Multi-Strikes Multi-Maturies Options Market Making

Zhou Fang[1]

[1]Department of Mathematics, The University of Texas at Austin

April 24, 2023

Abstract

Market making of options with different maturities and strikes is a challenging problem due to its high dimensional nature. In this paper, we propose a novel approach that combines a stochastic policy and reinforcement learning-inspired techniques to determine the optimal policy for posting bid-ask spreads for an options market maker who trades options with different maturities and strikes. When the arrival of market orders is linearly inverse to the spreads, the optimal policy is normally distributed.

1 Introduction

The market maker is a key player in the financial market, they buy

The option market maker is a key player in the financial market, serving as a dealer that buys and sells options. Their role is to provide liquidity to the market by offering prices for options and by taking positions to manage the risk associated with their trades. However, the task of setting optimal prices for options with different strikes and maturities is highly non-trivial. In this paper, we address the challenging problem of market-making for options with multiple strikes and maturities.

The studies of market making start from (Grossman & Miller, 1988), and (Ho & Stoll, 1981) in the 1980s. The idea in the (Ho & Stoll, 1981) was revived in (Avellaneda & Stoikov, 2008), which inspires a large number of subsequent literature in market making. There are two influential papers(Cartea, Jaimungal, & Ricci, 2014), (Cartea, Donnelly, & Jaimungal, 2017). Other papers include (Baldacci, Bergault, & Guéant, 2021), (Bergault, Evangelista, Guéant, & Vieira, 2021), (Stoikov & Sağlam, 2009).

There are some works that use reinforcement learning in market making, such as (Spooner & Savani, 2020), (Sadighian, 2020), (Beysolow II & Beysolow II, 2019), (Ganesh et al., 2019). Those papers are more engineering-oriented, and relatively simple models are assumed.

The use of stochastic policy is inspired by the reinforcement learning literature, and its first application in financial mathematics literature is in the (Wang, Zariphopoulou, & Zhou, 2020), and (Wang & Zhou, 2020) for portfolio management problems. The stochastic policy can improve the robustness, and balance the exploitation and exploration. In (Jia & Zhou, 2022a), and (Jia & Zhou, 2022b), the authors propose a unified policy evaluation and policy gradient framework that extend the previous two papers.

2 Model

The market maker will give bid-ask quotes on options ranging over multiple strikes with multiple maturity dates. The followings are some notations, $\epsilon_t^a(i,j)$ and $\epsilon_t^b(i,j)$ are the spreads for asking and bidding quotes posted on the option with strike prices K_i, and maturity date T_j at time t, denote the midprice of the option $\mathcal{O}^{i,j}$ as $\mathcal{O}^{i,j}(t, S, \sigma(i,j))$, where S, and $\sigma(i,j)$ is the mid-price of the asset and implied volatility. In order to simplify the model, we assume the implied volatility will stay constant over the entire trading period, which is a very short time period.

This model assumes the arrival of Market orders (MOs) for option $\mathcal{O}^{i,j}$ as Poisson processes with intensities $\lambda_t^a(i,j)$, $\lambda_t^b(i,j)$, where the intensities are functions of spreads $\epsilon_t^a(i,j)$, $\epsilon_t^b(i,j)$. Denote $N_t^+(i,j)$, and $N_t^-(i,j)$

as the counting process for the buy and sell MOs for option $\mathcal{O}^{i,j}$ respectively. Thus, the inventory for option $\mathcal{O}^{i,j}$ is

$$dq_t^{i,j} = dN_t^+(i,j) - dN_t^-(i,j) \tag{1}$$

Assume the underlying asset has the following dynamics

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t \tag{2}$$

Let ϵ_t be the half-spread of the underlying asset. Since with an alpha signal, the options market-maker could slightly take a directional bet, which means it doesn't need to hedge all of the existing inventory. Consider the following,

$$\Delta_t = \sum_i \sum_j \partial_S \mathcal{O}^{i,j}(t, S_t, \sigma(i,j)) q_t^{i,j} \tag{3}$$

In order to simplify the notation, denote $\mathcal{O}^{i,j} = \mathcal{O}^{i,j}(t, S_t, \sigma(i,j))$
Thus, the cash process becomes

$$\begin{aligned} dC_t = & \sum_i \sum_j \Big[\epsilon_t^b(i,j) dN_t^+(i,j) + \epsilon_t^a(i,j) dN_t^-(i,j) - \mathcal{O}^{i,j} dq_t^{i,j} \Big] \\ & + S_t d(\Delta_t) + d\langle \Delta, S \rangle_t \end{aligned} \tag{4}$$

Then the wealth has the following dynamics,

$$\begin{aligned} dX_t &= dC_t - d(\Delta_t S_t) + \sum_i \sum_j d(\mathcal{O}^{i,j} q_t^{i,j}) \\ &= \sum_i \sum_j \Big[\epsilon_t^b(i,j) dN_t^+(i,j) + \epsilon_t^a(i,j) dN_t^-(i,j) - \mathcal{O}^{i,j} dq_t^{i,j} \Big] - \Delta_t dS_t + \sum_i \sum_j \mathcal{O}^{i,j} dq_t^{i,j} + q_t^{i,j} d\mathcal{O}^{i,j} \\ &= \sum_i \sum_j \Big[\epsilon_t^b(i,j) dN_t^+(i,j) + \epsilon_t^a(i,j) dN_t^-(i,j) + \big(\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS} \mathcal{O}^{i,j} \big) q_t^{i,j} dt \Big] \end{aligned} \tag{5}$$

