ORF 474: High Frequency Trading
Notes 12b

Robert Almgren

April 29, 2020

Today we want to talk about computing optimal trajectories in models with market impact. We want to purchase a quantity X of shares or lots, starting at time $t = 0$ and completing by time $t = T$. Our goal is to determine a “trading trajectory” $x(t)$ for $0 \le t \le T$, where $x(t)$ is the quantity we have purchased by time t. Thus $x(0) = 0$ and $x(T) = X$ but $x(t)$ may have any shape in between.

We assume that underlying this trajectory is some “micro-algorithm” that handles the details of the order placement. Decisions between limit *vs* market orders, use of short-term signals such as microprice, and the like, go into the micro-algorithm. The market impact model we use will be a description of the results obtained by such a micro-algorithm. To fix ideas, imagine that the time horizon T is most of a day, and we divide time into bins of size say 5 minutes, 10 minutes, or so. We are trying to determine the allocation of trading across the entire day, that is, the quantity to executed in each bin. But this formulation does not determine how we execute within each bin.

We are making a pretty strong assumption, that the details of the micro-execution are independent of the overall scheduling. In practice, effects such as adverse selection (we get filled easily when the price is moving against us) would be important, but for these models we are ignoring that.

In any optimization problem there are three steps:

1. How do we model the environment?
2. What are we trying to achieve?
3. How do we calculate the optimal solution?

In trading problems, these questions and their answers have the three forms:

1. How does our trading affect the market? (market impact model)
2. What constitutes a good trade? (mean-variance criterion)
3. What mathematics do we use to compute solutions? (calculus of variations)

0.1 Market impact model

The time-dependent Kyle (1985) model suggests that the price change in each time interval $\Delta t_n = t_n - t_{n-1}$ is

$$\Delta p_n \;=\; \lambda\,(\,\Delta x_n \,+\, \Delta u_n\,) \;=\; \lambda\,\Delta x_n \,+\, \lambda\sigma_u\,\xi_n\,\sqrt{\Delta t_n}$$

where x_n is the cumulative position of the informed traders, which we intepret as our position, and u_n is the cumulative position of the uninformed traders. We assumed that the uninformed traders' position follows a random walk, so here ξ_n denotes an $\mathcal{N}(0,1)$ variable, indepedent on each step.

This model was derived as half of a strategic game, where the market makers who set the price assume that the informed traders are submitting orders to extract maximum value from a particular type of information. But we may take this as a model for general impact, for any trade list x_n that we choose to submit.

In this model, the price at step n is

$$\begin{aligned} p_n \;&=\; p_0 \,+\, \lambda\,x_n \,+\, \lambda\sigma_u \sum_{j=1}^{n} \sqrt{\Delta t_j}\,\xi_j \\ &=\; p_0 \,+\, \lambda\,x_n \,+\, \sigma\,B(t_n), \end{aligned}$$

with $\sigma = \lambda\sigma_u$ and where $B(t)$ denotes a Brownian motion (zero drift and unit variance per unit time). Viewed from $t = 0$, p_n is a normal random variable with

$$\begin{aligned} \mathbb{E}(p_n) \;&=\; p_0 \,+\, \lambda\,x_n \\ \mathrm{Var}(p_n) \;&=\; \sigma^2\,t_n. \end{aligned}$$

This assumes that our choice of trades $\{x_j\}_{j=1}^{n}$ is not random, that is, it does not depend on the realization of the uniformed trades or the Brownian motion.

Our trades impose a drift on the price process: if we buy ($x_n > 0$) we push it up and if we sell ($x_n < 0$) we push it down. The reason for this impact is that the market maker believes we have information, and when he[1] sees an imbalanced order flow, he adjusts his prices to compensate.

This discrete-time process has an immediate continuous-time limit

$$dp(t) \;=\; \lambda(\,dx(t) \,+\, du(t)\,) \;=\; \lambda\,dx(t) \,+\, \sigma\,dB(t)$$

and

$$p(t) \;=\; p_0 \,+\, \lambda\,x(t) \,+\, \sigma\,B(t).$$

The Kyle model for market impact has two important properties:

- It is *permanent:* the impact of our trading remains even after we stop trading, and

[1] I use a package that randomly inserts "he" or "she" each time I process the file.

