7. Transaction Costs

Prof. Gordon Ritter
October 23, 2017

7.1. **Limit Order Books.** This is not a course on market microstructure, but we will need to understand the basics of trading in limit-order books to continue. For subsection 7.1, a good background reference is Gould et al. (2013).

The majority of the world's transactions, by volume, are enacted in marketplaces which take the form of a continuous limit-order book (LOB). This means that almost anyone can participate in the market, by quoting prices at which they would buy or sell the asset. Quotes come flowing in, and an automated computer system listens to those quotes and uses them to build a representation of the supply and demand curves, more or less as you learned them in undergraduate micro-economics. The main difference is that the price levels are discrete, and the "curves" are continuously changing as new quotes come in and old ones are canceled.

Definition 7.1. A *limit order* or *quote* $x = (p_x, q_x, t_x)$ submitted at time t_x with price p_x and size $q_x > 0$ (respectively, $q_x < 0$) is a commitment to buy (respectively, sell) up to $|q_x|$ units of the traded asset at a price no greater than (respectively, no less than) p_x.

In this section, we will sometimes abbreviate "limit order" to simply "order" and assume all orders are limit orders.

Definition 7.2. The *lot size* ℓ of an LOB is the smallest amount of the asset that can be traded. All quotes must arrive with a size

$$q_x \in \{\pm k\ell \mid k = 1, 2, \dots\}$$

The *tick size* of an LOB is the smallest permissible price interval between different orders within it. All orders must arrive with a price that is specified to the accuracy of one tick.

The lot size and tick size are called *resolution parameters*.

When a buy (respectively, sell) order x is submitted, an LOB's matching engine checks whether it is possible to match x to some other previously submitted sell (respectively, buy) order. If so, the matching occurs immediately. If not, x becomes *active*, and it remains active until either it becomes matched to another incoming sell (respectively, buy) order or it is canceled. Cancellation usually occurs because the owner of an order no longer wishes to offer a trade at the stated price, but rules governing a market can also lead to the cancellation of active orders.

Let $\mathcal{L}(t)$ denote the set of all active orders in a market at time t. The active orders in an LOB $\mathcal{L}(t)$ can be partitioned into the set of active buy orders $\mathcal{B}(t)$, for which $q_x > 0$, and the set of active sell orders $\mathcal{A}(t)$, for which $q_x < 0$. An LOB can then be considered as a set of queues, each of which consists of active buy or sell orders at a specified price. Usually, but not always, in equity markets the queues are time-priority, or in other words they are FIFO queues. It follows that the first to join one of these queues (the trader who can react more quickly) has an advantage. Note that you cannot gain priority by posting a quote 10^{-9} cents above or below someone else's quote, due to the tick size.

Definition 7.3. The *bid price* at time t is the highest stated price among active buy orders at time t.

$$b(t) := \max_{x \in \mathcal{B}(t)} p_x.$$

The *ask price* at time t is the lowest stated price among active sell orders at time t.

$$a(t) := \min_{x \in \mathcal{A}(t)} p_x.$$

The bid-ask spread at time t is $s(t) := a(t) - b(t)$. The mid price at time t is $m(t) := [a(t) + b(t)]/2$.

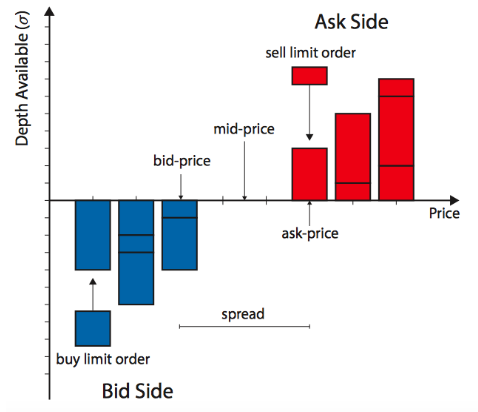

FIGURE 7.1. Schematic of an abstract limit-order book as depicted by Gould et al. (2013).

In the above schematic, note again the similarity to supply and demand curves from micro-economics.

Orders that result in immediate matching upon submission are known as *market orders*. This terminology is used only to emphasize whether an incoming order triggers an immediate matching or not. There is no fundamental difference between a limit order and a market order. Trading via limit orders which are not expected to immediately match is called *trading passively*. The opposite – submitting a sequence of market orders – is referred to as being *aggressive*. See also `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3057570` for further discussion of how to optimize the timing of aggressive or passive orders in the presence of a longer-term utility function.