Since the controls we use are $\epsilon_t(i,j)$ is high dimensional, it is very hard to solve mathematically, instead, we use reinforcement learning. Let $\boldsymbol{\epsilon}_t = (\boldsymbol{\epsilon}_t^a(i,j), \boldsymbol{\epsilon}_t^b(i,j))$, $\boldsymbol{q}_t = (q_t^{i,j})$, and $\boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, \boldsymbol{q})$ be the probability density that at time t, the control is $\boldsymbol{\epsilon}_t$ given the inventory $\boldsymbol{q}$,

Given a policy $\boldsymbol{\pi}$, let $\boldsymbol{q}_t^{\boldsymbol{\pi}}$ denote the inventory process under the policy $\boldsymbol{\pi}$, and the initial condition at time t is $\boldsymbol{q}_t^{\boldsymbol{\pi}} = \boldsymbol{q}$. Define the value function under policy $\boldsymbol{pi}$ to be

$$\begin{aligned} V^{\boldsymbol{\pi}}(t, \boldsymbol{q}) = \mathbb{E}\Big[& \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \sum_i \sum_j \big[\epsilon_u^b(i,j) dN_u^+(i,j) + \epsilon_u^a(i,j) dN_u^-(i,j) \big] d\boldsymbol{\epsilon}_u \\ & + \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \Big(\sum_i \sum_j \big(\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS} \mathcal{O}^{i,j} \big) q_u^{i,j} - \gamma \log \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \Big) d\boldsymbol{\epsilon}_u du \;\Big|\; \boldsymbol{q}_t^{\boldsymbol{\pi}} = \boldsymbol{q} \Big] \end{aligned} \tag{6}$$

Then the value function under optimal policy is as follows,

$$\begin{aligned} V(t, \boldsymbol{q}) = \max_{\boldsymbol{\pi}} \mathbb{E}\Big[& \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \sum_i \sum_j \big[\epsilon_u^b(i,j) dN_u^+(i,j) + \epsilon_u^a(i,j) dN_u^-(i,j) \big] d\boldsymbol{\epsilon}_u \\ & + \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \Big(\sum_i \sum_j \big(\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS} \mathcal{O}^{i,j} \big) q_u^{i,j} - \gamma \log \pi(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \Big) d\boldsymbol{\epsilon}_u du \;\Big|\; \boldsymbol{q}_t^{\boldsymbol{\pi}} = \boldsymbol{q} \Big] \end{aligned} \tag{7}$$

Dynamic Programming

For function $V(t,\boldsymbol{q})$, consider the following derivation, where $\Delta t \to 0$

$$
\begin{aligned}
V(t,\boldsymbol{q}) = &\max_{\boldsymbol{\pi}} \mathbb{E}\Big[\int_t^{t+\Delta t}\int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}})\sum_i\sum_j \big[\epsilon_u^b(i,j)dN_u^+(i,j)+\epsilon_u^a(i,j)dN_u^-(i,j)\big]d\boldsymbol{\epsilon}_u \\
&+\int_t^{t+\Delta t}\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}})\Big(\sum_i\sum_j\big(\partial_t\mathcal{O}^{i,j}+\frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j}\big)q_u^{i,j}-\gamma\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}})\Big)d\boldsymbol{\epsilon}_u du \\
&+V(t+\Delta t,\boldsymbol{q}_t^{\boldsymbol{\pi}}+\Delta\boldsymbol{q}_t^{\boldsymbol{\pi}})\ \Big|\ \boldsymbol{q}_t^{\boldsymbol{\pi}}=\boldsymbol{q}\Big] \\
=&\max_{\boldsymbol{\pi}} \mathbb{E}\Big[\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}_t^{\boldsymbol{\pi}})\sum_i\sum_j\big[\epsilon_t^b(i,j)dN_t^+(i,j)+\epsilon_t^a(i,j)dN_t^-(i,j)\big]d\boldsymbol{\epsilon}_t\Delta t \\
&+\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}_t^{\boldsymbol{\pi}})\Big(\sum_i\sum_j\big(\partial_t\mathcal{O}^{i,j}+\frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j}\big)q_t^{i,j}-\gamma\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}_t^{\boldsymbol{\pi}})\Big)d\boldsymbol{\epsilon}_t\Delta t \\
&+V(t+\Delta t,\boldsymbol{q}_t^{\boldsymbol{\pi}}+\Delta\boldsymbol{q}_t^{\boldsymbol{\pi}})\ \Big|\ \boldsymbol{q}_t^{\boldsymbol{\pi}}=\boldsymbol{q}\Big] \qquad (8)\\
=&\max_{\boldsymbol{\pi}}\Big\{\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}_t^{\boldsymbol{\pi}})\sum_i\sum_j\big[\epsilon_t^b(i,j)\lambda_t^+(i,j)+\epsilon_t^a(i,j)\lambda_t^-(i,j)\big]d\boldsymbol{\epsilon}_t\Delta t \\
&+\sum_i\sum_j\big(\partial_t\mathcal{O}^{i,j}+\frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j}\big)q_t^{i,j}-\gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}_t^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}_t^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_t\Delta t \\
&+\mathbb{E}\Big[V(t+\Delta t,\boldsymbol{q}_t^{\boldsymbol{\pi}}+\Delta\boldsymbol{q}_t^{\boldsymbol{\pi}})\ \Big|\ \boldsymbol{q}_t^{\boldsymbol{\pi}}=\boldsymbol{q}\Big]\Big\} \qquad (9)
\end{aligned}
$$