- It is *public:* there is only one market price $p(t)$ at which we and everyone else transact.

Below we will see that to get interesting solutions we need to expand upon these properties.

The second part of step 1 is to calculate the cost of a trajectory. Suppose we have discrete positions $x = (x_0, x_1, \dots, x_N)$ with $x_0 = 0$ and $x_N = X$. The total cost that we pay to acquire our X shares is

$$C[x] \;=\; \sum_{n=1}^{N} p_n \, \Delta x_n.$$

This is the number of shares we purchase at each step, times the price of each transaction. The notation $C[x]$ means that this cost depends on our entire trajectory x. Of course the realized cost will also depend on the random price motion $B(t)$, but we focus on the part that we can control.

It is simpler to go to the continuous-time limit, in which we can plug in the price model above and identify three separate terms:

$$\begin{aligned} C[x] \;&=\; \int_0^T p(t)\, dx(t) \\ &=\; \int_0^T \big(\, p_0 \;+\; \lambda\, x(t) \;+\; \sigma\, B(t)\,\big)\, dx(t) \\ &=\; \{1\} \;+\; \{2\} \;+\; \{3\}. \end{aligned}$$

We can evaluate each of these terms individually:

$$\{1\} \;=\; \int_0^T p_0 \, dx(t) \;=\; p_0 X.$$

This is what you would pay if you could execute the entire transaction immediately at the starting market price. As Perold (1988) says,

> … on paper you transact instantly, costlessly, and in unlimited quantities … simply look at the current bid and ask, and consider the deal done at the average of the two.

We will take the "cost of trading" to be $C - Xp_0$, the difference between what you actually pay and what you thought it would be at the beginning.

For the second term,

$$\{2\} \;=\; \lambda \int_0^T x(t)\, dx(t) \;=\; \lambda \int_0^T d\Big(\frac{1}{2} x(t)^2\Big) \;=\; \frac{1}{2}\lambda\, x(t)^2 \Big|_{t=0}^{T} \;=\; \frac{1}{2}\lambda X^2$$

since $x(T) = X$ and $x(0) = 0$. This depends only on the total quantity X and is independent of the choice of trajectory $x(t)$. The reason is that each share traded pays the market impact caused by all the shares before that. Because impact is permanent, it does not matter how much you space them out, that is, how rapidly or slowly you trade.

For the third term, it is convenient to introduce

$$y(t) \;=\; X - x(t),$$

the quantity remaining to execute at each time, with $y(0) = X$ and $y(T) = 0$. Then $dx(t) = -dy(t)$ and

$$\begin{aligned}
\{3\} &= -\sigma \int_0^T B(t)\, dy(t) \\
&= -\sigma \int_0^T \Big(d\big(B(t)\, y(t)\big) - y(t)\, dB(t) \Big) \\
&= -\sigma \Big[B(t)\, y(t) \Big]_{t=0}^T + \sigma \int_0^T y(t)\, dB(t) \\
&= \sigma \int_0^T y(t)\, dB(t).
\end{aligned}$$

The first term here is zero because $B(0) = 0$ and $y(T) = 0$. To understand this, suppose that at some time t during execution, you have $y(t) > 0$, meaning that you have shares you must buy but have not yet bought. If $dB(t) > 0$ at that time, meaning that the price ticks upward, then your total cost increases by exactly the product of how many shares you still have left to buy, times the price change $\sigma\, dB(t)$ on those remaining shares.

This third term is a random variable, since we do not know how $B(t)$ will evolve. But viewed from $t = 0$, it is normal, with

$$\begin{aligned}
\mathbb{E}\{3\} &= 0 \\
\mathrm{Var}\{3\} &= \sigma^2 \int_0^T y(t)^2\, dt
\end{aligned}$$

(because $B(t)$ has unit variance per unit time). The intuition here is that at each time, $y(t)^2$ is our exposure to price motion.