In an LOB, traders are able to choose between submitting limit orders and submitting market orders. Limit orders stand a chance of matching at better prices than do market orders, but they also run the risk of never being matched. Conversely, market orders never match at prices better than $b(t)$ and $a(t)$, but they do not face the inherent uncertainty associated with limit orders. An LOB's bid-ask spread $s(t)$ is a measure of how highly the market values the immediacy and certainty associated with market orders versus the waiting and uncertainty associated with limit orders.

Foucault, Kadan, and Kandel (2005) argued that the popularity of LOBs is due in part to their ability to allow some traders to demand immediacy, while simultaneously allowing others to supply it to those who later require it. Most traders use a combination of both limit orders and market orders;

they select their actions for each situation based on their individual needs at that time.

7.2. **Parent Orders and Slippage.**

Definition 7.4. A *parent order* $\mathfrak{o} = (q, \tau, \tau')$ will be an instruction to buy or sell a fixed quantity q of a certain asset over the time window $[\tau, \tau']$. Per convention, an order to *sell* has $q < 0$ while an order to buy has $q > 0$. A *fill* is a statement that part of the buying or selling in a particular order has been completed at a certain time and a certain price. A parent order is also called a *metaorder*.

This terminology is especially suited to cash equities, bonds, and equity options. For equities, the "quantity" is an *integer* in units of shares, while for options, the quantity is typically in units of contracts.

Thus the parent order $\mathfrak{o}$, if it were completed, would lead to a sequence of fills $\{f_i : i = 1 \dots n_f\}$ where each fill

$$f_i = (\mathfrak{o}, t_i, n_i, p_i)$$

is made up of the parent order, the time $t_i \in [\tau, \tau']$, the number n_i of shares filled, and the price p_i at which they were filled. Let $\mathrm{pa}(f)$ denote the parent order that generated the fill.

In other words a fill $f_i = (\mathfrak{o}, t_i, n_i, p_i)$ is a statement that n_i shares have been exchanged for cash in the amount of $n_i p_i$ dollars (or other numeraire currency) at time t_i, as part of the parent order $\mathfrak{o}$. Also, we assume all of the fills associated to a fixed parent order have $\mathrm{sgn}(n_i) = \mathrm{sgn}(q)$, as is logical. Thus necessarily

$$\sum_i |n_i| = |\sum_i n_i| \le |q|.$$

The total amount filled is $\sum_{i=1}^{n_f} n_i$ which could be smaller in magnitude than $|q|$; in other words, not every parent order is completely filled. For example, a parent order could be an unrealistically large number of shares in a very illiquid stock. The execution algo could either decide not to fully fill based on certain agreed-upon limits, or it could simply fail to locate the shares. If the order involves taking a short position in a stock that is hard to borrow, failing to locate is a common occurrence.

Definition 7.5. An *execution algorithm* or, in common usage, an *algo*, is a means of creating a sequence of fills for any given parent order, ie. a mapping from $\mathfrak{o} \to \{f_i\}$. Equivalently it is a means of choosing n_i and t_i in the sequence $f_i = (\mathfrak{o}, t_i, n_i, p_i)$.

For any list of pairs $L = \{(n_i, p_i) : i = 1, \dots, N_L\}$ where $p_i > 0$ are prices and $n_i \in \mathbb{N}$, we define

$$\mathrm{vwap}(L) = \frac{\sum_i n_i p_i}{\sum_i n_i}\,.$$

where vwap stands for "volume-weighted average price." In one common example, L is the list of all trade prices, with the volume transacted at each price, for a particular (stock, day). For analysis of intraday patterns, one could take L to be a subset of this data over a finer-grained time period, such as a minute.

The vwap of an order is defined to be the vwap of the sequence of fills used to fill the order:

$$\mathrm{vwap}(\mathfrak{o}) = \mathrm{vwap}\{f : \mathrm{pa}(f) = \mathfrak{o}\}$$

To the portfolio manager, the end result of executing the order is essentially the same as if the entire quantity q had been filled in one shot, at price $\mathrm{vwap}(\mathfrak{o})$. The goal for a *buy* parent order is for $\mathrm{vwap}(\mathfrak{o})$ to be as small as possible, and for a sell parent order, the goal is for $\mathrm{vwap}(\mathfrak{o})$ to be as large as possible.

FIGURE 7.2. Schematic of a parent sell order being executed by an algo. Each red triangle denotes a child order being filled. The green (resp blue) line is the national best offer (resp bid). This order has high slippage to arrival mid, because the price moved so quickly.

