'intérêt de ce projet est d'acquérir de l'expérience, par la lecture de l'état de l'art et par votre autoformation aux méthodes présentées. Au regard de votre expérience et de vos résultats intermédiaires, vous pouvez et êtes encouragés à modifier ou adapter la stratégie, et à redéfinir les objectifs.

2 Contexte et motivation

Gérer dynamiquement une position (combien d'unités d'un actif détenir à chaque instant) est le cœur du métier d'un gérant systématique ou d'un desk d'exécution : on veut profiter des mouvements de prix, mais chaque transaction coûte (spread, glissement, impact de marché) et la position expose au risque. Les méthodes classiques (contrôle optimal à la Almgren–Chriss, heuristiques de seuil) exigent un modèle explicite du marché. L'*apprentissage par renforcement* (RL) propose une alternative : un agent apprend une politique en interagissant avec un environnement, sans résoudre analytiquement le problème.

Le point central de ce sujet est que **la politique apprise dépend entièrement de la fonction objectif**. Choisir la récompense, c'est modéliser : une récompense qui ignore le risque produira un agent qui prend des paris extrêmes ; une récompense qui ignore la valeur résiduelle de l'inventaire produira un agent qui liquide tout ; une récompense sans coût d'inventaire produira un agent qui ne vend jamais. Vous allez donc *concevoir* cette fonction objectif, justifier chacun de ses paramètres, puis voir concrètement ce que deux algorithmes classiques (Q-learning et SARSA) en font.

Le sujet s'inspire d'un projet de cours de modélisation stochastique dans lequel un MDP de gestion d'inventaire est enrichi de *régimes de volatilité* et d'un *indicateur de momentum*. Vous partez de ce modèle de référence (annexe A), mais l'objectif est de le dépasser : adapter la récompense, calibrer l'environnement sur des données réelles, et finir par rejouer la stratégie apprise sur des prix historiques pour obtenir un **vrai P&L**.

Critère en une phrase. Un agent ajuste chaque jour son inventaire d'un actif (de 0 à Q unités) ; le P&L est la variation de valeur de sa position, nette des coûts, rejouée hors échantillon sur des prix historiques réels et comparée à une stratégie passive.

3 Type de projet et projet associé

— **Type principal :** quant trading (conception, backtest et simulation d'une stratégie), avec une composante de quant research (étude comparative de deux algorithmes de RL et de l'effet de la fonction objectif).

— **Projet appliqué associé :** intégré au sujet. La partie 5 des objectifs (rejouer la politique sur des prix réels avec coûts) constitue la chaîne complète *donnée* → *modèle* → *politique* → *décision* → *P&L*

4 Objectifs

4.1 Objectifs (initiaux, révisables)

1. **Modéliser.** Implémenter le MDP de référence à états (q, p, v, m) (annexe A), le valider (probabilités de transition qui somment à 1, comportement sur cas limites), et le résoudre exactement par *value iteration* pour disposer d'un étalon.
2. **Concevoir la fonction objectif.** Choisir ou construire une récompense dans (ou en dehors de) la famille proposée en section 6.2, et **justifier chaque paramètre introduit** (sens économique, unité, ordre de grandeur, source de calibrage, test de sensibilité).
3. **Résoudre par RL.** Appliquer Q-learning et SARSA (versions tabulaires) et comparer leurs politiques à l'optimum exact et aux heuristiques.
4. **Interpréter.** Expliquer pourquoi les politiques diffèrent : cartes de politique, distribution stationnaire, métriques par régime et par momentum, différence entre apprentissage *on-policy* et *off-policy*.
5. **Obtenir un P&L réel.** Calibrer l'environnement sur des prix historiques d'un actif liquide, entraîner sur une période, puis rejouer la politique hors échantillon avec des coûts réalistes, et comparer au *buy-and-hold*.

5 Données

— **Environnement synthétique (aucune donnée) :** le MDP de référence est un simulateur. Coût : nul. Limite : les rendements y sont indépendants, de loi en cloche à variance discrète par régime, ce qui n'a ni queues épaisses ni autocorrélation de la volatilité au-delà des deux régimes.

— **Prix historiques réels :** prix journaliers *ajustés* (dividendes et splits) d'un actif liquide choisi par l'équipe (par exemple un ETF actions large capitalisation ou une action du CAC 40), sur au moins 10 ans. Sources gratuites possibles : Stooq (export CSV), Yahoo Finance via la bibliothèque `yfinance` (non officielle : conditions d'utilisation à vérifier, usage pédagogique). Coût : nul.