Putting this all together, we have

$$C \;=\; p_0 X + \frac{1}{2}\lambda X^2 + \sigma \int_0^T y(t)\, dB(t)$$

This is a random variable: once we choose a fixed execution trajectory $x(t)$ or equivalently $y(t)$, the actual result depends on the realization of the price $B(t)$. But for a given trajectory $y(t)$, it is normal, with

$$\begin{aligned}
\mathbb{E}(C - p_0 X) &= \frac{1}{2}\lambda X^2 \\
\mathrm{Var}(C - p_0 X) &= \sigma^2 \int_0^T y(t)^2\, dt.
\end{aligned}$$

This completes the first step of our program: for any trajectory $x(t)$ we can assign a cost $C - p_0 X$.

0.2 Optimality criterion

The second step of our program is to decide what constitutes a “good” execution trajectory, and here we run into some problems with the pure Kyle model.

Clearly, a good trajectory should have a low value of cost $C - p_0 X$, on average. But here the expectation of cost does not depend on the trajectory! So we have nothing to optimize.

To salvage something from this model, we may recall that in portfolio optimization we talk a lot about variance. That is, in addition to reducing the value of cost itself, or its expected value, we also want to reduce *risk*, here measured as variance. Applying that idea here, we formulate the problem

$$\min_{y} \int_0^T y(t)^2 \, dt \qquad \text{over } y(t) \text{ on } 0 \le t \le 1 \text{ with } y(0) = X \text{ and } y(T) = 0.$$

This has a clear interpretation: get the trade done as quickly as possible to reduce the risk of the unexecuted quantity. The mathematical solution is clear:

$$y(t) \;=\; \begin{cases} X, & t = 0 \\ 0, & 0 < t \le 1 \end{cases} \qquad \text{(minimum variance of cost).} \tag{1}$$

It drops instantly to zero, executing the entire transaction in the shortest possible time.

But this is a terrible and unrealistic solution. It is not practical to assume that we can execute a large transaction completely in a very short time. We know that such a transaction will incur large cost, which we have so far left out of our model. To quote Perold (1988) again:

> What determines the amount of your execution costs and opportunity costs? … The chief factor is how quickly you trade.
>
> If you trade quickly and aggressively, you will tend to pay a bigger price to transact… The faster you trade, the larger your execution costs will be. On the other hand, you will have more of your ideal portfolio in place, and your opportunity costs consequently will be lower.

We need to add a term to the model to capture this effect.

1 Impact-dependent trajectories

We again need to go through our three steps of optimization.

1.1 Temporary market impact

We need to add some additional term to our model to capture the extra cost of rapid trading. One way to do this, which gives useful and interesting results, was introduced by Almgren and Chriss (2000). We suppose that we see a *private* price $\tilde{p}$ which is different from the public market price by an amount that depends on our instantaneous rate of trading:

$$\tilde{p}(t) \;=\; p(t) \;+\; g\big(x'(t)\big)$$

where $x'(t) = v(t)$ is our instantaneous *rate* of trading. You should think of this not as multiple prices at the same time, but as some form of adverse selection. For example, in a 10-minute window, the price will fluctuate. There may be one average price $p(t)$ observed by a neutral observer, but our own average execution price $\tilde{p}(t)$ may be slightly worse, because we somehow manage to buy at the top or sell at the bottom of all the little fluctuations.

The function $g(v)$ should be increasing, so that faster buying pushes the price up more, and faster selling pushes the price down more. There are several natural choices for the exact form.

- If we think that we always buy at the ask and sell at the bid, then we would intepret $p(t)$ as the bid-ask midpoint, and set

  $$g(v) \;=\; \frac{S}{2}\,\mathrm{sgn}(v)$$

  where sgn is the sign function, and S is the bid-ask spread. The problems with this model are that (a) large trades cannot always be filled at the inside quotes, so this will be an underestimate of the cost for large trade rates, and (b) small trades may be able to be filled passively to capture the spread, rather than aggressively to pay the spread, so this may overestimate the cost for small trade rates.

