LECTURE NOTES 2

Optimal Execution[1]

Petter Kolm
NYU Courant

In addition to these lecture notes, the required readings are:

- "Optimal Execution of Portfolio Transactions" by Almgren and Chriss (2001).
- "Optimal Execution with Nonlinear Impact Functions and Trading-Enhanced Risk" by Almgren (2003).

1. The Almgren & Chriss Model

In this section, we setup and solve the classic Almgren & Chriss (AC) model (see, Almgren and Chriss (2001) and Almgren (2003)) for a single security. The model can be extended to a portfolio of securities; refer to the original papers for details.

The objective is to liquidate X units of a security before time T at minimum cost. We will make "minimum cost" more precise later on.

We will use the following notation throughout:

- For some positive integer N let $\tau := T/N$. Then we refer to $t_k := k\tau$, for $k = 0, \ldots, N$, as the *time steps*.
- We refer to $x_0, \ldots, x_N$ as the *holdings*, where x_k is the number of units held of the security at time t_k.
- We assume initial and final holdings are given by $x_0 = X$ and $x_N = 0$.

[1] Version: February 15, 2022. These notes are copyrighted and intended only for students registered for the course *"Algorithmic Trading & Quantitative Strategies"* at NYU Courant. Please do not copy or post the notes on the web or pass them on to others by any means, whether digitally or otherwise. If you find any typos, I would much appreciate you email them to me.

- $n_1, \ldots, n_N$ is the *trade list.* Here $n_k = x_{k-1} - x_k$ is the number of units that we sell between t_{k-1} and t_k.

With this notation, we have

$$x_k = X - \sum_{j=1}^{k} n_j = \sum_{j=k+1}^{N} n_j\,, \quad \text{for } k = 0, \ldots, N\,. \tag{1}$$

Remark. With the initial and final holdings defined above, we are dealing with a liquidation problem. For an accumulation problem we would set $x_0 = 0$ and $x_N = X$. "Rebalancing trajectories" from X to Y also fit into this framework by setting $x_0 = X - Y$ and $x_N = 0$.

By a *trading strategy* we refer to a rule for determining the holdings n_k based on the information available. We distinguish between two types of trading strategies:

- *Static strategies*: These depend on information available before trading start at time t_0, and
- *Dynamic strategies*: These depend on information available up to and including time t_{k-1}.

We assume that the initial security price (at time t_0) is S_0 such that the market value of the portfolio is XS_0. In addition, we assume there are two kinds of market impact, temporary and permanent.

In earlier lecture notes, we discussed how market impact comes about, and the difference between temporary and permanent impact:

- Temporary impact: Temporary imbalances in supply and demand caused by our trading, leading to temporary price movements away from equilibrium in the current period only
- Permanent impact: Changes in the "equilibrium" price due to our trading, which remain at least for the life of our liquidation

When we are not trading, we assume security price evolves as the discrete arithmetic random walk

$$S_k = S_{k-1} + \sigma \tau^{1/2} \xi_k \tag{2}$$

for $k = 1, ..., N$, where σ is the instantaneous volatility of the security, and ξ_k are independent random variables with zero mean and unit variance, respectively. Note that by not having a drift term, we have no information about the direction of future price movements.

We incorporate permanent impact as a linear function of trade size into (2), obtaining

$$\begin{aligned} S_k &= S_{k-1} + \sigma\tau^{1/2}\xi_k - \gamma n_k && (3) \\ &= S_{k-1} + \sigma\tau^{1/2}\xi_k - \tau\gamma v_k && (4) \end{aligned}$$

where $\gamma > 0$ is the the linear permanent impact coefficient and $v_k := n_k/\tau$ is the average rate of trading during $[t_{k-1}, t_k]$.

Next, we incorporate temporary impact in the current period. Let us assume the price per share received for the sale in $[t_{k-1}, t_k]$ is

$$\tilde{S}_k := S_{k-1} - \eta \cdot \frac{n_k}{\tau} = S_{k-1} - \eta \cdot v_k \tag{5}$$

where $\eta > 0$ is the the linear temporary impact coefficient. Defined in this way, the temporary impact is linear in the rate of trading and it is gone in the next period. Therefore, it does not appear in the next market price S_k

$$S_k = S_{k-1} + \sigma\tau^{1/2}\xi_k - \gamma n_k\,. \tag{6}$$

1.1. Continuous Time Formulation. The discrete time formulation above is useful to obtain intuition about the model setup. For mathematical convenience, we introduce the continuous time formulation that will be useful for solving the resulting model.

In the continuous-time limit, $\tau \to 0$, the rate of trading becomes

$$v(t) := -\dot{x}(t) = \lim_{\tau\to 0} \frac{x(t-\tau) - x(t)}{\tau} \tag{7}$$

and therefore, our price process with permanent impact is given by

$$dS = \sigma dW - \gamma v dt = \sigma dW + \gamma dx \tag{8}$$

where $W(t)$ is a Brownian motion. Integrating, we obtain

$$S(t) = S_0 + \sigma W(t) - \gamma\left(X - x(t)\right) \tag{9}$$

Observe that permanent market impact is cumulative and proportional to the number of shares traded (as it should be).

