A few preliminaries

you should know market structure & terminology basics if you don’t already (e.g. order types, LOBs, spreads), this is a good intro:

https://www.machow.ski/posts/2021-07-18-introduction-to-limit-order-books/

The model here is from the paper “HFT in a limit order book” by Avellaneda & Stoikov. The derivation is non-trivial so I’ll focus on the motivation and results here

https://www.math.nyu.edu/~avellane/HighFrequencyTrading.pdf

Every MM and their mother knows this paper so there’s no unique alpha here, *but* it is the foundation of many models and you should know it too

Let's start with some motivation

Consider a pawn shop

When a customer arrives wanting to sell, you do a few things:
(1) Value their stuff best you can
(2) Offer to buy it below your valuation
(3) If they accept, store it until you can flip it for more than you paid for it

That’s roughly the MM business

Your store is your portfolio and the inventory you keep in it is financial assets, I like this comparison because it makes the term inventory risk clearer

If you’re running a business like this what are your major risks?

Principally, the risks are:
(1) bad valuations make you trade at bad prices
(2) you buy something but can't find anyone to sell it to
(3) price moves adversely b/w the time you buy and sell
(4) you bought too much and run out of cash until you sell some

A good market making model should manage for all of these

Let’s think about the flip side of what a profitable MM operation should look like again

Let’s say XYZ/USDT trades on exchanges A and B and you wanna make a market in it on A

You go long $100k XYZ on A and short $100k XZY on B at the start of the trading day, now you have a delta-neutral position on A to use for use for providing liquidity

Sanity check – why do we do this? Because trading happens both ways, and to sell something you must own it. Hedging like this lets you own it without assuming market risk

Now you only care about diff b/w the position on A and B (delta), which varies as your orders fill on A

f your buys/sells are filled evenly, your delta stays close to zero, but if flow is very one-sided, you could get into trouble if you don’t adapt

How do you adapt to one-sided flow?

A few ways:
(1) Resize your hedge with a market order on B
(2) Widen your quotes to adjust for uncertainty
(3) Skew your quotes based on your delta
(4) Skew your quotes based on alpha signals

Adjusting your hedge is the easiest way, but it’s relatively expensive

You gotta pay taker fees and cross the spread on B, hedging can eat into margins p badly

We should focus more on techniques 2-4 instead to make more money

For (2)

Suppose XYZ has a mid-price of $100, what’s the difference between quoting $99 @ $101 vs. $95 @ $105?

Namely the tradeoff is high volume + low margins vs. low volume + high margins

The latter also gives you more room for uncertainty in the future value of the asset

How do you quantify uncertainty in the future value of the asset?

You should’ve said volatility (i.e. variance of returns over a lagging period)

The takeaway is that to implement (2), our MM model should widen its pricing when volatility is high

Again, for XYZ, what’s the difference between quoting $95 @ $105 and $99 @ $105?

In the latter, you’re more likely to buy than sell because the bid is way closer to market value than the ask

This is attractive if your delta is too negative – the asymmetry adjusts the probability of getting fills in the direction you want

Important to emphasise this skew isn't based on future predictions, it’s only adjusting pricing based on the current state of *your* portfolio

It’s purely reactive, not predictive

If you have alpha signals, you could incorporate those into the MM model as well to position yourself more favourably for the future state of the market, but this is obvious

This can greatly help to reduce adverse selection but is hard

OOK if you’re still here and reading it is time to talk maths

I’m going to refer to this model as A&S going forward (after the authors initials)

A&S provides a 2-step procedure for optimally providing liquidity

(1) Compute an “indifference” price
(2) Compute the optimal spreads based on (1)

what the fuck is an indifference price

I’m glad you asked

it’s the price at which you’d be indifferent to buying/selling one more unit of the asset you’re trading

in simplest terms, it answers “how cheap would I need to this thing to be to reasonably buy more of it?”

the idea is that if you’ve bought a lot, you won’t want to buy more unless it’s a price far below market value

and conversely, you wouldn’t mind selling to others close to market value since you’ve got so much at the moment

A&S defines the indifference price as

$r(s,q,t) = s - q * gamma * sigma^2 * (T - t)$

where:
s = mid-price
q = inventory
gamma = risk-aversion param
sigma = volatility
T = end time
t = current time

If $q > 0$, $r < s$ indicating a desire to sell, and vice-versa with $q < 0$ indicating a desire to buy

If sigma >> 1, volatility is high and your pricing is further from mid-price

That captures the basics of what we want already

How do we derive the optimal spreads around this price?

What even makes a spread optimal?

A&S specifically defines the value function

$$u(s,x,q,t) = \max E \left[-\exp(-\gamma (X_t + q_T S_T)) \right]$$

where:
max is over delta(a), delta(b)
E = expected value
X_t = cash
q_t = inventory
S_t = mid-price

(X, q, s are all stochastic)

In other words, we're maxxing the EV of our net worth given the spreads we choose

The procedure to solve that for values of delta(a), delta(b) is non-trivial but I'll give the result

The optimal spreads are:

$$d(a) + d(b) = \gamma \sigma^2 (T-t) + \frac{2}{\gamma} \ln\left(1 + \frac{\gamma}{k}\right)$$

where:
d(a), d(b) = ask spread, bid spread
k = order-book density param
everything else the same

The value k comes from how we model the order flow we receive and there's some nuance in it

The simple way to think about it is high liquidity, high volume = higher value for k

I'll provide some more info but for now a quick recap

In A&S, the params you have control over are gamma (personal risk-aversion), k, and T (end-time)

With those you can compute the reservation price, the optimal spreads to quote around it, and begin market making with some inventory control

Finally let's talk about modeling order flow

Specifically, we'd consider the frequency, size distribution, and impact of market orders on price

Let's say frequency is constant, like total volume / avg. size of order

The distribution of size market order size is typically modelled as a power law, A&S gives it as $f(x) \sim x^{-1-\alpha}$ where alpha is a constant around 1.3–1.5 (derived empirically)

Market impact is harder, but A&S uses dP ~ ln(Q) where dP = change in price, Q = quantity

We can put all of these together to model order flow as a Poisson process parameterised by lambda

That parameter lambda is itself a function of the spread we choose

Based on the previous equations, A&S models lambda(delta) ~ A exp(-k delta)
where
A = frequency of MOs
k = alpha * K
K = proportionality constant

You’d typically do a regression against market data to get at a good value of K, but I’ve said enough for now

Future things to consider:
- how you define the fair value of an asset (is the mid-price really good or is it unreliable sometimes?)
- how do tick sizes and exchange rules affect your strategies?
- how do you incorporate alpha into your pricing?