# Execution Costs

Execution costs are the difference in value between an ideal trade and what was actually done. The execution cost of a single completed trade is typically the difference between the final average trade price, including commissions, fees and all other costs, and a suitable *benchmark* price representing a hypothetical perfectly executed trade. The sign is taken so that positive cost represents loss of value: buying for a higher price or selling for a lower price. If a trade is not completed either for endogeneous reasons (for example, the price moves away from an acceptable level) or for exogeneous reasons (a trader gets sick or a system fails), then some value must be assigned to the unexecuted shares. The cost of a portfolio transaction, or a series of transactions, is computed as a suitably weighted average of the costs of the individual executions.

Some of the costs of trading are direct and predictable, such as broker commissions, taxes, and exchange fees. Although these costs can be significant, they are commonly not included in the quantitative analysis of execution costs. "Indirect" costs include all other sources of price discrepancy, such as limited liquidity (market impact) and price motion due to volatility. These are much more difficult to characterize and measure, and are much more amenable to improvement. Since the technology of cost measurement is most advanced for equity trading, we shall use that language; cost measurement for other asset classes is very interesting but less developed.

The benchmark is commonly taken to be the *arrival price*, that is, the quoted market price in effect at the time the order was released to the trading desk. Using that benchmark is equivalent to saying that a perfect trade, one with zero execution cost, would be one that executed instantaneously at the arrival price. The cost measured using the arrival price benchmark is called the *implementation shortfall*, a term introduced by Perold [1988]: "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."

For example, suppose that an overnight investment decision assumed that a large quantity of stock could be purchased at the previous day's closing price of \$50 per share; if the trade was fully completed at an average price of \$50.25, then the execution cost would be reported as 25 cents per share. But execution costs are only part of the picture: if the stock closed that day at \$51, then the trade would be successful despite its positive cost; a naïve cost model might assign \$1 profit to the portfolio manager and \$0.25 cost to the trader. Execution costs can be negative, for example, if the price dropped in the course of a purchase program and the asset was acquired for a lower price than anticipated; or in the above example, if the benchmark price were the day's close. The forecast execution costs on any particular order typically have a very high degree of uncertainty, due to market volatility and other random effects.

A well-calibrated model for execution costs is an important part of the quantitative investment process. At a minimum, it is a tool for the portfolio manager to evaluate the performance of his or her trading desk and external brokers: were the results achieved on a particular execution compatible with the costs estimated from the pretrade model? Furthermore, anticipated transaction costs should be a component of the portfolio formation decisions: turnover should be minimized, and expected transition costs should be incorporated in the portfolio construction model along with expected alpha. Grinold and Kahn [1999] discuss in depth the use of transaction cost models in investment management.

This note is divided into three parts, corresponding to the order in which the three aspects should be addressed in designing an investment process, although the order is reverse chronological from the point of view of a single trade. First we shall look at post-trade cost reporting, then on optimal trading to minimise execution costs, and finally on pre-trade cost estimation.

## Post-trade reporting

"If you can not measure it, you can not improve it," said Lord Kelvin. The first step in any program to reduce execution costs is to measure them systematically. For each trade executed, the cost should be reported relative to a collection of benchmarks. In addition, the cost statistics should be computed across all trades in a suitable time period (daily or weekly) and broken down by any relevant parameters: primary market, size of trade, market capitalization of stock, *etc.*

As noted above, the most common benchmark is the pretrade arrival price. Another common choice is the "volume-weighted average price" (VWAP), taken across

the time interval during which the trade was executed. Although this most likely does not correspond directly to an investment goal, it is a popular benchmark for assessing the quality of the execution, because it largely filters out the effects of volatility. One would typically also use a post-trade price, for example the closing price on the day during which the trade was executed. Typical post-trade reporting systems display the execution price relative to all of these benchmarks: before, during, and after trading.

When aggregating cost numbers across a diverse variety of trades, the individual cost numbers should be weighted so that the result is representative of the overall change in portfolio value. If the individual costs are measured in cents per share, then they should be weighted using the number of shares in each individual trade. If the individual costs are measured in basis points, then they should be weighted using the dollar value of each trade.

If a pretrade cost model has been developed, then realized costs should also be compared with the forecast values, both on the level of individual execution, and on the overall portfolio level. This will help to identify trades that may have been badly executed, as well as maintaining accurate calibration of the pre-trade model.

In addition to the average, it is useful to report the standard deviation of the costs. This is useful as a reality check for the significance of the mean. For example, if the sample standard deviation were 25 basis points on 100 independent trades, then the expected error in the sample mean would be 2.5 basis points and a change of one or two basis points in the mean cost would not be significant. The standard deviation should also be weighted by trade size; the most reasonable weights are the same that are used for the average cost. The standard deviation does not have good properties under subdivision.