As a warm-up exercise, we calculate total dollar cost incurred from permanent impact, $\mathcal{C}_{\text{perm}}$, over the whole the time period $[0,T]$ as follows

$$\begin{aligned} \mathcal{C}_{\text{perm}} &= \int_0^T \gamma(X-x(t))v(t)dt && (10) \\ &= -\int_0^T \gamma(X-x(t))\dot{x}(t)dt && (11) \\ &= -\gamma X x(t)|_0^T + \frac{1}{2}\gamma x^2(t)\Big|_0^T && (12) \\ &= \frac{1}{2}\gamma X^2 \,, && (13) \end{aligned}$$

as $x(0) = X$ and $x(T) = 0$. This calculation shows that permanent impact is independent of the specific choice of $v(t)$. In particular, it does not effect the optimal strategy.

Including temporary impact into (9) , we obtain

$$S(t) = S_0 + \sigma W(t) - \gamma\left(X - x(t)\right) + \eta\dot{x}(t)\,, \tag{14}$$

or equivalently

$$dS = \sigma dW + \gamma\dot{x}(t)dt + \eta\ddot{x}(t)dt\,. \tag{15}$$

In Appendix 2.A we provide an informal explanation for why $\eta\dot{x}(t)$ represents temporary impact.

We calculate the total revenues $\mathcal{R}$ from the trading strategy as follows

$$
\begin{aligned}
\mathcal{R} &= \int_0^T v(t)S(t)dt \qquad (16)\\
&= \int_0^T -\dot{x}(t)S(t)dt\\
&= -x(t)S(t)|_0^T + \int_0^T x(t)\frac{dS(t)}{dt}dt\\
&= X(S_0 + \eta\dot{x}(0)) + \int_0^T x(t)dS(t)\\
&= X(S_0 + \eta\dot{x}(0)) + \sigma\int_0^T x(t)dW + \int_0^T x(t)\left(\gamma\dot{x}(t) + \eta\ddot{x}(t)\right)dt\\
&= X(S_0 + \eta\dot{x}(0)) + \sigma\int_0^T x(t)dW - \frac{\gamma}{2}X^2 + \eta x(t)\dot{x}(t)|_0^T\\
&\qquad - \eta\int_0^T \left(\dot{x}(t)\right)^2 dt
\end{aligned}
$$

Simplifying, the total revenues become

$$
\begin{aligned}
\mathcal{R} &= X(S_0 + \eta\dot{x}(0)) + \sigma\int_0^T x(t)dW - \frac{\gamma}{2}X^2 - \eta X\dot{x}(0)\\
&\qquad -\eta\int_0^T \left(\dot{x}(t)\right)^2 dt\\
&= XS_0 + \sigma\int_0^T x(t)dW - \frac{\gamma}{2}X^2 - \eta\int_0^T \left(\dot{x}(t)\right)^2 dt \qquad (17)
\end{aligned}
$$

Recall the implementation shortfall is the difference of the value of the portfolio at the time when the decision to trade was made (i.e. XS_0) less the (realized) total revenues from the actual trade. Therefore, the implementation shortfall, $\mathcal{C}$, becomes

$$
\begin{aligned}
\mathcal{C} &= XS_0 - R \qquad (18)\\
&= -\sigma\int_0^T x(t)dW + \frac{\gamma}{2}X^2 + \eta\int_0^T \left(\dot{x}(t)\right)^2 dt \qquad (19)\\
&= -\sigma\int_0^T x(t)dW + \eta\int_0^T \left(\dot{x}(t)\right)^2 dt + B \qquad (20)
\end{aligned}
$$

where $B := \frac{\gamma}{2}X^2$ is a constant.

$\mathcal{C}$ is a random variable with expectation and variance given by

$$\mathbb{E}[\mathcal{C}] = \eta \int_0^T (\dot{x}(t))^2 \, dt + B \,, \tag{21}$$

$$\begin{aligned}
\mathrm{var}[\mathcal{C}] &= \mathbb{E}\left[(\mathcal{C} - \mathbb{E}[\mathcal{C}])^2\right] \\
&= \mathbb{E}\left[\left(-\sigma \int_0^T x(t) dW\right)^2\right] = \mathbb{E}\left[\sigma^2 \int_0^T (x(t) dW)^2\right] \\
&= \sigma^2 \int_0^T (x(t))^2 \, dt \,,
\end{aligned} \tag{22}$$

where the last equality follows from Itô's lemma.

1.2. The Optimal Execution Problem. Almgren and Chriss (2001) propose solving the following problem

$$\min_{x(t)} \mathbb{E}[\mathcal{C}] + \lambda \mathrm{var}[\mathcal{C}] \tag{23}$$

$$\text{s.t. } x(0) = X \tag{24}$$

$$x(T) = 0 \tag{25}$$

where $\mathcal{C}$ is the implementation shortfall, and $\lambda \geq 0$ is the risk aversion to execution risk.

