# Rapport : Continuous-Time Mean–Variance Portfolio Selection

Aaron Clotaire Metchinjin Kombou, Ryan Witt, Raphaël Bittolo

17 mars 2025

## Table des matières



# 1 Analyse de l'article

L'article s'intéresse au problème de choix de portefeuille dans le cadre d'une optimisation moyenne-variance avec apprentissage par renforcement (RL) en temps continu.

## 1.1 Modèle classique

On s'intéresse dans un premier temps au problème classique, c'est-à-dire une optimisation moyenne variance sans exploration. On fixe un horizon $T > 0$.

On considère un actif financier modélisé par :

$$dS_t = S_t(\mu dt + \sigma dW_t), \quad S_0 = s_0 \in R$$

avec $(W_t)_{t\in[0,T]}$ un mouvement brownien, $\sigma > 0$ et $\mu \in R$. On suppose aussi qu'il existe un actif sans risque de dynamique $\frac{dS_t^0}{S_t^0} = rdt$.

La dynamique du portefeuille autofinançant actualisé est alors donnée par :

$$dx_t^u = \sigma u_t(\rho dt + dW_t)$$

avec $(u_t)_{t\in[0,T]}$ la stratégie du portefeuille (valeur actualisée mise dans S) et $\rho = (\mu - r)/\sigma$.

Le problème d'optimisation classique est :

$$\min_u \mathrm{Var}(x_T^u) \quad \text{sous contrainte} \quad E(x_T^u) = z$$

En considérant le Lagrangien $w$, on se ramène au problème :

$$\min_u E[(x_T^u - w)^2] - (w - z)^2$$

**Observations sur le modèle classique** Pour résoudre ce problème, il faut identifier les paramètres du modèle ($\mu$ et $\sigma$), ce qui est difficile numériquement. De plus, le modèle est très sensible aux paramètres (on l'observe expérimentalement avec le modèle MLE). L'avantage du modèle RL réside dans le fait qu'il n'a pas besoin d'apprendre les paramètres du modèle, mais qu'il se contente d'apprendre une allocation optimale d'après les données de marché.

### 1.2 Nouveau modèle

Dans le nouveau modèle, on change le processus de contrôle $(u_t)_{t\in[0,T]}$ pour le rendre aléatoire, suivant une loi $(\pi_t)_{t\in[0,T]}$.

Le portefeuille actualisé a maintenant une dynamique donnée par :

$$dX_t^\pi = \tilde{b}(\pi_t)dt + \tilde{\sigma}(\pi_t)dW_t$$

avec $\tilde{b}(\pi) = \int_{\mathbf{R}} \rho\sigma u\pi(u)du$, et $\tilde{\sigma}(\pi) = \sqrt{\int_{\mathbf{R}} \sigma^2 u^2 \pi(u)du}$, $\pi \in \mathbf{P}(\mathbf{R})$.

Comme dans la plupart des problème d'apprentissage par renforcement, on utilise l'entropie cumulée définie par $H(\pi) = -\int_0^T \int_{\mathbf{R}} \pi_t(u) \ln \pi_t(u)dudt$ pour quantifier le niveau d'exploration de $\pi$.

Le nouveau problème d'optimisation est :

$$\min_{\pi\in A(x_0,0)} E[(X_T^\pi - w)^2 - \lambda H(\pi)] - (w - z)^2$$

avec et $\lambda > 0$ un paramètre de température, permettant de contrôler le niveau d'exploration. En prenant $\lambda = 0$, il n'y a pas d'exploration, on retrouve un problème proche du problème classique. Plus $\lambda$ est grand, plus on pousse à l'exploration.

### 1.3 Résolutions

On considère $V$ et $V^\pi$ définies par :

$$V(s,x,w) = \inf_{\pi\in A(s,y)} E[(X_T^\pi - w)^2 - \lambda H(\pi)|X_s^\pi = y] - (w - z)^2$$

$$V^\pi(s,x,w) = E[(X_T^\pi - w)^2 + \lambda \int_s^T \int_{\mathbf{R}} \pi_t(u) \ln \pi_t(u)dudt|X_s^\pi = y] - (w - z)^2$$

En passant par HJB, on peut obtenir les expressions de V, $\pi^*$ (loi du contrôle optimal) et $w$ :

$$V(t,x,w) = (x-w)^2 \exp(-\rho^2(T-t)) + \frac{\lambda\rho^2}{4}(T^2 - t^2) - \frac{\lambda}{2}(\rho^2 T - \ln(\frac{\rho^2}{\pi\lambda}))(T-t) - (w-z)^2$$

$$\pi^*(u,t,x,w) = N(u|-\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2}\exp(\rho^2(T-t))$$

$$w = \frac{z\exp(\rho^2 T) - x_0}{\exp(\rho^2 T) - 1}$$

### 1.4 Lien entre le problème avec et sans exploration

En posant $\lambda = 0$, on récupère la solution classique. Cette équivalence montre que les deux problèmes partagent le même multiplicateur de Lagrange $w$ et que la politique optimale converge vers la solution classique lorsque $\lambda \to 0$.

