OTC market making

Zhou Fang[1]

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

April 24, 2023

Abstract

The over-the-counter (OTC) market is known for its unique characteristic of allowing market makers to set different bid-ask spreads based on the size of the order. However, the market-making problems in the OTC market is a challenging high-dimensional stochastic control problems. This paper proposes a stochastic policy approach for setting bid-ask prices and using reinforcement learning to optimize the strategy. Under some stylized assumptions, we demonstrate that the optimal stochastic policy follows a normal distribution.

1 Model

an OTC market-maker will set quotes on different price ladders, and can also choose whether or not to hedge part of its inventory by externalization. Let the dynamics of the underlying asset be

$$\frac{dS_t}{S_t} = \sigma dW_t \tag{1}$$

Modeling the successful deals of size z_k as Poisson processes $N_t^+(k)$, and $N_t^-(k)$, and denote the intensity of those Poisson processes as $\lambda^+(k)$, and $\lambda^-(k)$. Define a function $h(q)$ as follows

$$h(q) = \begin{cases} 0 & q < Q \\ 1 & q \geq Q \end{cases}$$

Let $\boldsymbol{\epsilon}_t = \left(\epsilon_t^{\pm}(k)\right)_{k=1}^N$ be the bid-ask spreads posted by the market maker at time t, and let $\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t, S, q)$ be the probability density for posting spreads $\boldsymbol{\epsilon}_t$. If the market maker posts the bid-ask spreads at time t as $\boldsymbol{\epsilon}_t$, then the inventory has the following dynamics

$$dq_t = \sum_{k=1}^N z_k \big(dN_t^+(k) - dN_t^-(k)\big) \tag{2}$$

If the market maker chooses to externalize dq_t inventory, then she needs to pay an additional transaction fee δ The wealth process is

$$dX_t = \sum_{k=1}^N z_k \left[\epsilon_t^b(k) dN_t^+(k) + \epsilon_t^a(k) dN_t^-(k)\right] + d(q_t S_t) - \delta h(q_t) dq_t \tag{3}$$

Value Function

Given a policy $\boldsymbol{\pi}$. Let $q_t^{\boldsymbol{\pi}}$ be the inventory process under policy $\boldsymbol{\pi}$, and the initial condition at time t be $S_t = S$, $q_t^{\boldsymbol{\pi}} = q$. Then the value function under policy $\boldsymbol{\pi}$ is

$$
\begin{aligned}
&V^{\boldsymbol{\pi}}(t,S,q) \\
&= \mathbb{E}\Big[\int_t^T \int_{\boldsymbol{\epsilon}_u} \Big[\sum_{k=1}^N z_k\big[\epsilon_u^b(k)dN_u^+(k) + \epsilon_u^a(k)dN_u^-(k)\big] + d(q_uS_u) - \delta h(q_u)dq_t\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u \\
&-\gamma\int_t^T\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u du \,\Big|\, S_t = S, q_t^{\boldsymbol{\pi}} = q\Big] \\
&= \mathbb{E}\Big[\int_t^T \int_{\boldsymbol{\epsilon}_u}\sum_{k=1}^N \Big[z_k\big(S_u + \epsilon_u^b(k) - \delta h(q_u)\big)dN_u^+(k) - z_k\big(S_u - \epsilon_u^a(k) - \delta h(q_u)\big)dN_u^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u \\
&-\gamma\int_t^T\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u du \,\Big|\, S_t = S, q_t^{\boldsymbol{\pi}} = q\Big]
\end{aligned} \tag{4}
$$