Definition 7.6. A *benchmark pricing method* is a way of assigning a theoretical price $p_0(\mathfrak{o}) \in \mathbb{R}$ to any order, ie a mapping $p_0 : \mathfrak{O} \to \mathbb{R}$ where $\mathfrak{O}$ denotes the space of all possible orders. This price need not represent the prices of any actual trades.

One popular benchmark is arrival (mid) price, or simply "arrival price." This is the last midpoint price available before the order begins being executed. This seems to be the benchmark used in Almgren et al. (2005), for example.

We now come to the most important definition of this section:

Definition 7.7. Given a benchmark pricing method p_0, the *slippage* of an order relative to this benchmark pricing method is defined as

$$\mathrm{slip}(\mathfrak{o}) = \left(\sum_{i=1}^{n_f} n_i \right) [\mathrm{vwap}(\mathfrak{o}) - p_0(\mathfrak{o})] \tag{7.1}$$

Note that $\mathrm{slip}(\mathfrak{o})$ has units of whatever currency the prices are denominated in, and the sign is such that *positive* slippage denotes a *worse* result for the trader than transacting at the benchmark price. If $\mathfrak{O}$ is an entire set of orders, then

$$\mathrm{slip}(\mathfrak{O}) = \sum_{\mathfrak{o} \in \mathfrak{O}} \mathrm{slip}(\mathfrak{o})$$

As stated, the benchmark pricing method is a mathematical construct and need not correspond to tradable prices. For example, if we take the zero mapping $p_0(\mathfrak{o}) = 0 \ \forall \ \mathfrak{o} \in \mathfrak{O}$ then the slippage is simply the total traded notional value. The arrival mid price is also clearly not achievable. Even the average of the midpoint price over the lifetime of a parent order is not achievable without some special short-term alpha. How exactly are we supposed to consistently transact between the bid and the offer on a sequence of trades that are all in the same direction? On a sequence of trades which includes buys and sells in roughly equal amounts, the vwap of those trades might be closer to the midpoint.

Another interesting benchmark is "actual duration vwap." This is computed after the algo has finished executing the parent order. Let $[t, t']$ be the interval over which, in the end, the order was executed. The actual-duration vwap is the vwap of all of the trades which occurred in the market over the same interval $[t, t']$. Low slippage to actual-duration VWAP is not necessarily a good thing! If you are causing huge market impact because you are a large percent of the volume, then the vwap of your order will be close to the actual-duration vwap.

For orders that are executed incrementally over the course of an entire day, the full-day vwap is a popular benchmark price. Like the other benchmark prices discussed above, this one is not exactly achievable for all orders. No algo can guarantee that the vwap of your order will equal the aggregate market vwap of the day.

A very simplistic implementation of an algo benchmarked to vwap might be as follows. If the length of the trading day is $5K$ minutes, divide the trading day up into K five-minute bins and let p_i be the fraction of the day's volume that you predict will occur within each bin, so $\sum_{i=1}^{K} p_i = 1$ and all $p_i \geq 0$. For an order with total quantity $N = \sum_j n_j$, plan to execute $p_i N$ within the i-th bin. When there is no "interesting" activity going on in a given stock, then algos of this sort can achieve the vwap plus noise, where the noise mostly comes from the difference between the predicted intraday volume pattern and the realized one.

Suppose bad news comes out at close minus 5 minutes. This causes the volume to rise continuously for the next 5 minutes as the price is falling. The VWAP algo did not foresee this, so already executed a fraction of $1 - p_K$ of the order, where p_K is the typical/predicted fraction of volume in the last 5 minutes. Due to the news, the fraction of volume in the last 5 minutes today is much larger than p_K. It follows that, by the time the dust settles at the end of the day, the vwap algo will buy at a higher price than the full-day vwap (so it will underperform the full-day vwap benchmark).

This effect is symmetrical – if the news was good, your vwap algo would have bought at a lower (hence better) price than the full-day vwap. Over time, the aggregate slippage to full-day vwap is then related to how often the direction of your trades were "on the right side" of the news announcements which cause the most volume. This is probably also related to your P/L over the period!

7.3. **Slippage as a Cost.** Let's now consider a concrete example in which we buy a stock that is going up, and then later sell it, and our benchmark is arrival mid price. Assume that immediately before the buy order begins execution, the bid and the ask are $b_0 < a_0$ and immediately after the subsequent sell order begins execution, the bid and the ask are $b_1 < a_1$. We also assume that both transactions are for 100 shares, and also that $a_0 < b_1$ so the transaction will be profitable.