— **Qualité et limites :** le spread bid-ask n'est pas fourni par ces sources : il doit être posé comme hypothèse (de l'ordre de quelques points de base pour un instrument très liquide) et justifié, ou estimé à partir des cours haut/bas (estimateur de Corwin–Schultz, cf. références). Jours fériés, valeurs manquantes et ajustements de dividendes sont à contrôler.

6 Grandes étapes et idées de stratégies (Exemple de solution et qui peut bien évidémment être étendu)

6.1 Le MDP de référence

L'état est $s_t = (q_t, p_t, v_t, m_t)$: inventaire $q \in \{0, \dots, Q\}$, prix discrétisé p, régime de volatilité $v \in \{\text{bas}, \text{haut}\}$, momentum $m \in \{\text{baisse}, \text{plat}, \text{hausse}\}$. L'action $a_t \in \{0, \dots, Q\}$ est l'*inventaire cible* ; la quantité échangée est $\delta_t = a_t - q_t$, et $q_{t+1} = a_t$. Deux tailles sont proposées : **niveau 1** ($Q = 5$, $P = 10$: $6 \times 10 \times 2 \times 3 = 360$ états) pour valider vite et comparer à la *value iteration* ; **niveau 2** ($Q = 20$, prix de 25 à 50 : $21 \times 26 \times 2 \times 3 = 3\,276$ états, 21 actions, soit 68 796 valeurs $Q(s, a)$).

Le régime suit une chaîne de Markov à deux états ($p_\uparrow$: bas $\to$ haut, $p_\downarrow$: haut $\to$ bas). Le rendement ρ_t est tiré dans $\{-r_{\max}, \dots, +r_{\max}\}$ avec une loi en cloche $w_v(\rho) \propto \exp(-\rho^2/2\eta_v^2)$ dont l'écart η_v dépend du régime, et le prix évolue selon

$$p_{t+1} = \text{clip}\Big(\text{round}\,(p_t(1 + \rho_t + \mu_{m_t}) + \beta\,\delta_t),\, p_{\min},\, p_{\max}\Big),$$

où $\mu_m \in \{-\mu, 0, +\mu\}$ est la dérive de momentum et β le coefficient d'impact de marché. Le momentum suivant est déterministe : $m_{t+1} = \text{sign}(p_{t+1} - p_t)$.

Artefact de grille. Avec des prix entiers entre 25 et 50 et des pas de rendement de 1 %, un mouvement de ± 1 % déplace le prix de 0,25 à 0,5 unité : l'arrondi le ramène souvent à zéro. Le momentum « plat » est alors sur-représenté, surtout à bas prix et en régime bas, et le modèle n'a plus les propriétés qu'on croit lui avoir données. Mesurez la distribution effective de $p_{t+1} - p_t$ et ajustez la finesse de la grille (ou passez à une grille en log-prix) avant de tirer des conclusions économiques.

6.2 Famille de fonctions objectif

Notons $a = q + \delta$ l'inventaire après transaction et $\Delta p = p_{t+1} - p_t$. Voici une famille de départ ; vous pouvez la combiner, la modifier ou proposer la vôtre.

Id	Récompense	Point d'attention
R0	$a\,\Delta p - \lambda\,\delta^2$ (valeur de marché de la nouvelle position, moins glissement quadratique)	Récompense « partie I » du projet d'origine ; ne pénalise ni le risque ni le portage.
R1	$-\delta\,p - \lambda\,\delta^2$ (flux de trésorerie : achat coûte, vente rapporte)	Ignore la valeur résiduelle de l'inventaire : l'agent est incité à liquider ; à compléter par une valeur terminale ou une valorisation $a\,p$.
R2	R0 $- \varepsilon\,p\,\lvert\delta\rvert$ (coût proportionnel, type spread/commission)	Plus réaliste que le seul coût quadratique pour un actif liquide ; mélange possible des deux.
R3	R2 $- \frac{\phi}{2}\,a^2 p^2 \sigma_v^2$ (pénalité de variance d'inventaire, σ_v dépendant du régime)	Rend le régime v économiquement pertinent ; la politique apprise ne maximise plus le P&L pur.
R4	R3 $- c\,a\,p$ (coût de portage ou de financement par pas)	Évite que « toujours au maximum » domine ; calibrer c sur un taux de financement.
R5	Transformation concave : $-\frac{1}{\eta}\exp(-\eta\,r)$ appliquée à R2	Sensibilité au risque ; non additive dans le temps, donc interprétation à discuter.