- A better model takes a concave nonlinear function

  $$g(v) \;=\; \eta\;\mathrm{sgn}(v)\,|v|^{\alpha}$$

  where the exponent α has $0 \le \alpha \le 1$ ($\alpha = 0$ is the spread model above). Empirical results (Almgren et al., 2005) suggest that $\alpha \approx 0.5$–0.6. This model is plausible, but the nonlinearity introduces analytical complexities that are not relevant for our current goals of understanding the basic tradeoffs of execution (Almgren, 2003).

- For this exposition, we will take the linear temporary cost model

  $$g(v) \;=\; \eta\, v$$

  where η is some coefficient.

This model semi-realistically captures the risk *vs.* reward tradeoff in optimal execution trajectories. We should note that more recent empirical evidence suggests a more complex structure (Bouchaud et al., 2018), which is essentially not captured by this temporary/permanent cost framework.

With this cost impact model, the cost of our trade program $y(t)$ becomes

$$\begin{aligned} C[x] &= \int_0^T \tilde{p}(t)\,dx(t) \\ &= \int_0^T \big(\,p_0 \;+\; \lambda\,x(t) \;+\; \sigma\,B(t) \;+\; \eta\,x'(t)\,\big)\,dx(t) \\ &= \{1\} \;+\; \{2\} \;+\; \{3\} \;+\; \{4\}. \end{aligned}$$

The first three terms here are exactly as above. The new fourth term is

$$\{4\} \;=\; \eta\int_0^T x'(t)\,dx(t) \;=\; \eta\int_0^T y'(t)^2\,dt$$

since $dx(t) = x'(t)\,dt$ and $y'(t) = -x'(t)$.

Putting this together as above, we find the total trading cost

$$C \;=\; p_0X \;+\frac{1}{2}\lambda X^2 \;+\; \eta\int_0^T y'(t)^2\,dt \;+\; \sigma\int_0^T y(t)\,dB(t)$$

Again, this is a random variable that depends on the trajectory $x(t)$ or $y(t)$. We can compute its expectation and variance, assuming $x(t)$ is fixed in advance:

$$E[x] \;=\; \mathbb{E}(C - p_0X) \;=\; \frac{1}{2}\lambda X^2 \;+\; \eta\int_0^T y'(t)^2\,dt \tag{2}$$

$$V[x] \;=\; \mathrm{Var}(C - p_0X) \;=\; \sigma^2\int_0^T y(t)^2\,dt. \tag{3}$$

We denote these as $E[x]$ and $V[x]$ to denote their dependence on the entire trajectory $x(t)$; we shall also denote them as $E[y]$ and $V[y]$ since specifying $x(t)$ is equivalent to specifying $y(t)$.

This completes the first step of our program: we can compute the cost of any chosen trajetory $x(t), y(t)$. Although the cost is a random variable, we can completely specify its expectation and its variance.

1.2 Optimality

The second step of our program is to decide what constitutes a “good” trajectory. Our first idea is simply to minimize the expected value of cost:

$$\min_y E[y] \qquad \text{over } y(t) \text{ on } 0 \le t \le 1 \text{ with } y(0) = X \text{ and } y(T) = 0.$$

With the form above, this is

$$\min_y \int_0^T y'(t)^2\,dt.$$

We leave off the term $\frac{1}{2}\lambda X^2$ which is fixed independently of $y(t)$; in fact, the Kyle coefficient has no effect on the final solution.

The solution to this problem is the straight-line trajectory

$$y(t) \;=\; \frac{T-t}{T}X, \qquad x(t) \;=\; \frac{t}{T}X \qquad \text{(minimum expected cost).} \tag{4}$$

If this is not obvious, it will emerge from our full solution below.

The problem with this solution is that it does not incorporate any incentive to trade rapidly. It suggests that you should stretch out your execution to take the full amount of time available to you. But in reality, there are reasons to get the trade done on a shorter time scale. Again from Perold (1988):

> If you trade slowly and patiently, your execution costs will tend to be lower… Nevertheless, although your execution costs will be lower, your opportunity costs will be higher. For the more slowly you trade, the more you will be forgoing the fruits of your research, and the more you will become prone to adverse selection (which shows up mostly in opportunity costs). The longer you are out there, the more time others have to act strategically against you.