Hence the benchmark prices are the mids

$$m_0 = (a_0 + b_0)/2 \text{ and } m_1 = (a_1 + b_1)/2.$$

Assume that we are trading aggressively, hence we buy at the ask and sell at the bid, so

$$\mathrm{vwap}(\mathfrak{o}_0) = a_0, \quad \mathrm{vwap}(\mathfrak{o}_1) = b_1.$$

Let π denote our P&L, then

$$\pi = 100(b_1 - a_0) = 100(m_1 - m_0) - \mathrm{slip}(\mathfrak{o}_0) - \mathrm{slip}(\mathfrak{o}_1) \tag{7.2}$$

where $\mathrm{slip}(\mathfrak{o}_0) = 100(a_0 - m_0)$ and $\mathrm{slip}(\mathfrak{o}_1) = 100(m_1 - b_1)$. Note that as per our convention $\mathrm{slip}(\mathfrak{o}_0)$ and $\mathrm{slip}(\mathfrak{o}_1)$ are both positive. Eq (7.2) is easily verified by simple arithmetic.

Furthermore,

$$\begin{aligned} \pi &= 100(b_1 - a_0) = hR - \mathrm{slip}(\mathfrak{o}_0) - \mathrm{slip}(\mathfrak{o}_1), \\ R &:= \frac{m_1 - m_0}{m_0}, \quad h := 100 m_0. \end{aligned} \tag{7.3}$$

We can interpret (7.3) as stating that if we price our intended holding of 100 shares at the arrival mid, so that our intended holding is worth $h = 100m_0$ dollars, then the P&L π can be represented as the holding value times the return R (which must be price-return using the benchmark price!) minus the total slippage from both orders.

We have proven the following.

Lemma 7.1. Over a sequence of trades which begin and end with zero holdings, the P/L can be represented as the holding value times the return R (defined as price-return using the benchmark price) minus the total slippage from all orders.

Eq.(7.3) generalizes to portfolios in the following way. Suppose that we hold portfolio $h_0 \in \mathbb{R}^n$ now and intend to trade into portfolio $h \in \mathbb{R}^n$. Suppose there are two times t_0 and t_1, and at t_0 we will begin trading from h_0 to h at t_0, we will reach h before t_1, and then at t_1 we will begin liquidating h. Let $\mathfrak{O}_i$ denote all orders started at t_i. Then the P&L can be written

$$\pi = h'R - \mathrm{slip}(\mathfrak{O}_1) - \mathrm{slip}(\mathfrak{O}_2) \tag{7.4}$$

where $R \in \mathbb{R}^n$ is the vector of returns over the interval $[t_0, t_1]$ computed with respect to benchmark price.

We could equivalently write

$$\pi = h'R - \mathrm{slip}(h_0, h) - \mathrm{slip}(h, 0) \tag{7.5}$$

where $\mathrm{slip}(x, y)$ denotes the slippage of the orders needed to trade from portfolio x into portfolio y. The second term $\mathrm{slip}(h, 0)$ is the slippage incurred from liquidating the final portfolio h. As in the simple example above, it is necessary to liquidate the final portfolio to actually realize all profits in dollars; otherwise some portion of the profits will be left as "unrealized" and any unrealized profits will be subject to slippage before they are "realized" or translated to dollars.

Definition 7.8. *Liquidation slippage* of a portfolio h is defined as $\mathrm{slip}(h, 0)$, i.e. the slippage incurred on the full set of orders necessary to convert the holdings entirely to cash. The liquidation slippage of h will be denoted by

$$\mathrm{liqslip}(h) := \mathrm{slip}(h, 0).$$

For a perspective on optimal execution algos with fairly similar notation to ours, see Almgren and Chriss (1999) and Almgren and Chriss (2001).

Note that $\mathrm{slip}(\mathfrak{o})$ is not knowable at the order creation time τ (as it involves future prices). For the same reason (it involves future prices), it is hard to predict with high R^2.

Definition 7.9. A *predictive slippage model* is a model for the conditional density $p(\mathrm{slip}(\mathfrak{o}) \,|\, I_\tau)$ where I_τ denotes the information set available at time τ. Many researchers simply model $\mathbb{E}[\mathrm{slip}(\mathfrak{o}) \,|\, I_\tau]$ directly without modeling the full distribution.