Then the value function under the optimal policy is

$$
\begin{aligned}
&V(t,S,q) \\
&= \max_{\boldsymbol{\pi}}\mathbb{E}\Big[\int_t^T \int_{\boldsymbol{\epsilon}_u}\sum_{k=1}^N \Big[z_k\big(S_u + \epsilon_u^b(k) - \delta h(q_u)\big)dN_u^+(k) - z_k\big(S_u - \epsilon_u^a(k) - \delta h(q_u)\big)dN_u^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u \\
&-\gamma\int_t^T\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u du \,\Big|\, S_t = S, q_t^{\boldsymbol{\pi}} = q\Big] \\
&= \max_{\boldsymbol{\pi}}\mathbb{E}\Big[\int_t^{t+\Delta t} \int_{\boldsymbol{\epsilon}_u}\sum_{k=1}^N \Big[z_k\big(S_u + \epsilon_u^b(k) - \delta h(q_u)\big)dN_u^+(k) - z_k\big(S_u - \epsilon_u^a(k) - \delta h(q_u)\big)dN_u^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u \\
&-\gamma\int_t^{t+\Delta t}\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u du + V(t+\Delta t, S_t+\Delta S_t, q_t + \Delta q_t^{\boldsymbol{\pi}}) \,\Big|\, S_t = S, q_t^{\boldsymbol{\pi}} = q\Big] \\
&= \max_{\boldsymbol{\pi}}\Big\{\int_{\boldsymbol{\epsilon}_t}\sum_{k=1}^N \Big[z_k\big(S_t + \epsilon_t^b(k) - \delta h(q_t)\big)\lambda_t^+(k) - z_k\big(S_t - \epsilon_t^a(k) - \delta h(q_t)\big)dN_t^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S_t,q_t^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_t\Delta t \\
&-\gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S_t,q_t^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S_t,q_t^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_t\Delta t + \mathbb{E}\Big[V(t+\Delta t, S_t + \Delta S_t, q_t^{\boldsymbol{\pi}} + \Delta q_t^{\boldsymbol{\pi}}) \,\Big|\, S_t = S, q_t^{\boldsymbol{\pi}} = q\Big]\Big\}
\end{aligned} \tag{5}
$$

Dynamic Programming

To make notation simplier, denote $\mathcal{L}V(t,S_t,q_t)$ as

$$
\mathcal{L}V(t,S_t,q_t) = V(t,S_t,q_t) + \big(\partial_t V(t,S_t,q_t) + \frac{1}{2}\sigma^2\partial_{SS}V(t,S_t,q_t)\big)\Delta t + \sigma\partial_S V(t,S_t,q_t)dW_t \tag{6}
$$

Since $dS_t = \sigma S_t dW_t$, and $dq_t = \sum_k z_k(dN_t^+(k) - dN_t^-(k))$, by the Ito formula, we have the following,

$$
\begin{aligned}
&V(t+\Delta t, S_t+\Delta S_t, q_t + \Delta q_t) \\
&= V(t+\Delta t, S_t+\Delta S_t, q_t)\prod_k(1-dN_t^+(k))(1-dN_t^-(k)) \\
&+ \sum_k\Big[V(t+\Delta t, S_t+\Delta S_t, q_t + z_k)dN_t^+(k) + V(t+\Delta t, S_t+\Delta S_t, q_t - z_k)dN_t^-(k)\Big] \\
&= \mathcal{L}V(t,S_t,q_t)\prod_k(1-dN_t^+(k))(1-dN_t^-(k)) + \sum_k \mathcal{L}V(t,S_t,q_t+z_k)dN_t^+(k) + \mathcal{L}V(t,S_t,q_t-z_k)dN_t^-(k)
\end{aligned} \tag{7}
$$

Notice that the above Ito formula is based on the assumption that the inventory process is $dq_t = \sum_k z_k(dN_t^+(k) - dN_t^-(k))$. Since the intensities of Poisson processes are determined by the quoted bid-ask spreads. So, the inventory process in the above Ito formula assumes the bid-ask spreads are already determined. Thus, when computing conditional expectation, $\mathbb{E}[V(t+\Delta t, S_t+\Delta S_t, q_t^{\boldsymbol{\pi}}+\Delta q_t^{\boldsymbol{\pi}})|S_t = S, q_t^{\boldsymbol{\pi}} = q]$, one should average over all possibilities. Then the conditional expectation is

$$\begin{aligned}
&\mathbb{E}\Big[V(t+\Delta t, S_t+\Delta S_t, q_t^{\boldsymbol{\pi}}+\Delta q_t^{\boldsymbol{\pi}}) \;\Big|\; S_t = S, q_t^{\boldsymbol{\pi}} = q\Big] \\
&= V(t,S,q) + \int_{\boldsymbol{\epsilon}_t} \boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)\Big[-\sum_k \big(\lambda_t^+(k)+\lambda_t^-(k)\big)V(t,S,q) + \partial_t V(t,S,q) + \frac{1}{2}\sigma^2\partial_{SS}V(t,S,q) \\
&\quad + \sum_k \big[\lambda_t^+(k)V(t,S,q+z_k) + \lambda_t^-(k)V(t,S,q-z_k)\big]\Big] d\boldsymbol{\epsilon}_t \Delta t
\end{aligned} \tag{8}$$