Homework 5 asks you to implement a more direct modeling of “the fruits of your research,” in the form of an expected drift term. But for now, we follow Almgren and Chriss (2000) in combining the risk-aversion that we introduced in the pure Kyle model, with the cost-aversion that we have just outlined.

Risk and reward We now propose that an optimal execution schedule should somehow minimize *both* the expected value of cost and *also* its variance. This is exactly the famous framework introduced by Markowitz (1952) for portfolio formation and for similar reasons: portfolios that simply give the highest expected return have strange properties, as do trajectories that simply give the lowest expected cost.

Thus, we propose that an optimal trajectory should either give the lowest cost for a given level of variance, or the lowest variance for a given level of cost. Mathematically, these would be either

$$\min E[y] \qquad \text{over } y(t) \text{ with } V[y] \le V_0$$

or

$$\min V[y] \qquad \text{over } y(t) \text{ with } E[y] \le E_0$$

These are constrained optimization problems and hence somewhat complicated. A standard mathematical approach is to introduce a “Lagrange multiplier,” that is, introduce an artificial parameter γ and solve the unconstrained problem

$$\min_{y(t)} U[y] \qquad \text{with} \qquad U[y] \;=\; E[y] \;+\; \gamma V[y].$$

If we can solve this problem for all values of γ, then we can adjust the value of γ to determine the solutions to either of the two constrained optimization problems.[2]

Alternatively, γ has a very natural economic interpretation as a “risk-aversion parameter.” It tells us how much additional variance we are willing to accept, in order to reduce our expected cost by a certain amount. You may think of it that way if you do not like Lagrange multipliers.

Plugging in the expressions (2,3) for $E[y]$ and $V[y]$ (again we neglect the constant $\frac{1}{2}\lambda X^2$), we have our final objective function

$$U[y] \;=\; \eta \int_0^T y'(t)^2\,dt \;+\; \gamma\sigma^2 \int_0^T y(t)^2\,dt.$$

This completes step 2 of our optimization program: we have decided how we will decide what constitutes a good solution. We only need to do some mathematics to find the function $y(t)$ that minimizes the functional $U[y]$ for give values of the parameters η, γ, and σ.

1.3 Calculus of Variations

The main idea is this: If some function $y(t)$ is a minimizer of $U[y]$, then small perturbations of y must not be able to cause reductions in U. It is exactly analogous to the condition $F'(x) = 0$ for a function of a single variable x, or $\nabla F = 0$ for a function of several variables, except that here the derivative of $U[y]$ is taken with respect to the infinitely many degrees of freedom represented by the function $y(t)$ for $0 < t < T$.

So suppose we modify $y(t)$ to

$$\hat{y}(t) \;=\; y(t) \;+\; r\, z(t)$$

where $z(t)$ is some perturbation function, and r is a real number. In order for $y(t)$ to be a candidate solution of the problem, we must have

$$z(0) \;=\; 0 \qquad \text{and} \qquad z(T) \;=\; 0,$$

[2] In the lecture slides I have interchanged λ and γ relative to these notes, sorry.

since $y(0) = X$ and $y(T) = 0$ are fixed. Then

$$
\begin{aligned}
U[\hat{y}] &= \eta \int_0^T [y'(t) + rz'(t)]^2 \, dt \; + \; \gamma\sigma^2 \int_0^T [y(t) + rz(t)]^2 \, dt \\
&= U[y] \; + \; 2r \left(\eta \int_0^T y'(t)\, z'(t)\, dt \; + \; \gamma\sigma^2 \int_0^T y(t)\, z(t)\, dt \right) \; + \; r^2[\cdots]
\end{aligned}
$$

where $r^2[\cdots]$ denotes terms quadratic in r (which are in fact not hard to calculate in this case). The first term may be rewritten