Finalement, le coût d'exploration est donné par :

$$C_{u^*,\pi^*}(0, x_0; w) = \frac{\lambda T}{2},$$

indépendant de $w$. Cela signifie que le coût de l'exploration augmente avec $\lambda$ et $T$, mais pas avec $w$.

### 1.5 Algorithme EMV (exploratory mean-variance)

En utilisant l'ensemble des résultats théoriques obtenus, l'article propose un algorithme pour apprendre les solutions du problème. Il énonce aussi quelques résultats qui permettent de prouver la convergence théorique de l'algorithme. L'algorithme se compose de trois procédures se déroulant simultanément : l'évaluation de la politique, l'amélioration de la politique et un mécanisme d'auto-correction pour l'apprentissage du multiplicateur de Lagrange w.

On observe que $V^\pi(t,x) = E(V^\pi(s, X_s) + \lambda \int_t^s \int_R \pi_v(u) \ln(\pi_v(u)) du dv | X_t = x)$

En arrangeant l'équation, on obtient :

$$\mathbf{E}(\frac{V^\pi(s, X_s) - V^\pi(t,x)}{s-t} + \frac{\lambda}{s-t} \int_t^s \int_R \pi_v(u) \ln(\pi_v(u)) du dv | X_t = x) = 0$$

En faisant tendre s vers t, on introduit $\delta_t$, définie par :

$$\delta_t = \dot{V}_t^\pi + \lambda \int_{\mathbf{R}} \pi_t(u) \ln(\pi_t(u)) du$$

avec $\dot{V}_t^\pi = \frac{V^\pi(t+\Delta t, X_{t+\Delta t}) - V^\pi(t, X_t)}{\Delta t}$. On note $V^\theta$ et $\pi^\phi$ les fonctions paramétriques qu'on utilise pour approcher $V$ et $\pi$. On définie par finalement une fonction coût $C$ par :

$$\begin{aligned} C(\theta, \phi) &= \frac{1}{2}\mathbb{E}\left[\int_0^T |\delta_t|^2 \, dt\right] \\ &= \frac{1}{2}\mathbb{E}\left[\int_0^T \left|\dot{V}_t^\theta + \lambda \int_{\mathbb{R}} \pi_t^\phi(u) \ln \pi_t^\phi(u) \, du\right|^2 dt\right] \\ &\approx \frac{1}{2} \sum_{(t_i, x_i)} \left(\dot{V}^\theta(t_i, x_i) + \lambda \int_{\mathbb{R}} \pi_{t_i}^\phi(u) \ln \pi_{t_i}^\phi(u) \, du\right)^2 \Delta t. \end{aligned}$$

On cherche à minimiser cette quantité en fonction des paramètres $\theta$ et $\phi$. Les résultats précédents sur la forme des fonctions optimales nous permettent de réduire l'ensemble des fonctions recherchées. On a donc $H(\pi_t^\phi) = \phi_1 + \phi_2(T-t)$ et $V^\theta(t,x) = (x-w)^2 \exp(-\theta_3(T-t)) + \theta_2 t^2 + \theta_1 t + \theta_0$. L'étude des conditions aux limites et du lien entre les paramètres amène ensuite à $V(T,.)^\theta = (.-w)^2 - (w-z)^2$ et $\theta_3 = 2\phi_2$. L'avantage de cette méthode est le faible nombre de paramètres à optimiser. De plus, l'expression de $C$ nous permet de facilement obtenir l'expression des différents gradients. Les méthodes alternatives consistent en l'estimation des paramètres $\mu$ et $\sigma$ afin de directement obtenir V et $\pi^*$ à l'aide des théorèmes précédents (MLE). Il est également possible d'utiliser des réseaux de neurones pour approximer les fonctions. Cependant, ces derniers possèdent de nombreux paramètres, ce qui peut rendre l'entraînement long. À l'inverse, nous avons ici trouvé un modèle ne comportant que cinq paramètres, permettant ainsi un entraînement plus rapide.

L'algorithme se déroule en deux étapes. Dans un premier temps on simule une trajectoire de $(S_t)_{t\in[0,T]}$ puis de $x_t^{\pi^\phi}$. On adapte ensuite les paramètres $\theta$ et $\phi$ pour minimiser C, puis on recommence. A intervalle régulier, on met à jour la valeur de $w$.

**Algorithm 1 EMV : Exploratory Mean-Variance Portfolio Selection**

**Require:** Market Simulator *Market*, learning rates $\alpha, \eta_\theta, \eta_\phi$, initial wealth $x_0$, target payoff $z$, investment horizon $T$, discretization $\Delta t$, exploration rate $\lambda$, number of iterations $M$, sample average size $N$.