Just like in classical mean-variance optimization of Markowitz (1952), this problem is posed as a trade-off between the mean and variance. Here it is formulated in the mean and variance of implementation shortfall rather than the security or portfolio return. Substituting the functional forms for $\mathbb{E}[\mathcal{C}]$ and $\mathrm{var}[\mathcal{C}]$, the mean-variance problem (23)–(25) becomes

$$\min_{x(t)} \int_0^T (\eta \dot{x}^2 + \lambda \sigma^2 x^2) dt \tag{26}$$

$$\text{s.t. } x(0) = X \tag{27}$$

$$x(T) = 0 \,. \tag{28}$$

This is a variational problem of the form

$$\min_{x(t)} \int_0^T F(x, \dot{x}, t) dt \tag{29}$$

$$\text{s.t. } x(0) = X\,, \quad x(T) = 0\,, \tag{30}$$

where

$$F(x, \dot{x}, t) = \eta \dot{x}^2 + \lambda \sigma^2 x^2\,. \tag{31}$$

It is well-known from variational calculus that the necessary condition for this problem to have an extremum for a given function x is that the *Euler-Lagrange equation* is satisfied[2]

$$F_x - \frac{d}{dt} F_{\dot{x}} = 0 \tag{32}$$

with boundary conditions $x(0) = X$ and $x(T) = 0$.

Notice that in our situation F does not explicitly depend on t. That is, $F(x, \dot{x}, t) \equiv F(x, \dot{x})$ so $\frac{\partial F}{\partial t} = 0$. Therefore,

$$F_x - \frac{d}{dt} F_{\dot{x}} = F_x - \dot{x} F_{x\dot{x}} - \ddot{x} F_{\dot{x}\dot{x}}\,. \tag{33}$$

Multiplying (33) by $\dot{x}$, we obtain

$$\begin{aligned} 0 &= \dot{x} F_x - \dot{x}\dot{x} F_{x\dot{x}} - \dot{x}\ddot{x} F_{\dot{x}\dot{x}} && (34) \\ &= \frac{d}{dt}(F - \dot{x} F_{\dot{x}})\,. && (35) \end{aligned}$$

Hence

$$F - \dot{x} F_{\dot{x}} = K \tag{36}$$

where K is a constant.

Inserting (31) and $F_{\dot{x}} = 2\eta\dot{x}$ into (36), we obtain

$$\begin{aligned} F - \dot{x} F_{\dot{x}} &= \eta \dot{x}^2 + \lambda \sigma^2 x^2 - 2\eta \dot{x}^2 && (37) \\ &= \lambda \sigma^2 x^2 - \eta \dot{x}^2 && (38) \\ &= K\,. && (39) \end{aligned}$$

[2] Standard texts on calculus of variations include Gelfand and Fomin (1963) and Dacorogna (2014). A brief introduction to the topic is given in Figueroa-O'Farrill (no year).

We recognize (38)-(39) is the first order inhomogeneous ODE

$$\lambda\sigma^2 x^2 - \eta\dot{x}^2 = K\,. \tag{40}$$

It is straightforward to solve the homogenous ODE

$$\lambda\sigma^2 x^2 - \eta\dot{x}^2 = 0 \tag{41}$$

by "taking square roots" to obtain

$$\dot{x} = \kappa x\,, \qquad \kappa := \sqrt{\frac{\lambda\sigma^2}{\eta}}\,. \tag{42}$$

Therefore

$$x = Ae^{\kappa t} \tag{43}$$

where A is a constant.

The solution to the inhomogenous ODE (41) is therefore of the form

$$x = Ae^{\kappa t} + Be^{-\kappa t} \tag{44}$$

for some constants A and B. From the boundary conditions

$$\begin{aligned} x(0) &= A + B \equiv X & (45)\\ x(T) &= Ae^{\kappa T} + Be^{-\kappa T} \equiv 0\,, & (46) \end{aligned}$$

we obtain

$$\begin{aligned} A &= X\frac{-e^{-\kappa T}}{e^{\kappa T} - e^{-\kappa T}} & (47)\\ B &= X\frac{e^{\kappa T}}{e^{\kappa T} - e^{-\kappa T}}\,. & (48) \end{aligned}$$

Altogether, the solution to our variational problem (26)–(28) is

$$\begin{aligned} x(t) &= \frac{X}{e^{\kappa T} - e^{-\kappa T}}\left(-e^{-\kappa(T-t)} + e^{\kappa(T-t)}\right) & (49)\\ &= X\frac{\sinh\left(\kappa(T-t)\right)}{\sinh(\kappa T)} & (50) \end{aligned}$$

where $\kappa = \sqrt{\frac{\lambda\sigma^2}{\eta}}$ is referred to as the *urgency parameter*. Figures 1 and 2 depict the resulting optimal trading trajectories and efficient frontier.

Concluding, we highlight some of the qualitative characteristics of the solution and its dependence on the urgency parameter (κ):

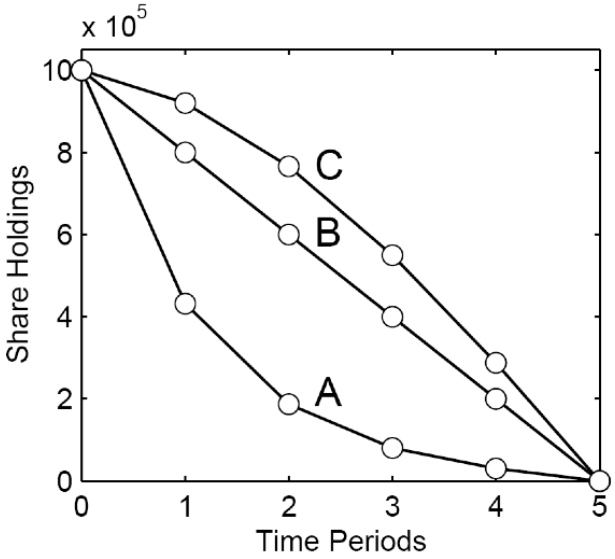

FIGURE 1. Optimal trading trajectories. (Source: Almgren and Chriss (2001).)