Ideally, aggregate cost numbers would be reported so that the result is indifferent to subdivision of trade blocks, but this is often not possible. For example, suppose that 100,000 shares are purchased throughout the day; a natural benchmark price would be the day's open price. But if this block were considered as 40,000 shares in the morning and 60,000 shares in the afternoon, the arrival price for the second block would be the midday price, and the overall reported cost would be lower. The choice can be made only with knowledge of the investor's overall goals. This difficulty does not arise with a VWAP benchmark or with a benchmark price at a fixed time such as at the close.

Additional difficulties in cost reporting come from price limits and incomplete trades. Suppose that a buy order is put in for a stock currently trading at \$50, but a condition is imposed that no shares are to be bought at a price higher than \$50.05. It will then be certain that the price impact on this trade will be at most 5 cents per share, but it may be that only a small fraction of the requested shares are actually executed. If the limit is below the initial price the situation may be even more extreme. Similar difficulties arise if the trade is halted for other reasons.

The solution to this difficulty rests on the valuation of unexecuted shares, but there is no simple rule. The most straightforward would be to imagine that the unexecuted shares were purchased at the day's closing price. In practice this rule is far too stringent and does not take account of the variety of possible reasons the trade might have been halted. Trade cost reporting must take account of the entire investment process.

## Optimal trading

Once a system is in place for measuring and reporting trade costs, and after it has been agreed what criteria define a good trade or collection of trades, then one can design optimal strategies to meet investment goals. "Best execution" is not only advantageous to investors, but also is required by regulation in most markets. The most important goal is always to reduce mean execution cost, which is largely due to market impact and uncaptured short-term alpha. One also often desires to reduce the standard deviation of trade costs in order to reduce overall investment volatility[Engle and Ferstenberg, 2007].

Reducing costs is achieved by searching for liquidity in "space" as well as in time, accessing as many pools of potential liquidity and as many potential counterparties as possible for each piece of the trade. This includes routing to non-exchange trading venues such as "dark pools" and block crossing services when possible. Rapid changes in market structure make this increasingly complicated (see Hasbrouck [2007, Appendix] for a discussion of the US equities markets).

Accessing liquidity in time means being willing to slow trading to give potential counterparties the opportunity to appear in the market. If short-term price drift is not expected to be significant, then average trading costs can generally be reduced by trading more slowly.

Other aspects of trading push for rapid trading, most significantly anticipated price drift and market volatility: "To trade a list of stocks efficiently, investors must balance opportunity costs and execution price risk against market impact costs. Trading each stock quickly minimizes lost alpha and price uncertainty due to delay, but impatient trading incurs maximum impact. In contrast, trading more patiently over a longer period reduces market impact but incurs larger opportunity costs and short-term execution price risk." [Alford et al., 2003] The actual trade schedule is determined using a quantitative balance of all of these factors. This schedule may need to be dynamically adapted depending on observed liquidity or other market components. For a portfolio, correlation between the assets should be included in addition to the volatility of each one, and the schedule may need to maintain strict neutrality to a market index or other risk factor.

In practice, optimal strategies for a mean-variance trade-off, and strategies that are calculated to optimize expected cost in the presence of short-term drift, are generally similar: at the beginning of each execution, rapid trading reduces the exposure to volatility or alpha, then the trading slows to reduce overall impact costs (Figure 1, from Almgren and Chriss [2000]). The most important question is the overall speed of the execution: should it be completed in minutes, hours or days? Regardless of the model, an approximate quantitative model for execution costs is an essential element of this decision.

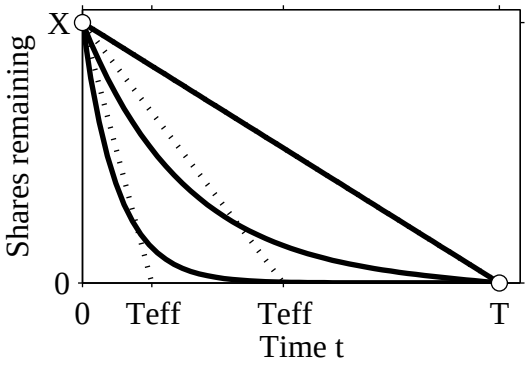


Figure 1: Optimal trading trajectories with varying degrees of urgency (solid curves). Regardless of the exact analytic form of the curve, the most important variable is the approximate effective duration $T_{\text{eff}}$ corresponding to a linear approximation of the trajectory and represented here by a broken line.