Initialize $\theta, \phi$ and $w$

**for** $k = 1$ to $M$ **do**

  **for** $i = 1$ to $\lfloor \frac{T}{\Delta t} \rfloor$ **do**

    Sample $(t_i^k, x_i^k)$ from *Market* under $\pi^\phi$

    Obtain collected samples $\mathcal{D} = \{(t_i^k, x_i^k), 1 \le i \le \lfloor \frac{T}{\Delta t} \rfloor\}$

    Update $\theta \leftarrow \theta - \eta_\theta \nabla_\theta C(\theta, \phi)$ using (47) and (48)

    Update $\theta_0$ using (51) and $\theta_3 \leftarrow 2\phi_2$

    Update $\phi \leftarrow \phi - \eta_\phi \nabla_\phi C(\theta, \phi)$ using (49) and (50)

  **end for**

  Update $\pi^\phi \leftarrow \mathcal{N}\left(u \middle| -\sqrt{\frac{2\phi_2}{\lambda\pi}} e^{-\frac{1}{2}(x-w)}, \frac{1}{2\pi} e^{2\phi_2(T-t)+2\phi_1-1}\right)$

  **if** $k \mod N == 0$ **then**

    Update $w \leftarrow w - \alpha\left(\frac{1}{N}\sum_{j=k-N+1}^{k} x^j_{\lfloor \frac{T}{\Delta t} \rfloor} - z\right)$

  **end if**

**end for**

## 2 Implémentation et résultats

Nous avons implémenté l'algorithme EMV et la première alternative MLE. Dans cette section, nous faisons part des difficultés techniques, des solutions utilisées et des résultats numériques.

### 2.1 MLE

L'algorithme MLE est basé sur l'estimation de $\mu$ et de $\sigma$ en utilisant la donnée des 100 derniers prix. Partant de la série des log rendements $r_t = \log(S_{t+1}) - \log(S_t)$, ces estimateurs sont donnés par :

$$\hat{\sigma} = \frac{1}{\sqrt{\Delta t}} \sqrt{\frac{1}{n-1}\sum_{t=1}^{n-1}(r_t - \bar{r})^2} \qquad \hat{\mu} = \frac{1}{\Delta t}\bar{r} + \frac{1}{2}\hat{\sigma}^2$$

Pour toutes les simulations, l'estimation de $\sigma$ était proche de la vraie valeur. On notait toutefois des instabilités dans l'estimation de $\mu$ (aussi appelé *mean-blur*) entraînant des allocations $u_t$ excessivement élevées et conduisant à des valeurs aberrantes de la richesse finale $x_T$. Pour résoudre ce problème, un **clipping** de l'allocation $u_t$ a été appliqué afin de limiter l'exposition à l'actif risqué. On utilise donc le contrôle définit par $\tilde{u} = (s \wedge u) \vee -s$ avec $s \in R_+$ (en prendra par exemple $s \approx 3x_0$). u représentant la somme actualisée investie dans l'actif risqué, il est cohérent de limiter cette valeur en fonction de la somme investie initialement (ou en fonction de la somme que l'on possède). Cette modification permet d'améliorer significativement les résultats. On ob-

serve aussi le résultat du MLE en supprimant les valeurs extrêmes de $x_T$ (5% des plus grandes valeurs et 5% des plus petites valeurs).

| | Classique | Suppression des extrêmes | Restriction de u |
|---|---|---|---|
| $\hat{\mathbf{E}}(x_T)$ | -3.2 | 1.1 | 1.1 |
| $\hat{\text{Var}}(x_T)$ | $1.7 \times 10^4$ | $5.2 \times 10^{-2}$ | $8.2 \times 10^{-2}$ |
| Ratio de Sharpe | $-3.2 \times 10^{-2}$ | $4.3 \times 10^{-1}$ | $3.5 \times 10^{-1}$ |

TABLE 1 – *Résultat obtenu en prenant 500 trajectoires,* $\sigma = 0.2$, $\mu = 0.3$, $r = 0.02$, $T = 1$, $dt = 1/252$, $z = 1.4$, $x_0 = 1$

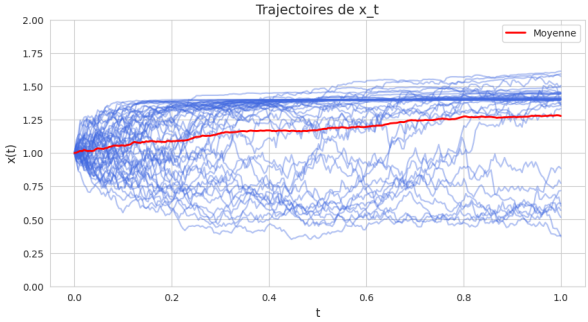



FIGURE 1 – *MLE avec modification, même paramètres que précédemment*

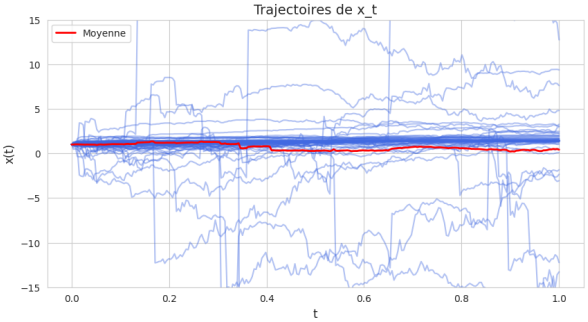



FIGURE 2 – *MLE sans modification, même paramètres que précédemment*

**Commentaire :** On peut observer graphiquement l'effet de la restriction de u. La restriction permet d'obtenir des courbes plus lisses, sans quoi on observe des variations extrêmes de $(x_t)_{t\in[0,T]}$, qui font exploser la variance.