- When the risk aversion increases ($\lambda \uparrow$), then the rate of trading increases,
- When the volatility of the security increases ($\sigma \uparrow$), then the rate of trading increases, and
- When the temporary price impact coefficient increases ($\eta \uparrow$), then the rate of trading decreases.

Of course, these characteristic empirically match the behavior observed in real-world trading.

Remark. How would a trader use this model? First, they need to specify their risk aversion, which is based on their own personal beliefs (i.e. this parameter is heterogeneous across traders). The security's volatility and temporary price impact coefficient need to be estimated from data, but they should be the same (or at least similar) across traders (i.e. these two parameters are homogenous across traders).

1.3. Model Assumptions And Extensions. The AC model is a building block for more general execution models used in the financial

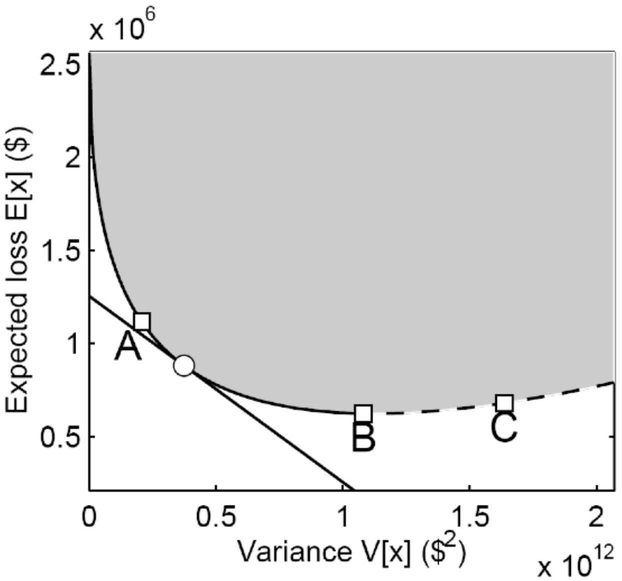

Figure 2. The efficient frontier. (Source: Almgren and Chriss (2001).)

industry today. Many broker dealers offer whole suites of models for different trading purposes clients may have.

Mathematically, the AC model has been generalized in a number of different directions. A way to generalize the security price dynamics (15) is to consider the following system

$$dB = rBdt - \dot{x}f(\dot{x})Sdt \qquad \text{(bank account)} \tag{51}$$

$$dS = (\mu + g(\dot{x}))Sdt + \sigma SdW \tag{52}$$

where $f(\cdot), g(\cdot)$ are the temporary and permanent market impact functions, respectively (see, for example, Forsyth, Kennedy, Tse, and Windcliff (2012) and Valle and Pacheco-Gonzalez (2009)). The trader's portfolio value P is the sum of holdings in the bank account and security, namely

$$P(t) = B(t) + x(t)S(t) \tag{53}$$

We observe that in this formulation, $-\dot{x}f(\dot{x})Sdt$ is the amount of cash required to buy $\dot{x}dt$ shares at a total price of $f(\dot{x})S$. For example, in the AC model above, we considered the special case $f(\dot{x}) = \eta\frac{\text{sign}(\dot{x})}{S}$.

In Almgren (2003), $f(\dot{x}) = 1 + \frac{h(\dot{x})S_0}{S}$ and $h(\dot{x}) = \operatorname{sign}(\dot{x})(\kappa_s + \kappa_t \left|\dot{x}\right|^{\beta})$ under arithmetic Brownian motion. Forsyth (2011) considers $f(\dot{x}) = (1 + \kappa_s \operatorname{sign}(\dot{x})) \exp\left(\kappa_t \operatorname{sign}(\dot{x}) \left|\dot{x}\right|^{\beta}\right)$ under geometric Brownian motion.

In more general models, permanent impact is typically still modeled as a linear function. This assumption is supported by the work of Huberman and Stanzl (2004) that suggests that there is a round-trip arbitrage possible when permanent impact is nonlinear. Empirically, it is known that temporary impact is nonlinear (see, for example, Almgren, Thum, Hauptmann, and Li (2005)). The most common "improvement" to the Almgren-Chriss model we derived above is to use nonlinear temporary impact.

What is the difference between arithmetic and geometric Brownian motion in the context of these models? First, at short time scales the two processes are hard to distinguish. However, over longer time periods there is a noticeable difference. The arithmetic process can of course become negative, which is undesirable. Mathematically, for specific forms of the permanent and temporary market impact functions, arithmetic Brownian motion will result in solutions that can be expressed in closed form. For geometric Brownian motion closed form solutions are seldom available and one has to resort to solving the full or a reduced form of the Hamilton-Jacobi-Bellman equation numerically.

2. Common Algorithmic Trading Strategies

This section follows Kolm and Maclin (2010) and Kolm and Maclin (2012). A small number of execution strategies have become de facto standards and are offered by most technology providers, banks, and institutional broker/dealers. However, even among these standards, the large number of input parameters makes it difficult to compare execution strategies directly.