A number of prominent academics have studied the problem of predicting $\mathrm{slip}(\mathfrak{o})$ as function of the order quantity q and attributes of the asset being traded. Some attributes that have been found to be predictive include that asset's volatility, volume, and the window $T = [\tau, \tau']$ over which the orders are filled. One of the most oft-cited such studies is Almgren et al. (2005).

Let's now suppose that our prior holding in some asset is h_0 dollars and we are considering a trade of

$$\delta := h - h_0$$

so that our new holding will be h. Suppose we translate δ into a quantity of shares q using the arrival price, so that up to roundoff errors, $q = \delta/p_0$.

If we assume that the order *will be fully executed* then we can algebraically manipulate the definition (7.1) to express it in terms of price return and order value:

$$\begin{aligned} \text{slip}(\mathfrak{o}) &= \delta \cdot R_S(\mathfrak{o}) \quad \text{where} \\ R_S(\mathfrak{o}) &:= \frac{\text{vwap}(\mathfrak{o}) - p_0}{p_0} \end{aligned} \tag{7.6}$$

The quantity $R_S(\mathfrak{o})$ will be referred to by me as *slippage price return* – I do not know if this is standard terminology. The quantity $R_S(\mathfrak{o})$ is defined in such a way that the dollar slippage equals return on (signed) dollars traded, using this number as the return.

The slippage price return is called J in the paper of Almgren et al. (2005), and is referred to there as "realized impact". The model of Almgren et al. (2005) for slippage price return is

$$J = R_S(\mathfrak{o}) = \frac{1}{2}\gamma\sigma\frac{q}{V}\left(\frac{\Theta}{V}\right)^{1/4} + \text{sgn}(q)\eta\sigma\left|\frac{q}{VT}\right|^{3/5}$$

where V is average daily share volume, Θ is shares outstanding, and γ, η are constants that must be fit to market data. Almgren et al. (2005) denote the quantity traded, q, instead by X. Note that the predicted slippage price return always has the same sign as the order, hence $\mathbb{E}[\text{slip}(\mathfrak{o})]$ is always positive in this model.

Noting that $\delta = p_0 q$ we can take the expression for J above and re-write it in terms of dollar-denominated quantities, by multiplying both terms by $1 = p_0/p_0$, which gives

$$\begin{aligned} R_S(\mathfrak{o}) &= \frac{\gamma}{2}\sigma\frac{\delta}{\text{advp}}\left(\frac{\Theta}{V}\right)^{1/4} + \text{sgn}(\delta)\eta\sigma\left|\frac{\delta}{\text{advp}\cdot T}\right|^{3/5} \\ \text{slip}(\delta) &= \delta \cdot R_S(\mathfrak{o}) = \sigma\left[\frac{\gamma}{2}\frac{\delta^2}{\text{advp}}\left(\frac{\Theta}{V}\right)^{1/4} + \eta|\delta|\left|\frac{\delta}{\text{advp}\cdot T}\right|^{3/5}\right] \end{aligned}$$

where $\text{advp} := p_0 V$ denotes average daily volume times price.

Someone desiring mathematical simplicity over empirical goodness-of-fit might take the approach of ignoring the more complicated second term in $\text{slip}(\delta)$ and re-fitting the coefficient γ in the first term to account for the second term's absence. In this modified model, the impact would be a purely quadratic function of δ, as assumed by Gârleanu and Pedersen (2013).

When a parent order (or metaorder) is finished executing, if you've had significant impact on the price, then typically that impact will revert somewhat once the price pressure that you were creating has been removed.