## 2.2 Algorithme EMV

L'implémentation de l'algorithme EMV a rencontré deux principales difficultés techniques. D'abord, il a fallu respecter le paradigme du Reinforcement Learning en évitant l'utilisation directe des paramètres du marché tels que $\mu$, $\sigma$ et $\rho$, car ces valeurs sont supposées inconnues en pratique. Une adaptation de certaines formules pour garantir une prise de décision basée uniquement sur les trajectoires de prix $S_t$ (simulées au préalable) était nécessaire. Ainsi, nous avons utilisé une approche **"model-free"** pour la mise à jour de la richesse actualisée, remplaçant $dx_t^u = \sigma u_t(\rho dt + dW_t)$ par :

$$dx_t^u = u_t\left(\frac{dS_t}{S_t} - r dt\right)$$

En outre, la policy d'allocation, donnée par $u_t \sim \mathcal{N}\left(-\sqrt{\frac{2\phi_2}{\lambda\pi}}\, e^{\frac{2\phi_1-1}{2}}(x-w),\ \frac{1}{2\pi}\, e^{2\phi_2(T-t)+2\phi_1-1}\right)$, repose sur l'hypothèse implicite d'un $\rho > 0$. En pratique, les simulations avec les $\rho < 0$. ont engendré des instabilités numériques et des valeurs aberrantes de la richesse terminale. Dans ce contexte, il s'est avéré difficile de maintenir une stricte adhésion au paradigme du RL : l'information sur le signe de $\rho$ a été utilisée comme variable d'entrée afin de stabiliser la policy.

$$u_t \sim \mathcal{N}\left(-sign(\rho)\sqrt{\frac{2\phi_2}{\lambda\pi}}\, e^{\frac{2\phi_1-1}{2}}(x-w),\ \frac{1}{2\pi}\, e^{2\phi_2(T-t)+2\phi_1-1}\right) \tag{1}$$

Enfin, la gestion de l'explosion des gradients a constitué un autre défi. Certaines mises à jour des paramètres (notamment $\phi_1$ et $\phi_2$) ont conduit à des valeurs numériques instables. Pour y faire face, des techniques de **gradient clipping** (avec une clipping value de quelques unités) ont été appliquées afin de contrôler l'amplitude des gradients.

De plus, notre algorithme converge bien vers la bonne sélection de portefeuille. Cependant, les paramètres obtenus à la fin ne correspondent pas aux paramètres théoriques. Pire encore, en initialisant notre EMV avec les paramètres théoriques, l'algorithme ne converge pas plus rapidement vers le bon résultat. Surpris par ce constat, nous ne savons pas encore comment l'interpréter.

Dans les simulations, on prendra sauf précision contraire : $\sigma = 0.1$, $\mu = -0.3$, $r = 0.02$, $T = 1$, $dt = 1/252$, $z = 1.4$, $x_0 = 1$, $\lambda = 2$, $\eta_\phi = 0.0005$, $\eta_\theta = 0.0005$, $\alpha = 0.05$, $M = 20000$, $N = 10$, comme suggéré dans l'article.

## 2.3 Effet de $\lambda$

On $\lambda$ peut tenter d'observer l'effet de $\lambda$ sur les trajectoires obtenues. Théoriquement, augmenter la valeur de $\lambda$ amène l'agent à plus explorer et moins exploiter, ce qui provoque une augmentation de la variance.

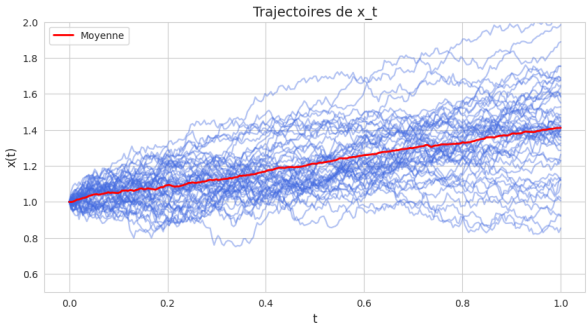



FIGURE 3 – $\lambda = 2$

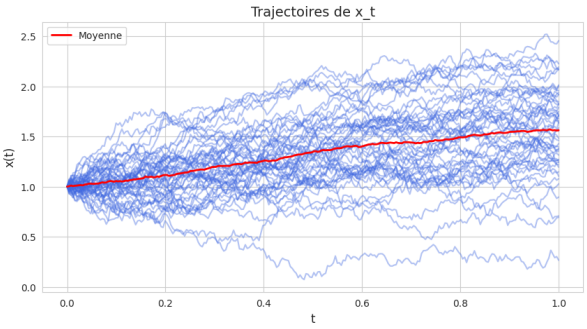



FIGURE 4 – $\lambda = 10$

| $\lambda$ | $\hat{\mathbb{E}}(x_T)$ | $\hat{\mathrm{Var}}(x_T)$ | Ratio de Sharpe |
|---|---|---|---|
| 1 | 1.38 | $7.14 \times 10^{-2}$ | 1.43 |
| 2 | 1.39 | $8.06 \times 10^{-2}$ | 1.37 |
| 10 | 1.34 | $5.71 \times 10^{-2}$ | 1.42 |
| 100 | 1.32 | $6.51 \times 10^{-2}$ | 1.25 |