Pistes avancées : *differential Sharpe ratio* (Moody & Saffell), pénalité de drawdown. Ces deux options exigent d'augmenter l'état (moments courants, maximum historique de la valeur du portefeuille), ce qui fait exploser la taille de l'espace d'états : à vous d'arbitrer.

Autres choix de modélisation à justifier : facteur d'actualisation γ (horizon effectif $1/(1-\gamma)$ pas : quelle durée de décision cela représente-t-il ?), longueur d'épisode $T_{\max}$, présence ou non d'une valeur terminale, ordre d'exécution (au cours de clôture ou à l'ouverture suivante).

6.3 Fiche de justification des paramètres

Chaque paramètre que vous introduisez ou modifiez doit apparaître dans un tableau de ce type. Un paramètre non justifié est un paramètre que le correcteur considérera comme arbitraire.

Param.	Sens économique, unité	Valeur retenue	Source / calibrage	Test de sensibilité
λ	Glissement quadratique par unité^2	…	…	…
ε	Coût proportionnel (fraction du notionnel)	…	Spread estimé ou hypothèse	…
ϕ	Aversion au risque	…	…	…
γ	Actualisation	…	Horizon de décision visé	…
…	…	…	…	…

6.4 Algorithmes et protocole d'apprentissage

Q-learning (off-policy) et **SARSA (on-policy)** mettent à jour la table Q à partir de la transition observée (s, a, r, s') :

$$\text{Q-learning :} \quad Q(s,a) \leftarrow Q(s,a) + \alpha\big[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\big],$$

$$\text{SARSA :} \quad Q(s,a) \leftarrow Q(s,a) + \alpha\big[r + \gamma\, Q(s',a') - Q(s,a)\big], \quad a' \sim \varepsilon\text{-greedy}.$$

La différence est fondamentale : SARSA évalue la politique *effectivement suivie*, exploration comprise ; Q-learning évalue la politique gloutonne. Quand explorer coûte cher (coût quadratique en δ, grosses transactions aléatoires), SARSA « voit » ce coût et tend à apprendre une politique plus prudente, comme dans l'exemple classique de la falaise (Sutton & Barto). **À vous de tester si c'est le cas ici**, en faisant varier $\varepsilon_{\text{final}}$ et la fonction objectif.

Ingrédients utiles : exploration ε décroissante (par exemple de 0,8 à 0,01 linéairement), initialisation de la table Q par l'évaluation d'une politique heuristique (*warm start*) plutôt que par des zéros, distribution initiale biaisée vers les états rares, nombre d'épisodes dimensionné pour une couverture moyenne d'environ 150 visites par couple (s, a). Pour une comparaison équitable, utilisez le **même** taux d'apprentissage et la **même** planification de ε pour les deux algorithmes, et plusieurs graines aléatoires (intervalles de confiance).

Étalon exact. Le modèle étant connu, la *value iteration* donne V^* et π^*. Mesurez l'écart des politiques RL : différence de J, pourcentage d'états où l'action coïncide *pondéré par la distribution stationnaire* (un désaccord dans un état jamais visité est sans importance).

6.5 Heuristiques de référence

— **Buy-and-hold :** acheter Q au départ puis ne plus échanger.

— **Seuils (buy low / sell high) :** $\pi(q,p,v,m) = Q$ si $p \le p_{\text{bas}}$, 0 si $p \ge p_{\text{haut}}$, q sinon ; seuils par défaut au tiers inférieur et supérieur de la grille, à ajuster *sur un jeu de validation*.

— **Toujours au maximum** et **ne jamais trader** (bornes de comparaison).

— **Ciblage de volatilité :** $a = \text{round}(Q\,\sigma_{\text{cible}}/\sigma_v)$, qui exploite le régime. Il constitue un meilleur adversaire que les heuristiques aveugles au régime.

6.6 Métriques d'analyse

Calculées par rollouts Monte-Carlo sous la distribution stationnaire de chaque politique :

— **Rendement actualisé moyen** $J(\pi) = \mathbb{E}_\pi\big[\sum_t \gamma^t r_t\big]$.