Since the general Ito formula for $V(t+\Delta t,\boldsymbol{q}_t+\Delta\boldsymbol{q}_t)=V(t+\Delta t,q_t^{1,1}+\Delta q_t^{1,1},...,q_t^{m,n}+\Delta q_t^{m,n})$ is as follows,

$$
\begin{aligned}
&V(t+\Delta t,q_t^{1,1}+\Delta q_t^{1,1},...,q_t^{m,n}+\Delta q_t^{m,n}) \\
=&V(t+\Delta t,q_t^{1,1},...,q_t^{m,n})\prod_{(i,j)}(1-dN_t^+(i,j))(1-dN_t^-(i,j)) \\
&+\sum_i\sum_j\Big[V(t+\Delta t,q_t^{1,1},...,q_t^{i,j}+1,...,q_t^{m,n})dN_t^+(i,j)+V(t+\Delta t,q_t^{1,1},...,q_t^{i,j}-1,...,q_t^{m,n})dN_t^-(i,j)\Big] \\
=&\big[V(t,q_t^{1,1},...,q_t^{m,n})+\partial_tV(t,q_t^{1,1},...,q_t^{m,n})\Delta t\big]\prod_{(i,j)}(1-dN_t^+(i,j))(1-dN_t^-(i,j)) \qquad (10)\\
&+\sum_i\sum_j\Big[V(t,q_t^{1,1},...,q_t^{i,j}+1,...,q_t^{m,n})dN_t^+(i,j)+V(t,q_t^{1,1},...,q_t^{i,j}-1,...,q_t^{m,n})dN_t^-(i,j)\Big]
\end{aligned}
$$

Notice that the above Ito formula is under the situation that $\boldsymbol{q}_t$ is known, which means that the above Ito formula is only valid when the $\boldsymbol{\epsilon}_t$ is already determined. For the conditional expectation of $\mathbb{E}[V(t+\Delta t,\boldsymbol{q}_t^{\boldsymbol{\pi}}+\Delta\boldsymbol{q}_t^{\boldsymbol{\pi}})|q_t^{\boldsymbol{\pi}}=\boldsymbol{q}]$, one should average over all possibilities, then we have the following derivation,

$$
\begin{aligned}
&\mathbb{E}\big[V(t+\Delta t,\boldsymbol{q}_t^{\boldsymbol{\pi}}+\Delta\boldsymbol{q}_t^{\boldsymbol{\pi}})\mid\boldsymbol{q}_t^{\boldsymbol{\pi}}=\boldsymbol{q}\big] \\
=&V(t,\boldsymbol{q})-\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\sum_i\sum_j\big(\lambda_t^+(i,j)+\lambda_t^-(i,j)\big)V(t,\boldsymbol{q})d\boldsymbol{\epsilon}_t\Delta t+\partial_tV(t,\boldsymbol{q})\Delta t \qquad (11)\\
&+\int_{\epsilon_t}\boldsymbol{\pi}(\epsilon_t|t,\boldsymbol{q})\sum_i\sum_j\Big[V(t,\boldsymbol{q}+\Delta_{i,j})\lambda_t^+(i,j)+V(t,\boldsymbol{q}-\Delta_{i,j})\lambda_t^-(i,j)\Big]d\boldsymbol{\epsilon}_t\Delta t
\end{aligned}
$$

Thus, the HJB equation will be, (to simplify the form of the equation,

$$\max_{\boldsymbol{\pi}}\Bigg\{\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\sum_i\sum_j\lambda_t^+(i,j)\Big[\epsilon_t^b-V(t,\boldsymbol{q})+V(t,\boldsymbol{q}+\Delta_{i,j})\Big]+\lambda_t^-(i,j)\Big[\epsilon_t^a-V(t,\boldsymbol{q})+V(t,\boldsymbol{q}-\Delta_{i,j})\Big]d\boldsymbol{\epsilon}_t$$
$$-\gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\Bigg\}+\sum_i\sum_j(\partial_t\mathcal{O}^{i,j}+\frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j})q^{i,j}+\partial_tV(t,\boldsymbol{q})=0 \tag{12}$$