TABLE 2 – *Simulations pour différentes valeurs de $\lambda$ (1000 trajectoires)*

| $\lambda$ | $\hat{\mathbb{E}}(x_T)$ | $\hat{\mathrm{Var}}(x_T)$ | Ratio de Sharpe |
|---|---|---|---|
| 1 | 1.39 | $1.86 \times 10^{-2}$ | 2.86 |
| 2 | 1.37 | $1.90 \times 10^{-2}$ | 2.68 |
| 10 | 1.37 | $1.89 \times 10^{-2}$ | 2.69 |
| 100 | 1.40 | $2.94 \times 10^{-2}$ | 2.33 |

TABLE 3 – *Simulations pour différentes valeurs de $\lambda$, $\sigma = 0.1$ (1000 trajectoires)*

**Commentaire :** Visuellement, la modification de la valeur de $\lambda$ n'est pas évidente lorsqu'on fait varier $\lambda$ entre 1 et 100. Cependant, on peut observer son effet dans la variance empirique : comme attendu une augmentation de $\lambda$ tend à la faire augmenter, mais cette variation n'est pas significative. On observe cependant une dégradation marquée du ratio de Sharpe.

## 2.4 Marché stationnaire

La figure 5 est un tableau qui présente les moyennes ($\bar{M}$), variances ($\bar{V}$) et Sharpes ($\frac{\bar{M}-1}{\sqrt{\bar{V}}}$) des 2000 dernières richesses terminales. Basé sur la comparaison des ratios de Sharpe, l'EMV

que nous avons implémenté surpasse le MLE dans 23 des 28 expériences.

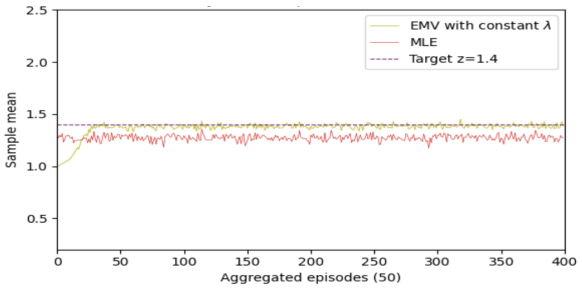
On observe par ailleurs que pour l'EMV, les résultats sont plus ou moins proche de ceux de l'articles avec un sharpe qui est d'autant plus petit que le drift $\mu$ est proche de 0 et que la volalité $\sigma$ est élevée. Cela suggère que l'algorithme EMV est plus performant lorsque le marché présente une forte tendance $\mu$ et une faible volatilité $\sigma$, tandis qu'il a des difficultés à générer des rendements élevés lorsque la tendance est faible et la volatilité élevée. Par ailleurs, bien que notre algorithme présente des résultats intéressants, il est environ trois fois plus lent que celui de l'article.

De ce qui est du MLE, la différence notoire entre nos résultats et ceux de l'article est probablement due à un choix très restrictif de la valeur de clipping de l'allocation.