— **Performance par régime** $J_v(\pi) = \mathbb{E}_\pi[r_t \mid v_t = v]$: une politique qui ignore v affiche $J_{\text{bas}} \approx J_{\text{haut}}$.

— **Score d'exploitation du momentum** $M(\pi) = \mathbb{E}_\pi[r_t \mid m_t = \text{hausse}] - \mathbb{E}_\pi[r_t \mid m_t = \text{baisse}]$.

— **Distribution stationnaire** (par valeurs propres de P_π ou par simulation) : inventaire moyen, valeur moyenne de la position $q\,p$, fréquence et taille moyenne des transactions, part des coûts dans le gain brut, temps passé à $q = 0$ et $q = Q$, pourcentage d'états jamais visités.

— **Risque :** écart-type et VaR/ES à 5 % du rendement total, ratio de Sharpe, drawdown maximal.

6.7 Du simulateur au P&L réel

Procédure proposée :

1. **Découpage chronologique** : entraînement, validation, test, sans mélange. Les seuils de discrétisation (quantiles de prix relatif, seuil de régime) sont calculés *sur l'entraînement uniquement*.
2. **Construction de l'état à partir des données** : prix discrétisé non pas en niveau brut (non stationnaire) mais en écart à une moyenne mobile, en quantiles ; régime v par seuil sur la volatilité réalisée ou par modèle de Markov caché à deux états ; momentum m par signe du rendement sur k jours avec zone neutre. Toute grandeur à la date t n'utilise que des données $\leq t$.
3. **Calibrage de l'environnement** (rendements par régime, probabilités de régime, dérive de momentum) sur l'entraînement, puis apprentissage de Q et π sur cet environnement.
4. **Rejeu hors échantillon** : à chaque date, observer s_t, jouer $a_t = \pi(s_t)$, exécuter avec les coûts réalistes, calculer le P&L journalier $a_t(P_{t+1} - P_t) - \text{coûts}$.
5. **Comparaison** : courbe de capital, Sharpe, drawdown, rotation (turnover), et **écart simulateur → réel** (la performance en simulation survit-elle sur données réelles ? pourquoi ?). Rapportez séparément la *récompense apprise* et le *P&L économique* : avec une pénalité de risque, ils diffèrent.

6.8 Extensions possibles

— **Deep Q-Network** (réseau de neurones à la place de la table, *experience replay*) sur un état continu non discrétisé ; comparer à la version tabulaire (Mnih et al.).

— **Deux actifs corrélés** : illustration de la malédiction de la dimension.

— **Algorithmes avec modèle** (Dyna-Q) ou **Expected SARSA**.

— **Réentraînement glissant** (walk-forward) pour tester la stabilité dans le temps.

6.9 Pièges classiques à éviter

- **Look-ahead bias :** état ou seuil de discrétisation calculé avec des informations postérieures à la date de décision ; exécution à un prix déjà connu.
- **Fuite de données (data leakage) :** quantiles ou régimes estimés sur toute la période, test utilisé pour régler les hyperparamètres.
- **Surapprentissage et p-hacking :** multiplier les fonctions objectif et paramètres testés puis ne présenter que le meilleur. Tenez le compte du nombre de configurations essayées (ratio de Sharpe *déflaté*) et ne touchez pas au jeu de test avant la fin.
- **Détournement de la récompense (reward hacking) :** l'agent exploite un défaut du modèle (arrondi de la grille, absence de coût de portage, valeur résiduelle ignorée) au lieu d'un phénomène de marché.
- **Coûts irréalistes :** impact de marché β nul ou énorme par rapport à la taille réellement négociée ; coût purement quadratique alors que les coûts réels sont surtout linéaires.
- **Non-stationnarité :** un régime appris sur une décennie peut disparaître ; testez sur plusieurs sous-périodes.
- **Comparaison déséquilibrée :** un taux d'apprentissage différent pour Q-learning et SARSA, une seule graine, des couples (s, a) rarement visités.