Typically, a strategy is motivated by a *theme*, or *style* of trading. The objective is to minimize either absolute or risk-adjusted costs relative to a *benchmark*. For strategies with mathematically defined objectives, an optimization is performed to determine how to best use the strategy to maximize a trader's or portfolio manager's utility. A *trade schedule* – or *trajectory* – is planned for strategies with a target quantity of shares to execute. The *order placement engine* – sometimes called the *microtrader* – translates from a strategy's broad objectives to individual orders. User defined input parameters control the trade schedule and order placement strategy.

In this section we review some of the most common algorithmic trading strategies.

2.1. Arrival Price. The *arrival price* strategy, also called the *implementation shortfall* strategy, attempts to minimize risk-adjusted costs using the arrival price benchmark. Arrival price optimization is the most sophisticated and popular of the commonly used algorithmic trading strategies. Most implementations are based on some form of the risk-adjusted cost minimization introduced by Almgren and Chriss (2001) that we derived in the previous section.

The ideal user of arrival price strategies has the following characteristics:

- They are benchmarked to the arrival price;
- They are risk averse and know their risk aversion parameter;
- They have high positive or high negative alpha; and,
- They believe that market impact is minimized by maintaining a constant rate of trading over the maximum execution period while keeping trade size small.

The parameters of an arrival price strategy are start time, end time, number of shares to execute, a risk aversion parameter, and alpha. In the most general terms, an arrival price strategy evaluates a series of trade schedules to determine which one minimizes risk-adjusted costs relative to the arrival price benchmark. As discussed in the section on optimal execution, under certain assumptions, this problem has a closed form solution.

The parameters in an arrival price optimization are alpha, number of shares to execute, start time, end time, and a risk aversion parameter. For buyers (sellers) positive (negative) alpha encourages faster trading. For both buyers and sellers, risk encourages faster trading, while market impact costs encourage slower trading.

For traders with positive alpha, the feasible region of trade schedules lies between the immediate execution of total target quantity and a constant rate of trading throughout the execution period.

A more general form of arrival price optimization allows for both buyers and sellers to have either positive or negative alpha. For example, under the assumption of negative alpha, shares held long and scheduled for liquidation are – without considering one's own trading – expected to go up in price over the execution period. This would encourage a trader to delay execution or stretch out trading. Hence, the feasible region of solutions that account for both positive and negative alpha includes back-weighted as well as front-weighted trade schedules.

Other factors that necessitate back-weighted trade schedules in an arrival price optimization are expected changes in liquidity and expected crossing opportunities. For example, an expectation of a later in the execution period may provide enough cost savings to warrant taking on some price risk and the possibility of a compressed execution if the cross fails to materialize. Similarly, if market impact costs are expected to be lower later in the execution period, a rational trader may take on some risk to obtain this cost savings.

A variant of the basic arrival price strategy is *adaptive arrival price*. A favorable execution may result in a windfall in which an accumulation of a large number of shares takes place at a price significantly below the arrival price. This can happen by random chance alone. Almgren and Lorenz (2007) demonstrated that a risk-averse trader should use some of this windfall to reduce the risk of the remaining shares. They do this by trading faster and thus incurring a higher market impact. Hence, the strategy is adaptive in that it changes its behavior based on how well it is performing.

2.2. Volume-Weighted Average Price. The *volume weighted average price* (VWAP) execution strategy is probably one of the most well-known strategies next to arrival price. The appeal of benchmarking to VWAP is that the benchmark is easy to compute and intuitively accessible.

The typical parameters of a VWAP execution are the start time, the end time, and the number of shares to execute. Additionally, optimized forms of this strategy require a choice of risk aversion.

The most basic form of VWAP trading uses a model of the fractional daily volume pattern over the execution period. A trade schedule is calculated to match this volume pattern. For example, if the execution period is one day, and 20% of a day's volume is expected to be transacted in the first hour, a trader using this basic strategy would trade 20% of their target accumulation or liquidation in the first hour of the day. Since the daily volume pattern has a U shape – with more trading in the morning and afternoon and less in the middle of the day – the volume distribution of shares executed in a VWAP pattern would also have this U shape.

VWAP is an ideal strategy for a trader who meets all of the following criteria:

- Their trading has little or no alpha during the execution period;
- They are benchmarked against the volume weighted average price;
- They believe that market impact is minimized when their own rate of trading represents the smallest possible fraction of all trading activity; and,
- They have a set number of shares to buy or sell.

Deviation from these criteria may make VWAP strategies less attractive. For example, market participants who trade over the course of a day and have strong positive alpha may prefer a front-weighted trajectory, such as those that are produced by an arrival price strategy.

The period of a VWAP execution is most typically a day or a large fraction of a day. Basic VWAP models predict the daily volume pattern using a simple historical average of fractional volume. Several weeks to several months of data are commonly used. However, this forecast is noisy. On any given day, the actual volume pattern deviates substantially from its historical average, complicating the strategy's objective of minimizing its risk-adjusted cost relative to the VWAP benchmark. Some models of fractional volume attempt to increase the accuracy of volume pattern prediction by making dynamic adjustments to the prediction based on observed trading results throughout the day.