Then one can get the HJB equation,

$$\begin{aligned}
&\max_{\boldsymbol{\pi}}\Bigg\{ \int_{\boldsymbol{\epsilon}_t}\sum_k \Big[\lambda_t^+(k)V(t,S,q+z_k) + \lambda_t^-(k)V(t,S,q-z_k) - \big(\lambda_t^+(k)+\lambda_t^-(k)\big)V(t,S,q)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)d\boldsymbol{\epsilon}_t \\
&+ \int_{\boldsymbol{\epsilon}_t}\sum_{k=1}^N \Big[z_k\lambda_t^+(k)\big(S+\epsilon_t^b(k)-\delta h(q)\big) - z_k\lambda_t^-(k)\big(S-\epsilon_t^a(k)-\delta h(q)\big)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)d\boldsymbol{\epsilon}_t \\
&- \gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)d\boldsymbol{\epsilon}_t\Bigg\} + \partial_t V(t,S,q) + \frac{1}{2}\sigma^2\partial_{SS}V(t,S,q) \\
&= 0
\end{aligned} \tag{9}$$

Optimal Stochastic Policy

In order to find the maximizer for the quantity inside the max bracket of the HJB equation, we apply calculus of variation

$$\begin{aligned}
0 &= \int_{\boldsymbol{\epsilon}_t}\sum_k \Big[\lambda_t^+(k)V(t,S,q+z_k) + \lambda_t^-(k)V(t,S,q-z_k) - \big(\lambda_t^+(k)+\lambda_t^-(k)\big)V(t,S,q)\Big]\delta\boldsymbol{\pi} d\boldsymbol{\epsilon}_t \\
&\quad + \int_{\boldsymbol{\epsilon}_t}\sum_{k=1}^N \Big[z_k\lambda_t^+(k)\big(S+\epsilon_t^b(k)-\delta h(q)\big) - z_k\lambda_t^-(k)\big(S-\epsilon_t^a(k)-\delta h(q)\big)\Big]\delta\boldsymbol{\pi} d\boldsymbol{\epsilon}_t \\
&\quad - \gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}\frac{\delta\boldsymbol{\pi}}{\boldsymbol{\pi}}d\boldsymbol{\epsilon}_t - \gamma\int_{\boldsymbol{\epsilon}_t}\delta\boldsymbol{\pi}\log\boldsymbol{\pi} d\boldsymbol{\epsilon}_t
\end{aligned} \tag{10}$$

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

$$\int_{\boldsymbol{\epsilon}_t}\delta\boldsymbol{\pi} d\boldsymbol{\epsilon}_t = 0 \tag{11}$$

Then equation (10) becomes

$$\begin{aligned}
0 &= \int_{\boldsymbol{\epsilon}_t}\delta\boldsymbol{\pi}\Bigg(\sum_k \Big[\lambda_t^+(k)V(t,S,q+z_k) + \lambda_t^-(k)V(t,S,q-z_k) - \big(\lambda_t^+(k)+\lambda_t^-(k)\big)V(t,S_t,q_t) \\
&\quad + z_k\lambda_t^+(k)\big(S+\epsilon_t^b(k)-\delta h(q)\big) - z_k\lambda_t^-(k)\big(S-\epsilon_t^a(k)-\delta h(q)\big)\Big] - \gamma(\delta\boldsymbol{\pi})\log\boldsymbol{\pi} d\boldsymbol{\epsilon}_t\Bigg)d\boldsymbol{\epsilon}_t
\end{aligned} \tag{12}$$

Then the quantity inside the bracket above is a constant

$$C = \sum_k \Big[\lambda_t^+(k)V(t,S,q+z_k) + \lambda_t^-(k)V(t,S,q-z_k) - \big(\lambda_t^+(k)+\lambda_t^-(k)\big)V(t,S,q)$$
$$+ z_k\lambda_t^+(k)\big(S+\epsilon_t^b(k)-\delta h(q)\big) - z_k\lambda_t^-(k)\big(S-\epsilon_t^a(k)-\delta h(q)\big)\Big] - \gamma \log \boldsymbol{\pi} \tag{13}$$