Numerical Solution

To get the maximizer $\boldsymbol{\pi}^*$, we apply the calculus of variation. For maximizer $\boldsymbol{\pi}^*$, the following is true

$$0=\int_{\boldsymbol{\epsilon}_t}\delta\boldsymbol{\pi}\sum_i\sum_j\Big[\lambda_t^+(i,j)\big[V(t,\boldsymbol{q}+\Delta_{i,j})-V(t,\boldsymbol{q})+\epsilon_t^b(i,j)\big]$$
$$+\lambda_t^-(i,j)\big[V(t,\boldsymbol{q}-\Delta_{i,j})-V(t,\boldsymbol{q})+\epsilon_t^a(i,j)\big]\Big]d\boldsymbol{\epsilon}_t-\gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}^*\frac{\delta\boldsymbol{\pi}}{\boldsymbol{\pi}^*}d\boldsymbol{\epsilon}_t-\gamma\int_{\boldsymbol{\epsilon}_t}\delta\boldsymbol{\pi}\log\boldsymbol{\pi}^*d\boldsymbol{\epsilon}_t \tag{13}$$

(14)

Since $\boldsymbol{\pi}$ is probability density distribution, then

$$\int_{\boldsymbol{\epsilon}_t}\delta\boldsymbol{\pi}d\boldsymbol{\epsilon}_t=0 \tag{15}$$

Then the above equation becomes

$$0=\int_{\boldsymbol{\epsilon}_t}\delta\boldsymbol{\pi}\Bigg(\sum_i\sum_j\Big[\lambda_t^+(i,j)\big[V(t,\boldsymbol{q}+\Delta_{i,j})-V(t,\boldsymbol{q})+\epsilon_t^b(i,j)\big]$$
$$+\lambda_t^-(i,j)\big[V(t,\boldsymbol{q}-\Delta_{i,j})-V(t,\boldsymbol{q})+\epsilon_t^a(i,j)\big]\Big]-\gamma\log\boldsymbol{\pi}^*(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\Bigg)d\boldsymbol{\epsilon}_t \tag{16}$$

Then The optimal policy is to maximize the quantity inside the above bracket, then it should satisfy the following equation

$$C=\sum_i\sum_j\Big[\lambda_t^+(i,j)\big[V(t,\boldsymbol{q}+\Delta_{i,j})-V(t,\boldsymbol{q})+\epsilon_t^b(i,j)\big]+\lambda_t^-(i,j)\big[V(t,\boldsymbol{q}-\Delta_{i,j})-V(t,\boldsymbol{q})+\epsilon_t^a(i,j)\big]\Big]$$
$$-\gamma\log\boldsymbol{\pi}^*(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}) \tag{17}$$

We assume the following stylized function that describes the relationship between intensity and spreads of option $\mathcal{O}^{i,j}$ to be

$$\lambda_t^+(i,j)=A_{i,j}-B_{i,j}\epsilon_t^b(i,j) \tag{18}$$
$$\lambda_t^-(i,j)=A_{i,j}-B_{i,j}\epsilon_t^a(i,j) \tag{19}$$

then the following derives the optimal policy $\boldsymbol{\pi}^*$

$$\begin{aligned}
\boldsymbol{\pi}^*(\boldsymbol{\epsilon}_t|t,\boldsymbol{q}) &\propto \exp\Big\{\frac{1}{\gamma}\sum_i\sum_j \lambda_t^{\pm}(i,j)\big[V(t,\boldsymbol{q}\pm\Delta_{i,j}) - V(t,\boldsymbol{q}) + \epsilon_t^{a,b}(i,j)\big]\Big\} \\
&= \exp\Big\{\frac{1}{\gamma}\sum_i\sum_j \lambda_t^{\pm}(i,j)\big[V(t,\boldsymbol{q}_t\pm\Delta_{i,j}) - V(t,\boldsymbol{q}) + \epsilon_t^{a,b}(i,j)\big]\Big\} \\
&= \prod_{i,j}\exp\Big\{\frac{1}{\gamma}\big(A_{i,j} - B_{i,j}\epsilon_t^{a,b}(i,j)\big)\big[V(t,\boldsymbol{q}\pm\Delta_{i,j}) - V(t,\boldsymbol{q}) + \epsilon_t^{a,b}(i,j)\big]\Big\} \\
&\propto \prod_{i,j}\exp\Big\{-\frac{B_{i,j}}{\gamma}\epsilon_t^{a,b}(i,j)^2 + \frac{1}{\gamma}\big[A_{i,j} + B_{i,j}\big(V(t,\boldsymbol{q}) - V(t,\boldsymbol{q}\pm\Delta_{i,j})\big)\big]\epsilon_t^{a,b}(i,j)\Big\} \\
&\propto \prod_{i,j}\exp\Big\{-\frac{B_{i,j}}{\gamma}\Big[\epsilon_t^{a,b}(i,j) - \frac{A_{i,j}}{2B_{i,j}} - \frac{1}{2}\big(V(t,\boldsymbol{q}) - V(t,\boldsymbol{q}\pm\Delta_{i,j})\big)\Big]^2\Big\} \\
&\propto \prod_{i,j}\mathcal{N}\Big(\epsilon_t^{a,b}(i,j) \mid \frac{A_{i,j}}{2B_{i,j}} + \frac{1}{2}\big(V(t,\boldsymbol{q}) - V(t,\boldsymbol{q}\pm\Delta_{i,j})\big), \frac{\gamma}{2B_{i,j}}\Big)
\end{aligned} \tag{20}$$