Several variations of the basic VWAP strategy are common. The ideal VWAP user (as defined above) can lower their expected costs by increasing their exposure to risk relative to the VWAP benchmark. For example, assuming an alpha of zero, placing limit orders throughout the execution period and catching up to a target quantity with a market order at the end of the execution period will lower expected cost while increasing risk. This is the highest risk strategy. Continuously placing small market orders in the fractional volume pattern is the lowest risk strategy, but has a higher expected cost. For a particular choice of risk aversion, somewhere between the highest and lowest risk strategies, is a compromise optimal strategy that perfectly balances risk and costs.

For example, a risk averse VWAP strategy might place one market order of 100 shares every twenty seconds while a less risk averse strategy might place a limit order of 200 shares, and 40 seconds later, place a market order for the difference between the desired fill of 200 and the actual fill (which may have been smaller). The choice of the average time between market orders in a VWAP execution implies a particular risk aversion.

For market participants with a positive alpha, a frequently used rule-of-thumb optimization is compressing trading into a shorter execution period. For example, a market participant may try to capture more profits by doing all of their VWAP trading in the first half of the day instead of taking the entire day to execute.

In another variant of VWAP referred to as *guaranteed VWAP*, a broker commits capital to guaranty their client the VWAP price in return for a pre-determined fee. The broker takes on a risk that the difference between their execution and VWAP will be greater than the fee they collect. If institutional trading volume and individual stock returns were uncorrelated, the risk of guaranteed VWAP trading could be diversified away across many clients and many stocks. In practice, managing a guaranteed VWAP book requires some complex risk calculations that include modeling the correlations of institutional trading volume.

2.3. Time Weighted Average Price. The *time weighted average price* (TWAP) execution strategy attempts to minimize market impact costs by maintaining an approximately constant rate of trading over the execution period. With only a few parameters – start time, end time, and target quantity – TWAP has the advantage of being the simplest execution strategy to implement. As with VWAP, optimized forms of TWAP may require a choice of risk aversion. Typically, the VWAP or arrival price benchmarks are used to gauge the quality of a TWAP execution. TWAP is hardly ever used as its own benchmark.

The most basic form of TWAP breaks a parent order into small child orders and executes these child orders at a constant rate. For example, a parent order of 300 shares with an execution period of 10 minutes could be divided into three child orders of 100 shares each. The child orders would be executed at the 3:20, 6:40, and 10:00 minute marks. Between market orders, the strategy may place limit orders in an attempt to improve execution quality.

An ideal TWAP user has almost the same characteristics as an ideal VWAP user, except that they believe that the lowest trading rate – not the lowest participation rate – incurs the lowest market impact costs.

TWAP users can benefit from the same type of optimization as VWAP users by placing market orders less frequently, and using resting limit orders to attempt improving execution quality.

2.4. Participation. The *participation* strategy attempts to maintain a constant fractional trading rate. That is, its own trading rate as a fraction of the market's total trading rate should be constant throughout the execution period. If the fractional trading rate is maintained exactly, participation strategies cannot guarantee a target fill quantity.

The parameters of a participation strategy are the start time, end time, fraction of market volume the strategy should represent, and max number of shares to execute. If the max number of shares is specified, the strategy may complete execution before the end time. Along with VWAP and TWAP, participation is a popular form of non-optimized strategies, though some improvements are possible with optimization.

VWAP and arrival price benchmarks are often used to gauge the quality of a participation strategy execution. The VWAP benchmark is particularly appropriate because the volume pattern of a perfectly executed participation strategy is the market's volume pattern during the period of execution. An ideal user of participation strategies has all of the same characteristics as an ideal user of VWAP strategies, except that they are willing to forego certain execution to maintain the lowest possible fractional participation rate.

Participation strategies do not use a trade schedule. The strategy's objective is to participate in volume as it arises. Without a trade schedule, a participation strategy can't guarantee a target fill quantity. The most basic form of participation strategies waits for trading volume to show up on the tape, and follows this volume with market orders. For example, if the target fractional participation rate is 10%, and an execution of 10,000 shares is shown to have been transacted by other market participants, a participation strategy would execute 1,000 shares in response.

Unlike a VWAP trading strategy, which for a given execution may experience large deviations from an execution period's actual volume pattern, participation strategies can closely track the actual – as opposed to the predicted – volume pattern. However, close tracking has a price. In the above example, placing a market order of 1,000 shares has a larger expected market impact than slowly following the market's trading volume with smaller orders. An optimized form of the participation strategy amortizes the trading shortfall over some period of time. Specifically, if an execution of 10,000 shares is shown to have been transacted by other market participants, instead of placing 1,000 shares all at once, a 10% participation strategy might place 100 share orders over some period of time to amortize the shortfall of 1,000 shares. The result is a lower expected shortfall, but a higher dispersion of shortfalls.

2.5. Market-On-Close. The *market-on-close* strategy is popular with market participants who either want to minimize risk-adjusted costs relative to the closing price of the day, or want to manipulate or game the close to create the perception of a good execution. The ideal market-on-close user is benchmarked to the close of the day and has low or negative alpha. The parameters of a market-on-close execution are the start time, the end time, and the number of shares to execute. Optimized forms of this strategy require a risk-aversion parameter.