We assume the relation between the intensity and spreads is

$$\lambda_t^{\pm}(k) = A_k - B_k \epsilon_t^{a,b}(k) \tag{14}$$

To simplify the notations, let

$$\mathcal{H}_k^+(t,S,q,\boldsymbol{\pi}) = V^{\boldsymbol{\pi}}(t,S,q+z_k) - V^{\boldsymbol{\pi}}(t,S,q) + z_k(S+\delta h(q)) \tag{15}$$
$$\mathcal{H}_k^-(t,S,q,\boldsymbol{\pi}) = V^{\boldsymbol{\pi}}(t,S,q-z_k) - V^{\boldsymbol{\pi}}(t,S,q) - z_k(S-\delta h(q)) \tag{16}$$

So, under optimal policy $\boldsymbol{\pi}^*$, we have

$$\mathcal{H}_k^+(t,S,q) = V(t,S,q+z_k) - V(t,S,q) + z_k(S+\delta h(q)) \tag{17}$$
$$\mathcal{H}_k^-(t,S,q) = V(t,S,q-z_k) - V(t,S,q) - z_k(S-\delta h(q)) \tag{18}$$

Then the optimal stochastic policy is

$$\begin{aligned}\boldsymbol{\pi}^*(\boldsymbol{\epsilon}_t|t,S,q) &\propto \exp\Big\{\frac{1}{\gamma}\sum_k \big(A_k - B_k\epsilon_t^{a,b}(k)\big)\big(z_k\epsilon_t^{a,b}(k) + \mathcal{H}_k^{\pm}(t,S,q)\big)\Big\} \\ &\propto \prod_k \exp\Big\{-\frac{z_kB_k}{\gamma}\Big[\epsilon_t^{a,b}(k) - \frac{A_k}{2B_k} + \frac{\mathcal{H}_k^{\pm}(t,S,q)}{2z_k}\Big]^2\Big\} \\ &\propto \prod_k \mathcal{N}\Big(\epsilon_t^{a,b} \mid \frac{A_k}{2B_k} - \frac{\mathcal{H}_k^{\pm}(t,S,q)}{2z_k}, \frac{\gamma}{2z_kB_k}\Big)\end{aligned} \tag{19}$$

Therefore, we know that the optimal policy will be a multi-dimensional Gaussian distribution. In order to simplify the notation, let

$$\boldsymbol{\mu}(t,S,q,\boldsymbol{\pi}) = \Big(\frac{A_1}{2B_1} - \frac{\mathcal{H}_1^{\pm}(t,S,q,\boldsymbol{\pi})}{2z_1}, ..., \frac{A_N}{2B_N} - \frac{\mathcal{H}_N^{\pm}(t,S,q,\boldsymbol{\pi})}{2z_N}\Big)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{\gamma}{2z_1B_1} & & & & \\ & \frac{\gamma}{2z_1B_1} & & & \\ & & \ddots & & \\ & & & \frac{\gamma}{2z_kB_k} & \\ & & & & \frac{\gamma}{2z_kB_k} \end{bmatrix}$$

Under the above notations, the optimal policy is

$$\boldsymbol{\pi}^* \sim \mathcal{N}(\cdot \mid \boldsymbol{\mu}(t,S,q,\boldsymbol{\pi}^*), \boldsymbol{\Sigma}) \tag{20}$$

Policy Improvement Theorem

Theorem 1.1 (policy improvement theorem). *Given any $\boldsymbol{\pi}$, let the new policy $\boldsymbol{\pi}_{new}$ to be*

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

then following inequality holds

$$V^{\boldsymbol{\pi}}(t,S,q) \le V^{\boldsymbol{\pi}_{new}}(t,S,q) \tag{22}$$

Proof. Let $q_t^{\boldsymbol{\pi}_{new}}$ be the inventory process under policy π_{new}. Let the initial condition at time t be $q_t^{\boldsymbol{\pi}_{new}} = q$, and $S_t = S$. Then by the Ito formula, and averaging over all possibilities, we have the following