Therefore, one can see that the optimal policy is multi-dimensional Gaussian distribution. To simplify the notation, let

$$\boldsymbol{\mu}(t,\boldsymbol{q},\boldsymbol{\pi}) = \Big(\frac{A_{1,1}}{2B_{1,1}} + \frac{1}{2}\big(V^{\boldsymbol{\pi}}(t,\boldsymbol{q}) - V^{\boldsymbol{\pi}}(t,\boldsymbol{q}\pm\Delta_{1,1})\big), ..., \frac{A_{m,n}}{2B_{m,n}} + \frac{1}{2}\big(V^{\boldsymbol{\pi}}(t,\boldsymbol{q}) - V^{\boldsymbol{\pi}}(t,\boldsymbol{q}\pm\Delta_{m,n})\big)\Big)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{\gamma}{2B_{1,1}} & & & & \\ & \frac{\gamma}{2B_{1,1}} & & & \\ & & \ddots & & \\ & & & \frac{\gamma}{2B_{m,n}} & \\ & & & & \frac{\gamma}{2B_{m,n}} \end{bmatrix}$$

The optimal policy is

$$\boldsymbol{\pi}^* \sim \mathcal{N}(\cdot \mid \boldsymbol{\mu}(t,\boldsymbol{q},\boldsymbol{\pi}^*), \boldsymbol{\Sigma}) \tag{21}$$

Policy Improvement Theorem

Theorem 2.1 (policy improvement theorem). *Given any $\boldsymbol{\pi}$, let the new policy $\boldsymbol{\pi}_{new}$ to be*

$$\boldsymbol{\pi}_{new} \sim \mathcal{N}(\cdot \mid \boldsymbol{\mu}(t,\boldsymbol{q},\boldsymbol{\pi}), \boldsymbol{\Sigma}) \tag{22}$$

then the value function

$$V^{\boldsymbol{\pi}}(t,\boldsymbol{q}) \le V^{\boldsymbol{\pi}_{new}}(t,\boldsymbol{q}) \tag{23}$$

Proof. Let $\boldsymbol{q}_t^{\boldsymbol{\pi}_{new}}$ be the inventory process under policy $\boldsymbol{\pi}_{new}$. Let the initial condition be $\boldsymbol{q}_t^{\boldsymbol{\pi}_{new}} = \boldsymbol{q}$. Then by the Ito formula, we have the following

$$\begin{aligned}
&V^{\boldsymbol{\pi}}(t,\boldsymbol{q}) \\
&= \mathbb{E}\Big[V^{\boldsymbol{\pi}}(s,\boldsymbol{q}_s^{\boldsymbol{\pi}_{new}}) + \int_t^s\int_{\boldsymbol{\epsilon}_u} V^{\boldsymbol{\pi}}(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})\boldsymbol{\pi}_{new}(\epsilon_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})\sum_i\sum_j[dN_u^+(i,j) + dN_u^-(i,j)]d\boldsymbol{\epsilon}_u \\
&- \int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})\sum_i\sum_j\big[V^{\boldsymbol{\pi}}(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}+\Delta_{i,j})dN_u^+(i,j) + V^{\boldsymbol{\pi}}(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}-\Delta_{i,j})dN_u^-(i,j)\big]d\boldsymbol{\epsilon}_u \\
&- \int_t^s \partial_t V(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})du \,\Big|\, \boldsymbol{q}_t^{\boldsymbol{\pi}_{new}} = \boldsymbol{q}\Big]
\end{aligned} \tag{24}$$

which becomes

$$
\begin{aligned}
&V^{\boldsymbol{\pi}}(t,\boldsymbol{q})\\
&=\mathbb{E}\Big[V^{\boldsymbol{\pi}}(s,\boldsymbol{q}_s^{\boldsymbol{\pi}_{new}})\;\Big|\;\boldsymbol{q}_t^{\boldsymbol{\pi}_{new}}=\boldsymbol{q}\Big]+\int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}_{new}(\epsilon_u|\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})\sum_i\sum_j[\lambda_u^+(i,j)+\lambda_u^-(i,j)]V^{\boldsymbol{\pi}}(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})d\boldsymbol{\epsilon}_u du\\
&-\int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})\sum_i\sum_j\big[V^{\boldsymbol{\pi}}(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}+\Delta_{i,j})\lambda_u^+(i,j)+V^{\boldsymbol{\pi}}(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}-\Delta_{i,j})\lambda^-(i,j)\big]d\boldsymbol{\epsilon}_u du\\
&-\int_t^s\partial_t V(u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})du
\end{aligned}
\tag{25}
$$