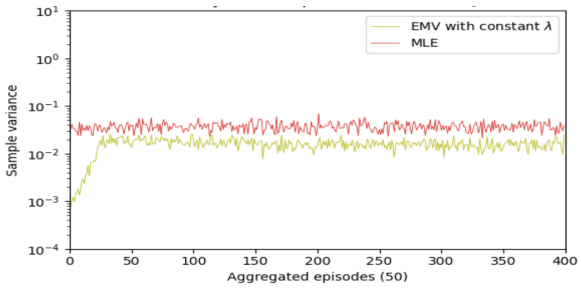
| Market scenarios | | EMV | | | MLE | | |
|---|---|---|---|---|---|---|---|
| **μ** | **σ** | **Mean** | **Var** | **Sharpe** | **Mean** | **Var** | **Sharpe** |
| -0,5 | 0,1 | 1,393 | 0,007 | 4,825 | 1,261 | 0,039 | 1,326 |
| -0,3 | 0,1 | 1,386 | 0,015 | 3,134 | 1,269 | 0,039 | 1,354 |
| -0,1 | 0,1 | 1,364 | 0,096 | 1,176 | 1,125 | 0,024 | 0,8 |
| 0 | 0,1 | 1,233 | 1,25 | 0,209 | 0,88 | 0,32 | -0,212 |
| 0,1 | 0,1 | 1,35 | 0,178 | 0,83 | 1,432 | 0,311 | 0,774 |
| 0,3 | 0,1 | 1,384 | 0,019 | 2,761 | 1,456 | 0,024 | 2,961 |
| 0,5 | 0,1 | 1,391 | 0,007 | 4,559 | 1,365 | 0,015 | 2,939 |
| -0,5 | 0,2 | 1,383 | 0,023 | 2,509 | 1,265 | 0,04 | 1,331 |
| -0,3 | 0,2 | 1,373 | 0,055 | 1,587 | 1,192 | 0,035 | 1,029 |
| -0,1 | 0,2 | 1,333 | 0,269 | 0,643 | 0,721 | 0,223 | -0,591 |
| 0 | 0,2 | 1,209 | 3,169 | 0,117 | 0,95 | 0,312 | -0,09 |
| 0,1 | 0,2 | 1,318 | 0,64 | 0,397 | 1,228 | 0,328 | 0,398 |
| 0,3 | 0,2 | 1,369 | 0,072 | 1,373 | 1,491 | 0,158 | 1,235 |
| 0,5 | 0,2 | 1,382 | 0,029 | 2,235 | 1,486 | 0,041 | 2,397 |
| -0,5 | 0,3 | 1,377 | 0,049 | 1,696 | 1,209 | 0,035 | 1,126 |
| -0,3 | 0,3 | 1,354 | 0,111 | 1,064 | 0,834 | 0,046 | -0,773 |
| -0,1 | 0,3 | 1,302 | 0,635 | 0,379 | 0,798 | 0,276 | -0,385 |
| 0 | 0,3 | 1,082 | 4,501 | 0,039 | 0,986 | 0,345 | -0,024 |
| 0,1 | 0,3 | 1,271 | 0,887 | 0,288 | 1,163 | 0,347 | 0,276 |
| 0,3 | 0,3 | 1,35 | 0,151 | 0,901 | 1,482 | 0,295 | 0,886 |
| 0,5 | 0,3 | 1,38 | 0,057 | 1,591 | 1,51 | 0,115 | 1,5 |
| -0,5 | 0,4 | 1,363 | 0,086 | 1,238 | 1,139 | 0,029 | 0,824 |
| -0,3 | 0,4 | 1,351 | 0,19 | 0,806 | 0,682 | 0,162 | -0,79 |
| -0,1 | 0,4 | 1,297 | 0,87 | 0,318 | 0,851 | 0,297 | -0,274 |
| 0 | 0,4 | 1,045 | 6,312 | 0,018 | 0,975 | 0,328 | -0,043 |
| 0,1 | 0,4 | 1,257 | 1,39 | 0,218 | 1,133 | 0,344 | 0,227 |
| 0,3 | 0,4 | 1,353 | 0,245 | 0,714 | 1,383 | 0,31 | 0,688 |
| 0,5 | 0,4 | 1,365 | 0,092 | 1,202 | 1,484 | 0,203 | 1,073 |
| **Training time** | | | **< 35 s** | | | **< 10 s** | |

FIGURE 5 – *Comparison de la moyenne, la variance, du Sharpe et du temps moyen d'entrainement (par expérience) pour EMV, MLE*

La figure 6 illustre les courbes d'apprentissage de la moyenne et de la variance échantillonnales dans un marché stationnaire. On y constate que l'algorithme EMV converge plus rapidement que le MLE, atteignant de bonnes performances dès les premières étapes de l'apprentissage. Cette rapidité de convergence se maintient tant que l'initialisation des paramètres de l'EMV reste dans des plages raisonnables, typiquement de l'ordre de l'unité.




(a) *Courbes d'apprentissage des moyennes échantillonnales des richesses terminales*

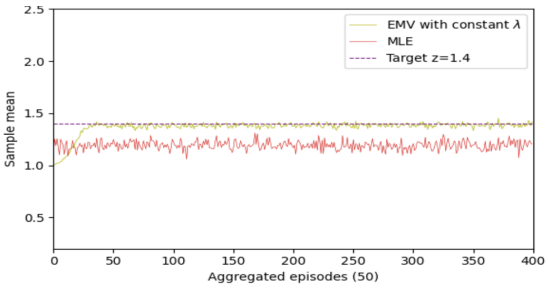



(b) *Courbes d'apprentissage des variances échantillonnales des richesses terminales*

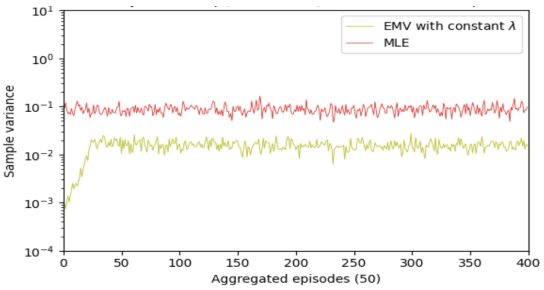
FIGURE 6 – *Courbes d'apprentissage pour EMV et MLE en marché stationnaire ($\mu = -30\%$, $\sigma = 10\%$).*

## 2.5 Marché non-stationnaire

Le marché non stationnaire a été modélisé tel que dans l'article. Comme le montre la figure 7, l'algorithme EMV conserve de bonnes propriétés de convergence, même en environnement non stationnaire. Le MLE (avec clipping de l'allocation), affiche comme dans le cas stationnaire une performance stable d'une simulation à l'autre. Toutefois, il devient encore plus éloigné de la cible $z$ et présente une variance plus élevée que dans le cas stationnaire.