$$
\begin{aligned}
& V^{\boldsymbol{\pi}}(t, S, q) \\
&= \mathbb{E}\Big[V^{\boldsymbol{\pi}}(s, S_s, q_s^{\boldsymbol{\pi}_{new}}) + \int_t^s \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, S_u, \boldsymbol{q}_u^{\boldsymbol{\pi}_{new}}) V^{\boldsymbol{\pi}}(u, S_u, q_u^{\boldsymbol{\pi}_{new}}) \sum_k [\lambda_u^+(k) + \lambda_u^-(k)] d\boldsymbol{\epsilon}_u du \\
&- \int_t^s \int_{\boldsymbol{\epsilon}_u} \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u | u, S_u, q_u^{\boldsymbol{\pi}_{new}}) \sum_k \Big[V^{\boldsymbol{\pi}}(u, S_u, q_u^{\boldsymbol{\pi}_{new}} + z_k)\lambda_u^+(k) + V^{\boldsymbol{\pi}}(u, S_u, q_u^{\boldsymbol{\pi}_{new}} - z_k)\lambda_u^-(k)\Big] d\boldsymbol{\epsilon}_u du \\
&- \int_s^t \Big(\partial_t V^{\boldsymbol{\pi}}(u, S_u, q_u^{\boldsymbol{\pi}_{new}}) + \frac{1}{2}\sigma^2 \partial_{SS} V^{\boldsymbol{\pi}}(u, S_u, q_u^{\boldsymbol{\pi}_{new}})\Big) du \;\Big|\; S_t = S, q_t^{\boldsymbol{\pi}_{new}} = q\Big]
\end{aligned}
\tag{23}
$$

Since at time t, under policy $\boldsymbol{\pi}$, the following equality holds,

$$
\begin{aligned}
& \int_{\boldsymbol{\epsilon}_t} \sum_k \Big[\lambda_t^+(k) V^{\boldsymbol{\pi}}(t, S, q + z_k) + \lambda_t^-(k) V^{\boldsymbol{\pi}}(t, S, q - z_k) - \big(\lambda_t^+(k) + \lambda_t^-(k)\big) V^{\boldsymbol{\pi}}(t, S, q)\Big] \boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t \\
&+ \int_{\boldsymbol{\epsilon}_t} \sum_{k=1}^N \Big[z_k \lambda_t^+(k)\big(S + \epsilon_t^b(k) - \delta h(q)\big) - z_k \lambda_t^-(k)\big(S - \epsilon_t^a(k) - \delta h(q)\big)\Big] \boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t \\
&- \gamma \int_{\boldsymbol{\epsilon}_t} \boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, S, q) \log \boldsymbol{\pi}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t + \partial_t V^{\boldsymbol{\pi}}(t, S, q) + \frac{1}{2}\sigma^2 \partial_{SS} V^{\boldsymbol{\pi}}(t, S, q) \\
&= 0
\end{aligned}
\tag{24}
$$

For $\boldsymbol{\pi}_{new}$, based on its construction, and by the same calculus of variation arguments as in equations $(10) - (13)$, $\boldsymbol{\pi}_{new}$ is the maximizer for the following quantity,

$$
\begin{aligned}
& \max_{\widetilde{\boldsymbol{\pi}}} \Big\{ \int_{\boldsymbol{\epsilon}_t} \sum_{k=1}^N \Big[z_k \lambda_t^+(k)\big(S + \epsilon_t^b(k) - \delta h(q)\big) - z_k \lambda_t^-(k)\big(S - \epsilon_t^a(k) - \delta h(q)\big)\Big] \widetilde{\boldsymbol{\pi}}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t \\
&+ \int_{\boldsymbol{\epsilon}_t} \sum_k \Big[\lambda_t^+(k) V^{\boldsymbol{\pi}}(t, S, q + z_k) + \lambda_t^-(k) V^{\boldsymbol{\pi}}(t, S, q - z_k) - \big(\lambda_t^+(k) + \lambda_t^-(k)\big) V^{\boldsymbol{\pi}}(t, S, q)\Big] \widetilde{\boldsymbol{\pi}}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t \\
&- \gamma \int_{\boldsymbol{\epsilon}_t} \widetilde{\boldsymbol{\pi}}(\boldsymbol{\epsilon}_t | t, S, q) \log \widetilde{\boldsymbol{\pi}}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t \Big\}
\end{aligned}
\tag{25}
$$

Which results in the following inequality,

$$
\begin{aligned}
& \int_{\boldsymbol{\epsilon}_t} \sum_k \Big[\lambda_t^+(k) V^{\boldsymbol{\pi}}(t, S, q + z_k) + \lambda_t^-(k) V^{\boldsymbol{\pi}}(t, S, q - z_k) - \big(\lambda_t^+(k) + \lambda_t^-(k)\big) V^{\boldsymbol{\pi}}(t, S, q)\Big] \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t \\
&+ \int_{\boldsymbol{\epsilon}_t} \sum_{k=1}^N \Big[z_k \lambda_t^+(k)\big(S + \epsilon_t^b(k) - \delta h(q)\big) - z_k \lambda_t^-(k)\big(S - \epsilon_t^a(k) - \delta h(q)\big)\Big] \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t \\
&- \gamma \int_{\boldsymbol{\epsilon}_t} \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_t | t, S, q) \log \boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_t | t, S, q) d\boldsymbol{\epsilon}_t + \partial_t V^{\boldsymbol{\pi}}(t, S, q) + \frac{1}{2}\sigma^2 \partial_{SS} V^{\boldsymbol{\pi}}(t, S, q) \\
&\geq 0
\end{aligned}
\tag{26}
$$