For a given policy $\boldsymbol{\pi}$, we have

$$
\begin{aligned}
&\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\sum_i\sum_j\lambda_t^+(i,j)\Big[\epsilon_t^b-V^{\boldsymbol{\pi}}(t,\boldsymbol{q})+V^{\boldsymbol{\pi}}(t,\boldsymbol{q}+\Delta_{i,j)}\Big]+\lambda_t^-(i,j)\Big[\epsilon_t^a-V^{\boldsymbol{\pi}}(t,\boldsymbol{q})+V^{\boldsymbol{\pi}}(t,\boldsymbol{q}-\Delta_{i,j})\Big]d\boldsymbol{\epsilon_t}\\
&-\gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})+\sum_i\sum_j(\partial_t\mathcal{O}^{i,j}+\frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j})q^{i,j}+\partial_t V^{\boldsymbol{\pi}}(t,\boldsymbol{q})=0
\end{aligned}
\tag{26}
$$

Based on the construction of $\boldsymbol{\pi}_{new}$, by the same calculus of variation arguments, the $\boldsymbol{\pi}_{new}$ is maximizer of the following quantity

$$
\begin{aligned}
&\max_{\tilde{\boldsymbol{\pi}}}\Big\{\int_{\boldsymbol{\epsilon}_t}\tilde{\boldsymbol{\pi}}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\Big(\sum_i\sum_j\lambda_t^+(i,j)\Big[\epsilon_t^b-V^{\boldsymbol{\pi}}(t,\boldsymbol{q})+V^{\boldsymbol{\pi}}(t,\boldsymbol{q}+\Delta_{i,j)}\Big]\\
&+\lambda_t^-(i,j)\Big[\epsilon_t^a-V^{\boldsymbol{\pi}}(t,\boldsymbol{q})+V^{\boldsymbol{\pi}}(t,\boldsymbol{q}-\Delta_{i,j)}\Big]\Big)d\boldsymbol{\epsilon}_t-\gamma\int_{\boldsymbol{\epsilon}_t}\tilde{\boldsymbol{\pi}}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\log\tilde{\boldsymbol{\pi}}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\Big\}
\end{aligned}
\tag{27}
$$

Then we have

$$
\begin{aligned}
&\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\Big(\sum_i\sum_j\lambda_t^+(i,j)\Big[\epsilon_t^b-V^{\boldsymbol{\pi}}(t,\boldsymbol{q})+V^{\boldsymbol{\pi}}(t,\boldsymbol{q}+\Delta_{i,j)}\Big]\\
&+\lambda_t^-(i,j)\Big[\epsilon_t^a-V^{\boldsymbol{\pi}}(t,\boldsymbol{q})+V^{\boldsymbol{\pi}}(t,\boldsymbol{q}-\Delta_{i,j)}\Big]\Big)d\boldsymbol{\epsilon}_t-\gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\log\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_t|t,\boldsymbol{q})\\
&+\sum_i\sum_j(\partial_t\mathcal{O}^{i,j}+\frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j})q^{i,j}+\partial_t V^{\boldsymbol{\pi}}(t,\boldsymbol{q})\geq 0
\end{aligned}
\tag{28}
$$

Then there is

$$
\begin{aligned}
&V^{\boldsymbol{\pi}}(t,\boldsymbol{q})\\
&\leq\mathbb{E}\Big[V^{\boldsymbol{\pi}}(s,\boldsymbol{q}_s^{\boldsymbol{\pi}_{new}})+\int_t^s\sum_i\sum_j(\partial_t\mathcal{O}^{i,j}+\frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j})q_u^{i,j,\boldsymbol{\pi}_{new}}\\
&+\int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})\sum_i\sum_j\big[\lambda_u^+(i,j)\epsilon_u^b(i,j)+\lambda_u^-(i,j)\epsilon_u^a(i,j)\big]d\boldsymbol{\epsilon}_u du\\
&-\gamma\int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})\log\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,\boldsymbol{q}_u^{\boldsymbol{\pi}_{new}})d\boldsymbol{\epsilon}_u du\;\Big|\;\boldsymbol{q}_t^{\boldsymbol{\pi}_{new}}=\boldsymbol{q}\Big]
\end{aligned}
\tag{29}
$$

Set $s=T$, then $V^{\boldsymbol{\pi}}(T,\boldsymbol{q}_T^{\boldsymbol{\pi}_{new}}) = V^{\boldsymbol{\pi}_{new}}(T,\boldsymbol{q}_T^{\boldsymbol{\pi}_{new}})$ then the equation (75) becomes