(a) *Courbes d'apprentissage des moyennes échantillonnales des richesses terminales*




(b) *Courbes d'apprentissage des variances échantillonnales des richesses terminales*

FIGURE 7 – *Courbes d'apprentissage pour EMV et MLE en marché non stationnaire ($\mu_0 = -30\%$, $\sigma_0 = 10\%$, $\delta = 0.0001$, $\gamma = 0$).*

## 2.6 Insigths sur les paramètres de l'algorithme EMV : cas du marché stationnaire

Dans cette section, nous analysons les paramètres indépendants $w$, $\phi_1$ et $\phi_2$, obtenus à l'issue de l'exécution de l'algorithme EMV en marché stationnaire. Bien que leurs valeurs numériques puissent varier en fonction de l'initialisation et que nous n'ayons observé aucune convergence vers leurs valeurs théoriques en marché stationnaire, des comportements récurrents émergent, offrant des éclairages sur le fonctionnement de l'agent EMV.

La figure 8 présente l'évolution du multiplicateur de Lagrange $w$ en fonction du drift $\mu$ et de la volatilité $\sigma$. Le multiplicateur de Lagrange $w$ représente le coût d'imposer la contrainte sur l'espérance de la richesse finale. Numériquement, $w$ augmente avec la volatilité : atteindre une cible d'espérance devient plus coûteux dans un marché incertain. Pour une volatilité donnée, $w$ est maximal quand le drift est proche de zéro, car l'actif n'offre alors aucune tendance exploitable, rendant la contrainte plus difficile à satisfaire sans prendre de risque. Enfin, le profil quasi symétrique de $w$ autour de $\mu = 0$ reflète le fait que l'effort d'atteinte d'une espérance cible est similaire pour un drift positif ou négatif, l'agent pouvant inverser le sens de son allocation pour compenser.

La figure 9 présente l'évolution des paramètres d'allocation $\phi_1$ et $\phi_2$ en fonction de $\mu$, $\sigma$ et $\rho$. Comme le montre l'équation 1, le paramètre $\phi_1$ contrôle le niveau de base de la variance des décisions d'allocation, indépendamment du temps. On observe que $\phi_1$ reste la plus part du temps négatif ce qui reflète un comportement naturellement conservateur de l'agent. Cela suggère que l'agent privilégie spontanément des stratégies peu risquées. De plus, on observe que $\phi_1$ diminue lorsque la volatilité de l'actif augmente. Selon l'article, cela s'explique par le fait qu'un marché plus volatil fournit davantage d'informations observables, permettant à l'agent d'apprendre plus efficacement tout en réduisant son besoin d'exploration active, coûteuse en termes de performance et de risque. Par ailleurs, nous pensons qu'un environnement très incertain incite l'agent à adopter un comportement plus prudent, en limitant à la fois la variabilité et l'intensité moyenne de ses décisions pour se protéger du risque accru. Ces deux points de vue

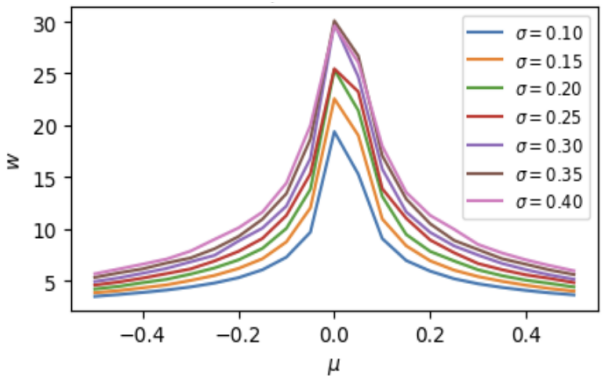
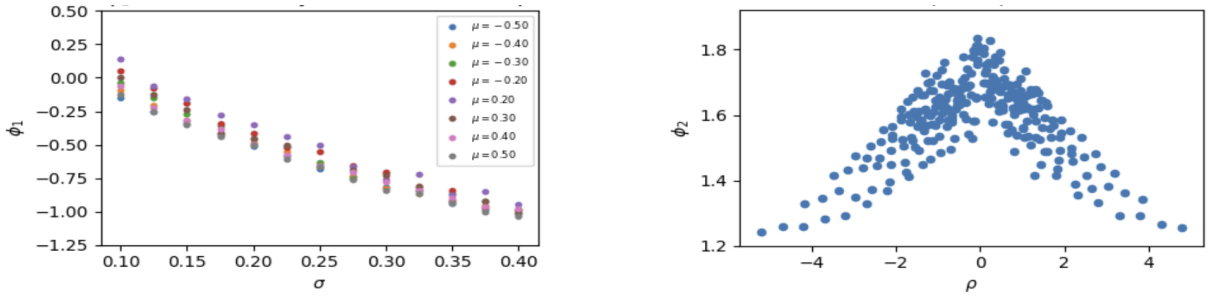



FIGURE 8 – *Evolution de $w$ en fonction de $\mu$ et $\sigma$*

nous semblent complémentaires : une volatilité élevée réduit le besoin d'exploration active, ce qui conduit naturellement à une exploration plus mesurée et sécurisée.

Le paramètre $\phi_2$, censé théoriquement évoluer dans le même sens que $\rho^2$, possède un comportement inverse dans nos simulations. Puisque $\phi_2$ contrôle la vitesse de décroissance de la variance au cours du temps, on pense qu'une baisse de $\phi_2$ avec $\rho^2$ suggère que dans un marché performant, l'incertitude de l'agent décroît beaucoup plus lentement dans le temps. En d'autres termes, plus le marché est attractif, plus l'agent garde la porte ouverte à des décisions flexibles plus longtemps dans l'horizon. Il retarde le gel de sa stratégie, car les opportunités sont plus nombreuses. Cela ralentit aussi l'intensité moyenne de l'allocation, confirmant qu'un bon Sharpe ratio n'entraîne pas forcément une prise de risque immédiate.