When market-on-close is used as an optimized strategy, it is similar in its formulation to an arrival price strategy. However, with market-on-close, a back-weighted trade schedule incurs less risk than a front-weighted one. With arrival price, an infinitely risk averse trader would execute everything in the opening seconds of the execution period. With market-on-close, an infinitely risk averse trader would execute everything at the closing seconds of the day. For typical levels of risk aversion, some trading would take place throughout the execution period. As with arrival price optimization, positive alpha increases urgency to trade and negative alpha encourages delayed execution.

In the past, market-on-close strategies were used to manipulate – or game – the close, but this has become less popular as the use of VWAP and arrival price benchmarks have increased. Gaming the close is achieved by executing rapidly near the close of the day. The trade print becomes the closing price or very close to it, and hence shows little or no shortfall from the closing price benchmark. The true cost of the execution is hidden until the next day when temporary impact dissipates and prices return to a new equilibrium.

2.6. Crossing. Though *crossing networks* have been around for some time, their use in algorithmic trading strategies is a relatively recent development. The idea behind crossing networks is that large limit orders – the kind of orders that may be placed by large institutional traders – are not adequately protected in a public exchange. Simply displaying large limit orders in the open book of an electronic exchange may leak too much information about institutional traders' intentions. This information is used by prospective counter-parties to trade more passively in the expectation that time constraints will force traders to replace some or all of large limit orders with market orders. In other words, information leakage encourages gaming of large limit orders. Crossing networks are designed to limit information leakage by making their limit books opaque to both their clients and the general public.

A common form of cross is the *mid-quote cross*, in which two counter-parties obtain a mid-quote fill price. The mid-quote is obtained from a reference exchange, such as the NYSE or other public exchange. Regulations require that the trade is then printed to a public exchange to alert other market participants that it has taken place. The cross has no market impact but both counter-parties pay a fee to the crossing network. These fees are typically higher than the fees for other types of algorithmic trading because the market impact savings are significant while the fee is contingent on a successful cross.

More recently, crossing networks have offered their clients the ability to place limit orders in the crossing networks' dark books. Placing a limit order in a crossing network allows a cross to occur only at a certain price. This makes crossing networks much more like traditional exchanges, with the important difference that their books are opaque to market participants.

To protect their clients from price manipulation, crossing networks implement anti-gaming logic. As previously explained, opaqueness is itself a form of anti-gaming, but there are other strategies. For example, some crossing networks require orders above a minimum size, or orders that will remain in the network longer than a minimum time commitment. Other networks will cross only orders of similar size. This prevents traders from pinging – sending small orders to the network to determine which side of the network's book has an order imbalance.

Another approach to anti-gaming prevents crosses from taking place during periods of unusual market activity. The assumption is that some of this unusual activity is caused by traders trying to manipulate the spread in the open markets to get a better fill in a crossing network.

Some networks also attempt to limit participation by active traders, monitoring their clients' activities to see if their behavior is more consistent with normal trading than with gaming.

There are several different kinds of crossing networks. A *continuous crossing network* constantly sweeps through its book in an attempt to match buy orders with sell orders. A *discrete crossing network* specifies points in time when a cross will take place, say every half hour. This allows market participants to queue up in the crossing network just prior to a cross instead of committing resting orders to the network for extended periods of time. Some crossing networks allow scraping – a one time sweep to see if a single order can find a counter-party in the crossing network's book – while others allow only resting orders.

In *automated crossing networks*, resting orders are matched according to a set of rules, without direct interaction between the counter-parties. In *negotiated crossing networks*, the counterparties first exchange indications of interest, then negotiate price and size via tools provided by the system.

Some traditional exchanges may allow the use of *invisible orders*, resting orders that sit in their order books but are not visible to market participants. These orders are also referred to as *dark liquidity*. The difference between these orders and those placed in a crossing network is that traditional exchanges offer no special anti-gaming protection.

Private dark pools are collections of orders that are not directly available to the public. For example, a bank or pension manager might have enough order flow to maintain an internal order book that, under special circumstances is exposed to external scraping by a crossing network or *crossing aggregator*.

A crossing aggregator charges a fee for managing a single large order across multiple crossing networks. Order placement and anti-gaming rules differ across networks, making this task fairly complex. A crossing aggregator may also use information about historical and real-time fills to direct orders. For example, failure to fill a small resting buy order in a crossing network may betray information of a much larger imbalance in the network's book. This makes the network a more attractive destination for future sell orders. In general, the management of information across crossing networks should give crossing aggregators higher fill rates than exposure to any individual network.

Crossing lends itself to several optimization strategies. Longer exposure to a crossing network increases the chances of an impact-free fill, but also increases the risk of a large and compressed execution if an order fails to obtain a fill. Finding an optimal exposure time is one type of crossing optimization. A more sophisticated version of this approach is solving for a *trade-out*, a schedule for trading shares out of the crossing network into the open markets. As time passes and a cross is not obtained, the strategy mitigates the risk of a large, compressed execution by slowly trading parts of the order into the open markets.

2.7. Other Algorithms. Two other algorithms are typically included in the mix of standard algorithmic trading offerings. The first is *liquidity seeking* where the objective is to soak up available liquidity. As the order book is depleted, trading slows. As the order book is replenished, trading speeds up.

The second algorithm is *financed trading*. The idea behind this strategy is to use a sale to finance the purchase of a buy with the objective of obtaining some form of hedge. This problem has all of the components of a full optimization. For example, if, after a sell, a buy is executed too quickly, it will obtain a less favorable fill price. On the other hand, executing a buy leg too slowly increases the tracking error between the two components of the hedge and increases the dispersion of costs required to complete the hedge.

