Pricing and Inventory Control For High Frequency Market Making

Dar En Tang

March 17, 2020

1 Pricing

The inventory at time t is denoted as $q(t)$. A discount δ is applied to the bid and ask prices for inventory control. It is defined by the polynomial:

$$\delta = -\text{sign}(q(t))\lfloor a|q(t)|^n \rfloor \tag{1}$$

Where a and n determines the "harshness" of the penalty. For example, these two parameters can be defined by specifying the inventory level at which one penalisation level is applied, q_1, and the penalisation level at the maximum allowable inventory level, $\delta_{\max}$ at $q_{\max}$. For instance, if q_1 is 15 and $q_{\max}$ and

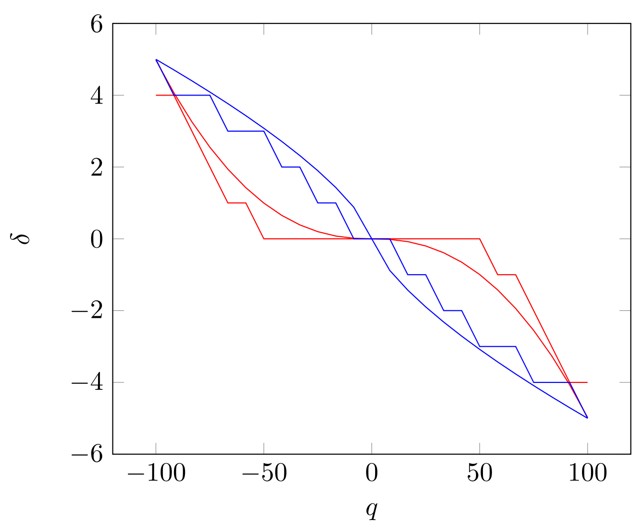

Figure 1: Penalty levels in response to inventory

$\delta_{\max}$ are 100 and 5 respectively, then $n = 2.322, a = 0.000113$. If we wish to penalise heavier, then q_1 can be set to a stricter value of 10, resulting in

$n = 0.699, a = 0.2$. In general, setting a smaller q_1 will result in a smaller fluctuation in inventory, and a higher $q_{\max}$ will result in a quicker "unstuck". Although please note that this is a trade off between stability and profitability, as any penalty applied to the price will result in suboptimal ordering.

Now, let the best ask and best bid in the etf order book be p_e^a and p_e^b respectively. In the future order book, same convention applies, i.e. p_f^a and p_f^b. The asking price is defined by:

$$p^a = \begin{cases} \max(p_e^a + \delta, p_f^a) & q < q_{\text{thres}} \\ \min(p_e^a, p_f^a) & \text{otherwise} \end{cases} \tag{2}$$

Where p_{thres} is the threshold inventory volume. In general, smaller q_{thres} will result in a more stable inventory. This value can set to be dependent on the current movement in the market:

$$q_{\text{thres}} = \min(q_{\text{thres}_{\max}}, f|\dot{p_e}|) \tag{3}$$

Where $q_{\text{thres}_{\max}}$ is the maximum threshold value, and f is some multiplier, typically $f = 0.4$. In a similar fashion, the bidding price can be defined as:

$$p^b = \begin{cases} \max(p_e^b, p_f^b + \delta) & q < -q_{\text{thres}} \\ \min(p_e^b, p_f^b) & \text{otherwise} \end{cases} \tag{4}$$

Now, equations 2 and 4 are discontinuous functions. To smooth the transition between the two, a mixing multiplier λ can be used, which can be defined from a logistic curve:

$$\lambda^a = \frac{1}{1 + \exp\left(-\kappa(q - q_{\text{thres}})\right)} \tag{5}$$

$$\lambda^b = \frac{1}{1 + \exp\left(-\kappa(-q - q_{\text{thres}})\right)} \tag{6}$$

Where κ is a mixing constant. A larger κ results in a stepper transition in the sigmoid curve.

In turn, equations 2 and 4 can be redefined as:

$$p^a = \lambda \max(p_e^a + \delta, p_f^a) + (1 - \lambda) \min(p_e^a, p_f^a) \tag{7}$$

$$p^b = \lambda \max(p_e^b, p_f^b + \delta) + (1 - \lambda) \min(p_e^b, p_f^b) \tag{8}$$