Then equation (23) yields

$$
\begin{aligned}
&V^{\boldsymbol{\pi}}(t,S,q)\\
&\le \mathbb{E}\Big[\int_t^s \int_{\boldsymbol{\epsilon}_t} \sum_{k=1}^N \Big[z_k\lambda_u^+(k)\big(S_u+\epsilon_u^b(k)-\delta h(q_u^{\boldsymbol{\pi}_{new}})\big)-z_k\lambda_u^-(k)\big(S_u-\epsilon_u^a(k)-\delta h(q_u^{\boldsymbol{\pi}_{new}})\big)\Big]\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}_{new}})d\boldsymbol{\epsilon}_u du\\
&-\gamma\int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}_{new}})\log\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}_{new}})d\boldsymbol{\epsilon}_u du+V^{\boldsymbol{\pi}}(s,S_s,q_s^{\boldsymbol{\pi}_{new}})\ \Big|\ S_t=S,q_t^{\boldsymbol{\pi}_{new}}=q\Big] \quad (27)
\end{aligned}
$$

Set $s=T$, we have $V^{\boldsymbol{\pi}}(T,S_T,q_T^{\boldsymbol{\pi}_{new}})=V^{\boldsymbol{\pi}_{new}}(T,S_T,q_T^{\boldsymbol{\pi}_{new}})$ then we have

$$
\begin{aligned}
&V^{\boldsymbol{\pi}}(t,S,q)\\
&\le \mathbb{E}\Big[\int_t^T \int_{\boldsymbol{\epsilon}_t} \sum_{k=1}^N \Big[z_k\lambda_u^+(k)\big(S_u+\epsilon_u^b(k)-\delta h(q_u^{\boldsymbol{\pi}_{new}})\big)-z_k\lambda_u^-(k)\big(S_u-\epsilon_u^a(k)-\delta h(q_u^{\boldsymbol{\pi}_{new}})\big)\Big]\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}_{new}})d\boldsymbol{\epsilon}_u du\\
&-\gamma\int_t^T\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}_{new}})\log\boldsymbol{\pi}_{new}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}_{new}})d\boldsymbol{\epsilon}_u du+V^{\boldsymbol{\pi}}(T,S_T,q_T^{\boldsymbol{\pi}_{new}})\ \Big|\ S_t=S,q_t^{\boldsymbol{\pi}_{new}}=q\Big]\\
&\le V^{\boldsymbol{\pi}_{new}}(t,S,q) \quad (28)
\end{aligned}
$$

□

Martingale Loss

Given a policy $\boldsymbol{\pi}$, and $q_t^{\boldsymbol{\pi}}$ is inventory process under policy $\boldsymbol{\pi}$. Let the initial condition at time t to be $S_t=S$, $q_t^{\boldsymbol{\pi}}=q$, the value function under policy $\boldsymbol{\pi}$ is

$$
\begin{aligned}
&V^{\boldsymbol{\pi}}(t,S,q)\\
&=\mathbb{E}\Big[\int_t^s\int_{\boldsymbol{\epsilon}_u}\sum_{k=1}^N\Big[z_k\big(S_u+\epsilon_u^b(k)-\delta h(q_u^{\boldsymbol{\pi}})\big)dN_u^+(k)-z_k\big(S_u-\epsilon_u^a(k)-\delta h(q_u^{\boldsymbol{\pi}})\big)dN_u^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u\\
&-\gamma\int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u du+V(s,S_s,q_s^{\boldsymbol{\pi}})\ \Big|\ S_t=S,q_t^{\boldsymbol{\pi}}=q\Big] \quad (29)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{s-t}\int_t^s\int_{\boldsymbol{\epsilon}_u}\sum_{k=1}^N\Big[z_k\big(S_u+\epsilon_u^b(k)-\delta h(q_u^{\boldsymbol{\pi}})\big)dN_u^+(k)-z_k\big(S_u-\epsilon_u^a(k)-\delta h(q_u^{\boldsymbol{\pi}})\big)dN_u^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u\\
&-\frac{\gamma}{s-t}\int_t^s\int_{\boldsymbol{\epsilon}_u}\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_u|u,S_u,q_u^{\boldsymbol{\pi}})d\boldsymbol{\epsilon}_u du+\frac{V^{\boldsymbol{\pi}}(s,S_s,q_s^{\boldsymbol{\pi}})-V^{\boldsymbol{\pi}}(t,S_t,q_t^{\boldsymbol{\pi}})}{s-t}\ \Big|\ S_t=S,q_t^{\boldsymbol{\pi}}=q\Big]=0\\
&\quad (30)
\end{aligned}
$$

