Consider a Market Maker, i.e. an agent who is obliged to quote a bid b_t and an ask a_t, $a_t, b_t \in \mathbb{N}$, $b_t < a_t$. As a first approach, assume that the quoted volumes are always unit and that the time interval between the decision points is constant, without loss of generality equal to one.

We assume that, at any time, the MM can consume a unit of cash, preferring today's consumption, hence maximizing

$$\sum_{\tau=0}^{\infty} \gamma^{-\tau} c_\tau$$

where $\gamma < 1$ is a discount factor very close to one. In addition to c_t, the actions at t include setting the quotes $b_t < a_t$ so that the amount of money and stocks including reservations is non-negative.

The maximalistic DP

Denoting m_t and n_t the amount of money, stocks, respectively, the action space is

$$\begin{aligned} A_t = \{(c, b, a) \in \{0,1\} \times \mathbb{N}_0 \times \mathbb{N}_0 \\ : m_t - b_t - c_t \geq 0, n_t \geq 0, n_t = 0 \Leftrightarrow a_t = \infty, m_t = 0 \Leftrightarrow b_t = 0\} \end{aligned}$$

where quoting $b = 0$ or $a = \infty$ means not quoting (yet the MM is obliged to quote, it may happen that he cannot quote b or a due to the lack of money, stocks, respectively).

In addition to m_t and n_t,the state space includes the state of the market which we denote by R_t, i.e.

$$S = \mathbb{N}_0 \times \mathbb{N}_0 \times \mathcal{R}$$

where $\mathcal{R}$ is the state space of the market, which may in extreme contain all the history of the market including limit orders. The transition function is

$$m_t = m_{t-1} - c_{t-1} - D_t b_{t-1} + C_t a_{t-1}$$

$$n_t = n_{t-1} + D_t - C_t$$

$$R_t = F(r_{t-1}, a_{t-1}, b_{t-1})$$

where $D_t \in \{0,1\}$,$C_t \in \{0,1\}$ are indicators of a buy transaction, sell transaction, respectively, by the MM, and F is a very complex random function.

The Bellman equation is then

$$V(m, n, r) = \max_{(c,b,a) \in A(m,n)} c + \gamma \mathbb{E}_{C,D,R} V(m - c - Db + Ca, n + D - C, R)$$

where

$$\begin{aligned} A(m, n) = \{(c, b, a) \in \{0,1\} \times \mathbb{N}_0 \times \mathbb{N}_0 : \\ b < a, m - b - c \geq 0, n \geq 0, m = 0 \Leftrightarrow b = 0, n = 0 \Leftrightarrow a = \infty\} \end{aligned}$$

Note that, in practice, the distribution of C, D, it clearly depends on the quotes a and b (quoting near or inside the spread clearly increases the probability of transaction). Specially, given no quoting a or b, it is $C \equiv 0$, $D \equiv 0$, respectively.

Clearly, the problem suffers from all the three curses of dimensionality.

A minimalistic DP

Restriction of A

Clearly, quoting far from the spread leads to a negligible probability of transaction and the cross-quoting leads to the same results regardless of the quoted price, so we can restrict the action space to

$$A(m,n,\beta,\alpha) = \{(c,b,a) \in \{0,1\}\times\{\alpha-\Delta,\dots,\alpha+\delta\}\cup\{0\}\times\{\beta-\delta,\dots\beta+\Delta\}\cup\{0\} : \\ b < a, m-b-c \geq 0, n \geq 0, m = 0 \Leftrightarrow b = 0, n = 0 \Leftrightarrow a = \infty\}.$$

Here,

β is the best buy limit order of the other participants (if there is any) or own previous value of b (if some was set) or the initial price minus one

α is the best sell limit order of the other participants (if there is any) or own previous value of a (if some was set) or the initial price plus one

and $\delta, \Delta \in \mathbb{N}$ are given constants.