$$
\begin{aligned}
&V^{\boldsymbol{\pi}}(t,\boldsymbol{q}) \\
&\leq \mathbb{E}\Big[V^{\boldsymbol{\pi}_{new}}(T,\boldsymbol{q}_T^{\boldsymbol{\pi}_{new}}) + \int_t^T \sum_i \sum_j (\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS}\mathcal{O}^{i,j}) q_u^{i,j,\boldsymbol{\pi}_{new}} \\
&+ \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}) \sum_i \sum_j \big[\lambda_u^+(i,j)\epsilon_u^b(i,j) + \lambda_u^-(i,j)\epsilon_u^a(i,j)\big] d\boldsymbol{\epsilon}_u du \\
&- \gamma \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}) \log \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}) d\boldsymbol{\epsilon}_u du \;\Big|\; \boldsymbol{q}_t^{\boldsymbol{\pi}_{new}} = \boldsymbol{q}\Big] \\
&= \mathbb{E}\Big[V^{\boldsymbol{\pi}_{new}}(T,\boldsymbol{q}_T^{\boldsymbol{\pi}_{new}}) + \int_t^T \sum_i \sum_j (\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS}\mathcal{O}^{i,j}) q_u^{i,j,\boldsymbol{\pi}_{new}} \\
&+ \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}) \sum_i \sum_j \big[dN_u^+(i,j)\epsilon_u^b(i,j) + dN_u^-(i,j)\epsilon_u^a(i,j)\big] d\boldsymbol{\epsilon}_u du \\
&- \gamma \int_t^T \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}) \log \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}) d\boldsymbol{\epsilon}_u du \;\Big|\; \boldsymbol{q}_t^{\boldsymbol{\pi}_{new}} = \boldsymbol{q}\Big] \\
&= V^{\boldsymbol{\pi}_{new}}(t,\boldsymbol{q})
\end{aligned} \tag{30}
$$

□

Martingale Loss

$$
\begin{aligned}
V^{\boldsymbol{\pi}}(t,\boldsymbol{q}) = \; &\mathbb{E}\Big[\int_t^s \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \sum_i \sum_j \big[\epsilon_u^b(i,j) dN_u^+(i,j) + \epsilon_u^a(i,j) dN_u^-(i,j)\big] d\boldsymbol{\epsilon}_u \\
&+ \int_t^s \sum_i \sum_j \big(\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS}\mathcal{O}^{i,j}\big) q_u^{i,j,\boldsymbol{\pi}} du - \gamma \int_t^s \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \log \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) d\boldsymbol{\epsilon}_u du \\
&+ V^{\boldsymbol{\pi}}(s, \boldsymbol{q}_s^{\boldsymbol{\pi}}) \;\Big|\; \boldsymbol{q}_t^{\boldsymbol{\pi}} = \boldsymbol{q}\Big]
\end{aligned} \tag{31}
$$

Then we have

$$
\begin{aligned}
0 = \; &\mathbb{E}\Big[\frac{1}{s-t}\int_t^s \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \sum_i \sum_j \big[\epsilon_u^b(i,j) dN_u^+(i,j) + \epsilon_u^a(i,j) dN_u^-(i,j)\big] d\boldsymbol{\epsilon}_u \\
&+ \frac{1}{s-t}\int_t^s \sum_i \sum_j \big(\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS}\mathcal{O}^{i,j}\big) q_u^{i,j,\boldsymbol{\pi}} du - \frac{1}{s-t}\gamma \int_t^s \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) \log \boldsymbol{\pi}(\boldsymbol{\epsilon}_u | u, \boldsymbol{q}_u^{\boldsymbol{\pi}}) d\boldsymbol{\epsilon}_u du \\
&+ \frac{V^{\boldsymbol{\pi}}(s, \boldsymbol{q}_s^{\boldsymbol{\pi}}) - V^{\boldsymbol{\pi}}(t, \boldsymbol{q}_t^{\boldsymbol{\pi}})}{s-t} \;\Big|\; \boldsymbol{q}_t^{\boldsymbol{\pi}} = \boldsymbol{q}_t\Big]
\end{aligned} \tag{32}
$$

When $s \to t$, and we parametrize the value function under policy $\boldsymbol{\pi}$ as $V_\theta^{\boldsymbol{\pi}}$, then we can define the temporal difference in continuous-time as

$$
\begin{aligned}
\delta_t^\theta = \; &\mathbb{E}\Big[\frac{V_\theta^{\boldsymbol{\pi}}(s, \boldsymbol{q}_s^{\boldsymbol{\pi}}) - V_\theta^{\boldsymbol{\pi}}(t, \boldsymbol{q}_t^{\boldsymbol{\pi}})}{s-t} \;\Big|\; \boldsymbol{q}_t^{\boldsymbol{\pi}} = \boldsymbol{q}_t\Big] + \int_{\boldsymbol{\epsilon}_t} \boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, \boldsymbol{q}_t^{\boldsymbol{\pi}}) \sum_i \sum_j \big[\epsilon_t^b(i,j) dN_t^+(i,j) + \epsilon_t^a(i,j) dN_t^-(i,j)\big] d\boldsymbol{\epsilon}_t \\
&+ \sum_i \sum_j \big(\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2 \partial_{SS}\mathcal{O}^{i,j}\big) q_t^{i,j,\boldsymbol{\pi}} dt - \gamma \int_{\boldsymbol{\epsilon}_t} \boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, \boldsymbol{q}_t^{\boldsymbol{\pi}}) \log \boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, \boldsymbol{q}_t^{\boldsymbol{\pi}}) d\boldsymbol{\epsilon}_t
\end{aligned} \tag{33}
$$

So we need to minimize the following loss function

$$\mathbf{ML}(\theta) = \frac{1}{2}\mathbb{E}\big[\int_0^T |\delta_t^\theta|^2 dt\big] \tag{34}$$