Let $s\to t$, and parametrize the value function $V_\theta^{\boldsymbol{\pi}}$, define the temporal difference error as,

$$
\begin{aligned}
\delta_t^\theta&=\lim_{s\to t}\mathbb{E}\Big[\frac{V_\theta^{\boldsymbol{\pi}}(s,S_s,q_s^{\boldsymbol{\pi}})-V_\theta^{\boldsymbol{\pi}}(t,S_t,q_t^{\boldsymbol{\pi}})}{s-t}\ \Big|\ S_t=S,q_t^{\boldsymbol{\pi}}=q\Big]-\gamma\int_{\boldsymbol{\epsilon}_t}\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)d\boldsymbol{\epsilon}_t\\
&+\int_{\boldsymbol{\epsilon}_t}\sum_{k=1}^N\Big[z_k\big(S+\epsilon_t^b(k)-\delta h(q)\big)dN_t^+(k)-z_k\big(S-\epsilon_t^a(k)-\delta h(q)\big)dN_t^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_t|t,S,q)d\boldsymbol{\epsilon}_t \quad (31)
\end{aligned}
$$

Using the Monte Carlo method to generate a set of sample paths $\mathcal{D} = \{(t_i, S_i^d, q_i^d)_{i=1}^T\}_{d=1}^D$, then define the martingale loss as

$$
\begin{aligned}
\mathbf{ML}(\theta) = \frac{1}{2}\sum_{\mathcal{D}}\sum_{i}\Bigg(&\frac{V_\theta^{\boldsymbol{\pi}}(t_{i+1}, S_{t_{i+1}}^d, q_{t_{i+1}}^d) - V(t_i, S_{t_i}^d, q_{t_i}^d)}{\Delta t} - \gamma\int_{\boldsymbol{\epsilon}_{t_i}}\boldsymbol{\pi}(\boldsymbol{\epsilon}_{t_i}|t_i, S_{t_i}^d, q_{t_i}^d)\log\boldsymbol{\pi}(\boldsymbol{\epsilon}_{t_i}|t_i, S_{t_i}^d, q_{t_i}^d)d\boldsymbol{\epsilon}_{t_i} \\
&\int_{\boldsymbol{\epsilon}_{t_i}}\sum_{k=1}^{N}\Big[z_k\big(S_{t_i}^d + \epsilon_{t_i}^b(k) - \delta h(q_{t_i}^d)\big)dN_{t_i}^+(k) - z_k\big(S_{t_i}^d - \epsilon_{t_i}^a(k) - \delta h(q_{t_i}^d)\big)dN_{t_i}^-(k)\Big]\boldsymbol{\pi}(\boldsymbol{\epsilon}_{t_i}|t_i, S_{t_i}^d, q_{t_i}^d)d\boldsymbol{\epsilon}_{t_i}\Bigg)^2\Delta t
\end{aligned} \tag{32}
$$

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 $\mathcal{D} = \{(t_i, S_{t_i}, q_{t_i})_{i=1}^T\}$ under policy $\boldsymbol{\pi}^\phi$
    - Compute $\mathbf{ML}(\theta)$
    - Updates $\theta \leftarrow \theta - \alpha\nabla_\theta\mathbf{ML}(\theta)$
  - **end for**
  - Update $\boldsymbol{\pi}^\phi \leftarrow \mathcal{N}\Big(\boldsymbol{\epsilon}\ \Big|\Big(\frac{A_k}{2B_k} - \frac{\mathcal{H}_\theta^{\pm}(t,S,q,\boldsymbol{\pi}^\phi)}{2z_k}\Big), \Sigma\Big)$
- **end for**