Restriction of S

As it is clear from the definition, the MM's objective depends directly only on process $c_t, b_t D_t, a_t C_t$. However, C and D depend on the state of the market, which is very complex in reality. In our opinion, a minimal reasonable market state space $\mathcal{R}$ should contain the the bid β and ask α save own quotes, i.e. $\mathcal{R} = \{\alpha \in \mathbb{N}_\infty, \beta \in \mathbb{N}_0 : \alpha > \beta\}$

Formulation of the problem

The Bellman equation is

$$V(m,n,\beta,\alpha) = \max_{(c,b,a)\in A(m,n,\beta,\alpha)} c + \gamma \mathbb{E}_{D,C,B,A} V(m-c-Db+Ca, n+D-C, B, A)$$

Here, B and A are new values of β, α, respectively.

Approximation of V

The future earnings of the MM do not depend much on the actual situation on the market, rather they depend on his resources held (if $m = n = 0$ for instance, then all future earnings are zero). Thus, we approximate $V(m,n,\beta,\alpha) \doteq W(M,N)$

As such an approximation does not depend on B and A, we need only the distribution of C and D for our computations. We assume that this distribution depends only on $\Delta a := a - \alpha$ (i.e. the new value of ask minus the present value of α) and $\Delta b := b - \beta$; thus, we have to estimate/learn only $(\Delta + \delta + 1)^2$ probabilities.

Note that $W(0,0) = 0$. As V is monotonic both in m and n (with more resources, there are more options, hence higher income), it is reasonable that its approximation W is monotonic too. We may use [AN APPROXIMATE DYNAMIC PROGRAMMING ALGORITHM FOR MONOTONE VALUE FUNCTIONS by Jiang and Powell][1] [2]to achieve the monotonicity and use value approximation:

1. Initialize W: $W(m,n) \leftarrow \frac{1-\gamma^{m+np}}{1-\gamma}$ for each m, n where p is the current price of the asset (corresponds to the situation when all is consumed as quickly as possible)

2. Initialize
$$N(C=1, D=1|\Delta b, \Delta a) = 0,$$
$$N(C=1, D=0|\Delta b, \Delta a) = K \times \begin{cases} 1/3 & \Delta a < 0 \\ 1/4 & \Delta a = 0 \\ 0 & \Delta a > 0 \end{cases}$$
$$N(C=0, D=1|\Delta b, \Delta a) = K \times \begin{cases} 1/3 & \Delta b > 0 \\ 1/4 & \Delta b = 0 \\ 0 & \Delta b < 0 \end{cases}$$
$$N(C=0, D=0) = K - N(C=1, D=0) - N(C=0, D=1)$$
where K is a number divisible by 3 (e.g. 300)

3. Initialize $m = m_0$, $n = n_0$, $a = p+1, b = p-1$ ´

4. Repeat until the end of the experiment
   (a) Get (C, D, β, α) from the simulator
   (b) Update $N(C = \bullet, D = \bullet|\Delta b_{last}, \Delta a_{last})$
   (c) Find $v \leftarrow \max_{(c,b,a)\in A(m,n,\beta.\alpha)}[c + \gamma\mathbb{E}(W(m - c - Db + Ca, n + D - C)|\Delta b, \Delta a)]$ where the expectation is computed by means of $p(\bullet|\Delta b, \Delta a)$ which is got from N.
   (d) $w \leftarrow \epsilon v + (1-\epsilon)W(m,n)$
   (e) $W \leftarrow \Pi_M((m,n), w, W)$ (4.1 of Jjang and Powell)

[1]https://castlelab.princeton.edu/html/Papers/Jiang-MonotoneADP_arxiv_V4_May252015.pdf

[2]Other options could be nearest neighbor. We can also approximate V by means of a piece-wise linear function during the ADP algorithm, namely to compute its values by interpolation of values (approximations) in existing points. Efficient implementation of this approximation is a non-trivial task, library https://www.sintef.no/globalassets/upload/ikt/9011/geometri/ttl/ttl_1.1.0_doc/index.html will hopefully help.