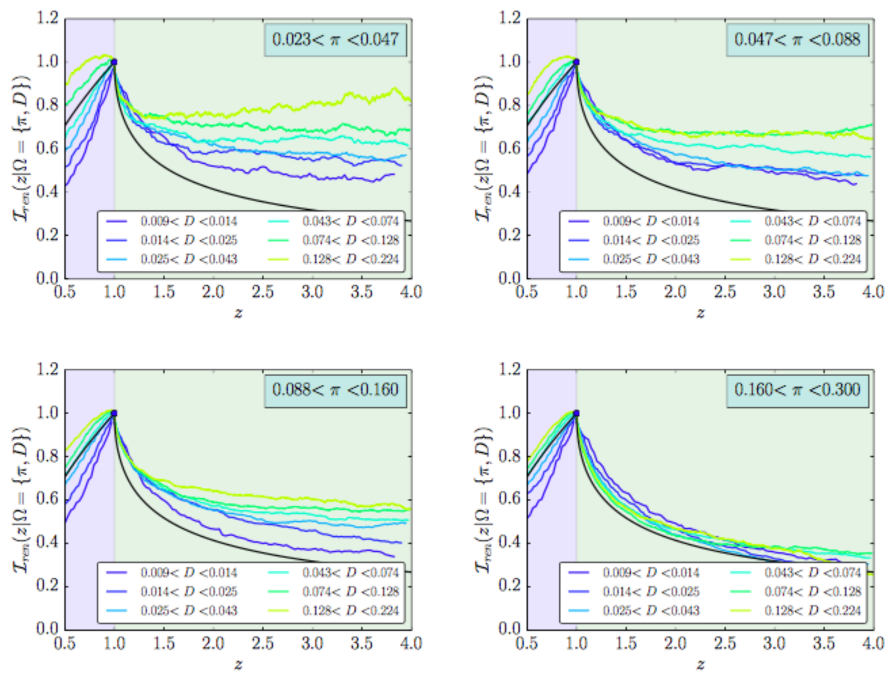

FIGURE 7.3. From Zarinelli et al. (2015). Decay of temporary market impact after the execution of a metaorder.

Within each panel the solid lines correspond to the average market impact trajectory for metaorders with different durations D; the four panels correspond to different participation rates. The black line corresponds to the prediction of the transient impact model with $\delta = 0.5$.

7.4. **Optimization with Costs.** Mean-variance optimization is concerned with the problem

$$\max_h \left\{ \mathbb{E}[\pi(h)] - \frac{\kappa}{2}\mathbb{V}[\pi(h)] \right\}$$

where

$$\pi(h) = h'R - [\text{slip}(h_0, h) + \text{liqslip}(h)]$$

and the latter expression for π is of course copied from (7.5). Suppose we assume that $\mathbb{V}[\pi(h)]$ is well approximated by the variance of the term $h'R$

in $\pi(h)$, in other words

$$\mathbb{V}[\pi(h)] \approx h'\Sigma h \text{ where } \Sigma := \mathrm{cov}(R).$$

Then the mean-variance problem becomes

$$\max_h \left\{ h'\mathbb{E}[R] - \frac{\kappa}{2} h'\Sigma h - \mathbb{E}[\mathrm{slip}(h_0, h) + \mathrm{liqslip}(h)] \right\} \tag{7.7}$$

Note that (7.7) can equivalently be considered as an optimization over the *trade list*

$$\delta := h - h_0 \in \mathbb{R}^n$$

since h_0 is fixed, and is not a parameter in the optimization. The trade list δ is the more natural variable; $\mathrm{slip}(h_0, h)$ is a function of δ, not a function of h. The same is true for commissions.

Financing costs, on the other hand, are functions of h rather than δ. You pay to finance the portfolio you end up with. On the short side, financing costs are usually called *borrow costs*, and can be quite high for stocks that are hard to borrow. On the long side, you will pay to finance a portfolio that is larger (in notional terms) than your capital, usually at the rate of 25bps per year times financed notional value.

Note that there are (at least) two functional forms for the latter two terms in (7.7) which allow for easy solution of the mean-variance maximization problem: (1) purely quadratic, and (2) quadratic plus absolute-value type penalty terms. In the first case, the entire problem remains quadratic, while in the second case, the problem becomes mathematically equivalent to a Lasso regression. The Almgren et al. (2005) form does not lead to such a well-known procedure as Lasso, but the associated problem is convex and differentiable, hence standard optimization routines can be expected to perform well.

References

Almgren, Robert and Neil Chriss (1999). “Value under liquidation”. In: *Risk* 12.12, pp. 61–63.

– (2001). “Optimal execution of portfolio transactions”. In: *Journal of Risk* 3, pp. 5–40.

Almgren, Robert et al. (2005). “Direct estimation of equity market impact”. In: *Risk* 18.7, pp. 58–62.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel (2005). "Limit order book as a market for liquidity". In: *The review of financial studies* 18.4, pp. 1171–1217.

Gârleanu, Nicolae and Lasse Heje Pedersen (2013). "Dynamic trading with predictable returns and transaction costs". In: *The Journal of Finance* 68.6, pp. 2309–2340.

Gould, Martin D et al. (2013). "Limit order books". In: *Quantitative Finance* 13.11, pp. 1709–1742.

Zarinelli, Elia et al. (2015). "Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate". In: *Market Microstructure and Liquidity* 1.02, p. 1550004.