2.A. Appendix. Why Does the Term $\eta\dot{x}(t)$ Represent Temporary Impact?

Recall that

$$dS = \sigma dW + \gamma\dot{x}(t)dt + \eta\ddot{x}(t)dt\,. \tag{54}$$

Equivalently,

$$S(t) = S_0 + \sigma W(t) - \gamma\left(X - x(t)\right) + \eta\dot{x}(t)\,. \tag{55}$$

In discrete form this looks like

$$S_k = S_0 + \gamma(x_k - X) + \sigma\tau^{1/2}\sum_{i=1}^{k}\xi_i + \eta\Delta x_k \tag{56}$$

where $\eta\Delta x_k$ is the temporary impact. Note, that there is no temporary impact from previous time periods, i.e. terms like $\Delta x_{k-1}, \Delta x_{k-2}, \ldots$, etc.

Maybe to see this "even clearer," we can derive the expression for S_k above from the recursive relationship

$$S_k = S_{k-1} + \sigma\tau^{1/2}\xi_k + \gamma(x_k - x_{k-1}) + \eta(\Delta x_k - \Delta x_{k-1})\,. \tag{57}$$

Successive substitution shows

$$\begin{aligned}
S_k &= S_{k-1} + \sigma\tau^{1/2}\xi_k + \gamma(x_k - x_{k-1}) + \eta(\Delta x_k - \Delta x_{k-1}) && (58)\\
&= S_{k-2} + \sigma\tau^{1/2}\xi_{k-1} + \gamma(x_{k-1} - x_{k-2}) + \eta(\Delta x_{k-1} - \Delta x_{k-2}) && (59)\\
&\qquad +\sigma\tau^{1/2}\xi_k + \gamma(x_k - x_{k-1}) + \eta(\Delta x_k - \Delta x_{k-1}) && (60)\\
&= S_{k-2} + \sigma\tau^{1/2}(\xi_{k-1} + \xi_k) + \gamma(x_k - x_{k-2}) \\
&\qquad +\eta(\Delta x_k - \Delta x_{k-2}) && (61)\\
&\ \ \vdots \\
&= S_0 + \sigma\tau^{1/2}\sum_{i=1}^{k}\xi_i + \gamma(x_k - x_0) + \eta(\Delta x_k - \Delta x_0) && (62)\\
&\equiv S_0 + \sigma\tau^{1/2}\sum_{i=1}^{k}\xi_i + \gamma(x_k - X) + \eta\Delta x_k && (63)
\end{aligned}$$

as $x_0 = X$ and $\Delta x_0 = 0$. We see that all the temporary impact terms but the one in the current period cancel out.

References

Almgren, Robert and Neil Chriss (2001). "Optimal Execution Of Portfolio Rransactions". In: *Journal of Risk* 3, pp. 5–40.

Almgren, Robert and Julian Lorenz (2007). "Adaptive Arrival Price". In: *Algorithmic trading III: Precision, Control, Execution* 2007.1, pp. 59–66.

Almgren, Robert, Chee Thum, Emmanuel Hauptmann, and Hong Li (2005). "Equity Market Impact". In: *Risk* 18.7, pp. 57–62.

Almgren, Robert F. (2003). "Optimal Execution With Nonlinear Impact Functions And Trading-Enhanced Risk". In: *Applied Mathematical Finance* 10.1, pp. 1–18.

Dacorogna, Bernard (2014). *Introduction to the Calculus of Variations.* World Scientific Publishing Company.

Figueroa-O'Farrill, José Miguel (no year). "Brief Notes On the Calculus of Variations". URL: https://www.maths.ed.ac.uk/~jmf/Teaching/Lectures/CoV.pdf.

Forsyth, Peter A. (2011). "A Hamilton–Jacobi–Bellman Approach To Optimal Trade Execution". In: *Applied Numerical Mathematics* 61.2, pp. 241–265.

Forsyth, Peter A., J. Shannon Kennedy, S.T. Tse, and Heath Windcliff (2012). "Optimal Trade Execution: A Mean Quadratic Variation Approach". In: *Journal of Economic Dynamics and Control* 36.12, pp. 1971–1991.

Gelfand, I.M. and S.V. Fomin (1963). "Calculus Of Variations. Revised English Edition Translated And Edited By Richard A. Silverman". In: *Prentice Hall, Englewood Cli s, NJ* 7, pp. 10–11.

Huberman, Gur and Werner Stanzl (2004). "Price Manipulation And Quasi-Arbitrage". In: *Econometrica* 72.4, pp. 1247–1275.

Kolm, Petter N. and Lee Maclin (2010). “Algorithmic Trading”. In: *Encyclopedia of Quantitative Finance*.

— (2012). “Algorithmic Trading, Optimal Execution, and Dyna Mic Port Folios”. In: *The Oxford Handbook of Quantitative Asset Management*.

Markowitz, Harry (1952). “Portfolio Selection”. In: *The Journal of Finance* 7.1, pp. 77–91.

Valle, Gerardo Hernandez-del and Carlos Pacheco-Gonzalez (2009). “Optimal Execution Of Portfolio Transactions With Geometric Price Process”. In: *arXiv preprint arXiv:0908.1211*.