$$
\begin{aligned}
\int_0^T y'(t)\, z'(t)\, dt &= \int_0^T \Big[\big(y'(t) z(t)\big)' \; - \; y''(t)\, z(t) \Big] \, dt \\
&= y'(t)\, z(t) \Big|_{t=0}^{T} \; - \; \int_0^T y''(t)\, z(t)\, dt \; = \; - \int_0^T y''(t)\, z(t)\, dt
\end{aligned}
$$

since $z(0) = 0$ and $z(T) = 0$. Thus the overall change in U is

$$
U[y + rz] \; - \; U[y] \; = \; -2r \int_0^T \Big\{ \eta\, y''(t) \; - \; \gamma\sigma^2\, y(t) \Big\} \, z(t)\, dt \; + \; r^2[\cdots]
$$

Suppose the term in curly brackets were nonzero at any $0 < t < T$. Then we could choose a perturbation function $z(t)$ that was nonzero at the same value of t, zero elsewhere.[3] There would be small values of r for which $U[y + rz] < U[y]$, which would mean that $y(t)$ was not a minimizer. Therefore, a necessary condition for $y(t)$ to minimize the functional $U[y]$ is that

$$
y''(t) \; = \; \kappa^2\, y(t) \qquad \text{for each } 0 < t < 1, \text{ with } \kappa^2 = \frac{\gamma\sigma^2}{\eta}.
$$

This is a differential equation for $y(t)$, whose solutions are of the form

$$
y(t) \; = \; A\, e^{\kappa t} \; + \; B\, e^{-\kappa t}
$$

for coefficients A and B chosen to satisfy the boundary conditions $y(0) = X$ and $y(T) = 0$. Specifically, the unique solution is

$$
y(t) \; = \; \frac{\sinh \kappa(T - t)}{\sinh \kappa T}\, X, \qquad \text{with } \sinh x = \frac{1}{2}\Big(e^x - e^{-x}\Big).
$$

The coefficient κ has units of inverse time, and represents the "time scale of optimal liquidation." It is not too hard to see the behavior of the solution in two limiting cases:

[3] For mathematicians: there are assumptions about continuity that we are not making explicit.

- Suppose κ is large. This means one or more of the following: volatility σ is large, γ is large (we are very risk-sensitive), or temporary impact η is small. The time scale is short. The solutions are very close to $y(t) = Xe^{-\kappa t}$ (since $e^{-\kappa T}$ is almost zero). As $\kappa \to \infty$ these solutions approach the pure Kyle solutions (1) in Section **??**, with extremely rapid trading.

- Suppose κ is small. This means one or more of the following: volatility σ is small, γ is small (we don't care about risk), or temporary impact η is large. Our desired time scale is long, but it will be constrained by the given time horizon T. The solutions are close to the straight-line form (4), which minimizes expectation of cost without regard to variance.

It is perhaps non-intuitive that the optimal time scale does not depend on the portfolio size. This is because for a given time scale, the expected cost increases as the square of the portfolio size, since you pay a cost per share that is linear in trade size, times the size of the portfolio itself. And variance of the result also increases as the square of the portfolio size, so the balance between these terms is independent of size. If you change either the form of the market impact, or how you measure risk, then this balance can change, and large portfolios are optimally traded either more quickly or more slowly than small ones.

References

Almgren, R. and N. Chriss (2000). Optimal execution of portfolio transactions. *J. Risk 3*(2), 5–39.

Almgren, R., C. Thum, E. Hauptmann, and H. Li (2005). Equity market impact. *Risk 18*(7, July), 57–62.

Almgren, R. F. (2003). Optimal execution with nonlinear impact functions and trading-enhanced risk. *Appl. Math. Fin. 10*, 1–18.

Bouchaud, J.-P., J. Bonart, J. Donier, and M. Gould (2018). *Trades, Quotes, and Prices: Financial Markets Under the Microscope.* Cambridge University Press.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica 53*, 1315–1336.

Markowitz, H. (1952). Portfolio selection. *J. Finance 7*, 77–91.

Perold, A. F. (1988). The implementation shortfall: Paper versus reality. *J. Portfolio Management 14*(3), 4–9.