7 Références et pistes de lecture

- R. Sutton, A. Barto, *Reinforcement Learning : An Introduction*, 2e éd., MIT Press, 2018 (chap. 3 et 6 : MDP, SARSA, Q-learning, exemple de la falaise).
- C. Watkins, P. Dayan, « Q-learning », *Machine Learning*, 8, 1992. G. Rummery, M. Niranjan, « On-line Q-learning using connectionist systems », rapport technique, Cambridge, 1994 (origine de SARSA).
- R. Almgren, N. Chriss, « Optimal execution of portfolio transactions », *Journal of Risk*, 3, 2000 (coûts d'impact et risque).
- J. Moody, M. Saffell, « Learning to trade via direct reinforcement », *IEEE Trans. Neural Networks*, 12(4), 2001 (differential Sharpe ratio).
- B. Hambly, R. Xu, H. Yang, « Recent advances in reinforcement learning in finance », *Mathematical Finance*, 2023.
- P. Kolm, G. Ritter, « Modern perspectives on reinforcement learning in finance », *Journal of Machine Learning in Finance* (version de travail SSRN, 2019).
- V. Mnih et al., « Human-level control through deep reinforcement learning », *Nature*, 518, 2015.
- J. Hamilton, « A new approach to the economic analysis of nonstationary time series and the business cycle », *Econometrica*, 57, 1989 (régimes).
- D. Bailey, M. López de Prado, « The deflated Sharpe ratio », *Journal of Portfolio Management*, 40(5), 2014.
- S. Corwin, P. Schultz, « A simple way to estimate bid-ask spreads from daily high and low prices », *Journal of Finance*, 67(2), 2012.
- Outils : `POMDPs.jl` et `TabularTDLearning.jl` (Julia), ou implémentation tabulaire directe en Python/NumPy (la table reste petite).

8 Jalons et résultats intermédiaires attendus

Jalon	Livrable intermédiaire	Échéance
J1	Revue de littérature (RL en finance, SARSA vs Q-learning) et plan de travail ; premières fonctions objectif envisagées	Semaines 1–2
J2	MDP niveau 1 implémenté et validé ; données réelles collectées, nettoyées, statistiques descriptives ; distribution de $p_{t+1} - p_t$ (artefact de grille)	Semaines 3–4
J3	Baseline bout en bout : une récompense, Q-learning et SARSA, *value iteration*, heuristiques ; premiers résultats et fiche de paramètres	Semaines 5–6
J4	Plusieurs fonctions objectif comparées ; niveau 2 ; calibrage sur données ; premier P&L hors échantillon ; analyse de robustesse et des échecs	Semaines 7–9
J5	Livrable final : code reproductible, rapport, présentation	Semaine 10

Les échéances sont indicatives ; un point court et régulier avec un référent du comité est recommandé.

9 Prérequis et niveau de difficulté estimés

— **Niveau :** intermédiaire à avancé ; équipe de 3 ou 4 étudiants ; environ 10 semaines.

— **Prérequis :** probabilités (chaînes de Markov, espérance conditionnelle), programmation Python ou Julia, notions d'apprentissage par renforcement (MDP, équation de Bellman) ; la lecture des chapitres 3 et 6 de Sutton & Barto suffit pour démarrer.

— **Points durs :** justification honnête des paramètres, construction d'un état sans fuite d'information, interprétation de la différence entre Q-learning et SARSA.

10 Confidentialité

Ce document est un document de travail interne au comité scientifique et aux équipes de BlaCS-Scholes. Merci de ne pas le distribuer en dehors.

A Annexe A : paramètres de référence du MDP

Paramètre	Symbole	Niveau 1	Niveau 2
Inventaire maximal	Q	5	20
Prix (grille)	p	$\{1, \ldots, 10\}$	$\{25, \ldots, 50\}$
Amplitude des rendements	$r_{\max}$	5 %	5 %
Pas de rendement	r_{step}	1 %	1 %
Écart en régime bas / haut	$\eta_{\text{bas}}, \eta_{\text{haut}}$	2 % / 5 %	1,5 % / 5 %
Probabilité bas → haut	$p_{\uparrow}$	0,05	0,05
Probabilité haut → bas	$p_{\downarrow}$	0,20	0,20
Dérive de momentum	μ	0,5 %	1 %
Impact de marché	β	0,3	0 (à discuter)
Glissement quadratique	λ	0,3	0,1
Actualisation	γ	0,95	0,999
Taux d'apprentissage	α	0,1	0,1
Longueur d'épisode	$T_{\max}$	…	1 000

Ces valeurs sont des points de départ, **non des valeurs justifiées** : les remplacer par des valeurs argumentées fait partie du travail demandé.