(a) $\phi_1$ *en fonction de* $\mu$ *et* $\sigma$ (b) $\phi_2$ *en fonction de* $\rho$

FIGURE 9 – $\phi_1$ *et* $\phi_2$ *en fonction de* $\rho$

## 2.7 Cas de l'exploration décroissante

Un schéma d'exploration décroissant est souvent souhaitable en apprentissage par renforcement (RL). L'article propose une loi d'exploration décroissante définie, à l'épisode $k$, par :

$$\lambda_k = \lambda_0 \left(1 - \exp\left(\frac{\beta(k-M)}{M}\right)\right) \tag{2}$$

L'implémentation de cette loi avec $\beta = 200$, comme suggéré dans l'article, a conduit systématiquement à des pics à la fin de la courbe d'apprentissage, comme l'illustre la figure 10. Ce facteur $\beta$ contrôle l'amplitude de ces pics, la moyenne de la politique d'allocation étant inversement liée à $\lambda_k$. Nous comprenons ainsi pourquoi un tel choix de $\beta$ (grand) a été fait dans l'article. Cependant, pour notre implémentation de l'EMV, cette valeur de $\beta$ n'était peut-être

pas suffisamment grande. En l'augmentant, nous avons réussi à réduire considérablement le pic final de la courbe d'apprentissage. Néanmoins, aucune amélioration notable du ratio de Sharpe n'a été observée.

Nous avons donc exploré une autre loi de décroissance, définie par :

$$\lambda_k = \frac{\lambda_0}{1 + \beta' \frac{k}{M}} \tag{3}$$

Dans cette nouvelle formulation, nous avons fixé $\beta' = 0.33$. Cette modification a permis d'éviter le pic final tout en améliorant le ratio de Sharpe (de 3.134 à 3.246). Des résultats similaires ont été observés dans des conditions de marché instationnaires.

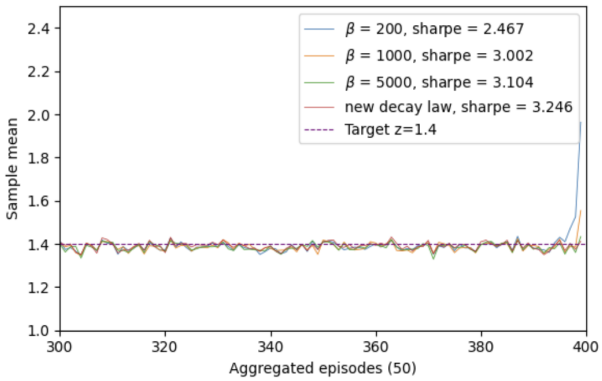
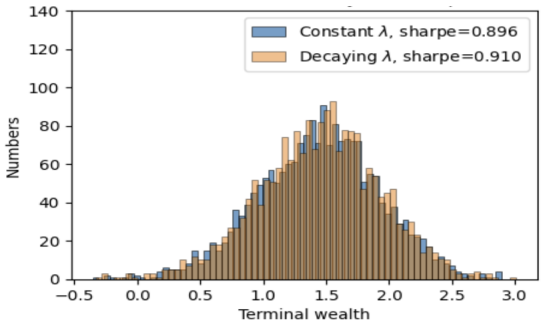
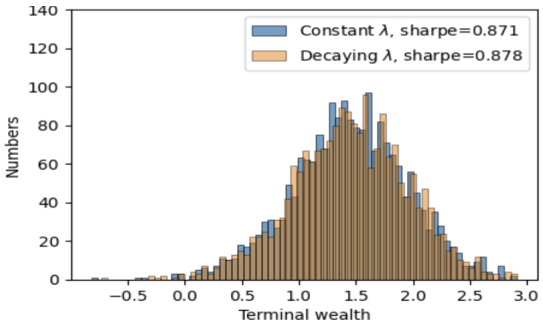



FIGURE 10 – *Exploration décroissante dans le cas du marché stationnaire ($\mu = -30\%$, $\sigma = 10\%$). Le sharpe pour $\lambda$ constant est* 3.134

La figure 11 présente l'histogramme de la richesse terminale dans des environnements de marché stationnaire et instationnaire (ce dernier étant légèrement plus volatile avec une tendance moins marquée que dans les exemples précédents) et pour une cible plus grande ($z = 1.5$). Nos résultats numériques indiquent que le choix de cette loi de décroissance de l'exploration entraîne une légère amélioration des performances de l'algorithme EMV.




(a) *Marché stationnaire : $\mu = 20\%$, $\sigma = 20\%$*




(b) *Marché instationnaire : $\mu_0 = 20\%$, $\sigma_0 = 20\%$*

FIGURE 11 – *Histogramme des 2000 dernières valeurs de la richesse terminale pour EMV avec $\lambda$ constant et $\lambda$ décroissant. La target est $z = 1.5$*