Using Monte Carlo method, given the policy $\boldsymbol{\pi}$, there are sample paths, $\mathcal{D} = \{(t_k, \boldsymbol{q}_{t_k}^d)_{k=1}^K\}_{d=1}^D$. Then the discrete version of the loss function to be

$$\begin{aligned}\widehat{\mathbf{ML}}(\theta) = \frac{1}{2}\sum_{\mathcal{D}}\sum_{k=0}^{K-1}\Bigg(&\frac{V_\theta^{\boldsymbol{\pi}}(t_{k+1}, \boldsymbol{q}_{k+1}^d) - V_\theta^{\boldsymbol{\pi}}(t_k, \boldsymbol{q}_k^d)}{\Delta t}\\ &+ \int_{\boldsymbol{\epsilon}_{t_k}} \boldsymbol{\pi}(\boldsymbol{\epsilon}_{t_k}|t_k, \boldsymbol{q}_{t_k}^d)\sum_i\sum_j \big[\epsilon_{t_k}^b(i,j)\Delta N_{t_k}^+(i,j) + \epsilon_{t_k}^a(i,j)\Delta N_{t_k}^-(i,j)\big]d\boldsymbol{\epsilon}_{t_k}\\ &+ \sum_i\sum_j \big(\partial_t \mathcal{O}^{i,j} + \frac{1}{2}\sigma^2\partial_{SS}\mathcal{O}^{i,j}\big)q_{t_k}^{i,j,d}dt - \gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_{t_k}|t_k, \boldsymbol{q}_{t_k}^d)\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_{t_k}|t_k, \boldsymbol{q}_{t_k}^d)d\boldsymbol{\epsilon}_t\Bigg)^2\Delta t\end{aligned} \tag{35}$$

The following is a summary of the training process

Algorithm 1 EMM: Exploratory Market Making

Require: Initialize hyperparameters
  for l = 1 to L **do**
    for m = 1 to M **do**
      Generate one sample path $\mathcal{D} = \{(t_k, \boldsymbol{q}_{t_k})_{k=0}^K\}$ under policy $\boldsymbol{\pi}^\phi$
      Compute $\widehat{\mathbf{ML}}(\theta)$
      Updates $\theta \leftarrow \theta - \alpha\nabla_\theta\widehat{\mathbf{ML}}(\theta)$
    end for
    Update $\boldsymbol{\pi}^\phi \leftarrow \mathcal{N}\Big(\boldsymbol{\epsilon}\ \big|\big(\frac{A_{i,j}}{2B_{i,j}} + \frac{1}{2}\big[V_\theta^{\boldsymbol{\pi}}(t,\boldsymbol{q}) - V_\theta^{\boldsymbol{\pi}}(t,\boldsymbol{q}\pm\Delta_{i,j})\big]\big), \Sigma\Big)$
  end for

References

Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224.

Baldacci, B., Bergault, P., & Guéant, O. (2021). Algorithmic market making for options. *Quantitative Finance*, *21*(1), 85–97.

Bergault, P., Evangelista, D., Guéant, O., & Vieira, D. (2021). Closed-form approximations in multi-asset market making. *Applied Mathematical Finance*, *28*(2), 101–142.

Beysolow II, T., & Beysolow II, T. (2019). Market making via reinforcement learning. *Applied Reinforcement Learning with Python: With OpenAI Gym, Tensorflow, and Keras*, 77–94.

Cartea, Á., Donnelly, R., & Jaimungal, S. (2017). Algorithmic trading with model uncertainty. *SIAM Journal on Financial Mathematics*, *8*(1), 635–671.

Cartea, Á., Jaimungal, S., & Ricci, J. (2014). Buy low, sell high: A high frequency trading perspective. *SIAM Journal on Financial Mathematics*, *5*(1), 415–444.

Ganesh, S., Vadori, N., Xu, M., Zheng, H., Reddy, P., & Veloso, M. (2019). Reinforcement learning for market making in a multi-agent dealer market. *arXiv preprint arXiv:1911.05892*.

Grossman, S. J., & Miller, M. H. (1988). Liquidity and market structure. *the Journal of Finance*, *43*(3), 617–633.

Ho, T., & Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial economics*, *9*(1), 47–73.

Jia, Y., & Zhou, X. Y. (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, *23*(154), 1–55.

Jia, Y., & Zhou, X. Y. (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, *23*(154), 1–55.

Sadighian, J. (2020). Extending deep reinforcement learning frameworks in cryptocurrency market making. *arXiv preprint arXiv:2004.06985*.

Spooner, T., & Savani, R. (2020). Robust market making via adversarial reinforcement learning. *arXiv preprint arXiv:2003.01820*.

Stoikov, S., & Sağlam, M. (2009). Option market making under inventory risk. *Review of Derivatives Research*, *12*, 55–79.

Wang, H., Zariphopoulou, T., & Zhou, X. Y. (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *The Journal of Machine Learning Research*, *21*(1), 8145–8178.

Wang, H., & Zhou, X. Y. (2020). Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, *30*(4), 1273–1308.