## Pre-trade cost estimation

After a measurement system has been running for long enough to generate a useful amount of data, then one can begin to think about developing an analytic model. The goal of the model is to forecast the execution cost to be expected on any anticipated trade, along with an estimate of the uncertainty in the forecast. In its most general form, such a model must contain a full explanation of how trading in markets affects prices. This is a rich area of research with a large literature (see Madhavan [2000] for an extensive survey; see Bouchaud et al. [2004] and Lillo et al. [2003] for more subtle models). Thus one typically poses the problem in the more prosaic terms of estimating the future in terms of past data.

The goal is to give the predicted cost $C$ (in cents per share or basis points) in terms of input variables, including at a minimum the number of shares traded $X$ and the daily volume $V$ of the stock (either historical average or volume on the day of trading), and the effective duration $T$ over which the trade was executed as in Figure 1. In order to normalize across many different stocks, one should also include properties such as volatility $\sigma$, and perhaps market capitalization, primary exchange or country, and other economic parameters. In addition, one might include trade information such as the anticipated short term alpha, and which algorithm or other means was used to trade the stock.

For any proposed trade, the pre-trade model predicts the cost based on cost that was measured on "similar" trades in the past. "Supervised learning" algorithms [Poggio and Smale, 2003] could in theory be used to carry out this classification. In practice, because the dimension of the parameter space is so large, there may be no, or very few, trades that are sufficiently similar to the proposed ones in all variables. This argues for a regression approach in which a specific functional form is proposed, calibrated to data, and evaluated using standard statistical tests of significance.

The most detailed such study was done by Almgren et al. [2005]. This model is based on separation of the

cost into a "permanent" and a "temporary" component. Although both components contribute to the realized cost on every trade, it is useful to calibrate each separately. The calibration relies on three observed prices. $S_{\text{pre}}$ denotes the arrival price, that is, a market price just before the order begins trading; typically this would be the bid-ask midpoint before the first execution. $S_{\text{exec}}$ is the actual average price for which the order finally executes, which is of course the quantity of most interest to the trader. $S_{\text{post}}$ denotes a market price just after the order has finished executing, possibly with a short time lag to allow transient effects to dissipate.

The model of Almgren et al. [2005] measures two quantities $I$ and $J$, for permanent and temporary impact respectively. For a buy order, these are

$$I = \log \frac{S_{\text{post}}}{S_{\text{pre}}} \approx \frac{S_{\text{post}} - S_{\text{pre}}}{S_{\text{pre}}}$$
$$J = \log \frac{S_{\text{exec}}}{S_{\text{pre}}} \approx \frac{S_{\text{exec}} - S_{\text{pre}}}{S_{\text{pre}}},$$

and for a sell order, the signs would be reversed. The approximate equalities are valid for moderate-sized orders, for which the price does not move more than a few percent during execution. These quantities represent price changes relative to the pre-trade price, as fractions of that benchmark.

The permanent component of cost is interpreted as the net displacement of the market due to the buy-sell imbalance introduced by this particular trade. The simplest model sets

$$I \;=\; \gamma\sigma \frac{X}{V} + \langle \text{noise} \rangle.$$

In this expression, we normalize the trade size $X$ by the day's volume $V$, so $X/V$ is the trade size as a fraction of a typical day's flow. We also normalize the impact $I$ by the volatility $\sigma$, so that we represent the impact as a fraction of a typical amount that the stock moves without our trading. The coefficient $\gamma$ is taken to be constant across all stocks, with widely varying daily volumes and volatilities. The noise term arises from intraday volatility: the actual cost experienced on any particular trade may be very different from the mean prediction from the model.

In this form of the model, the permanent impact is linear in the total number of shares traded. This model is particularly convenient for theoretical modeling but must be justified by empirical analysis of real trade data. Almgren et al. [2005] found a reasonable agreement with their data, but other recent empirical studies have suggested that the linear form might not be justified.

The temporary impact component of cost is interpreted as the additional premium that must be paid for execution in a finite time, above a suitably prorated fraction of the permanent cost. We model it as

$$J \;=\; \frac{I}{2} + \eta\,\sigma \left( \frac{X}{VT} \right)^{\beta} + \langle \text{noise} \rangle$$

where $T$ represents the effective duration of the trade as a fraction of the trading day. Thus $X/VT$ represents the "participation rate", or the fraction of market flow that this trade constitutes during the time that it is active. The factor $1/2$ on the permanent cost component represents the fraction of post-trade impact that is paid on the trade itself. As for the permanent cost, the impact cost is expressed as a fraction of typical volatility; the coefficient $\eta$ is taken constant across all stocks. Other factors such as bid-ask spread and market capitalization were not determined to be significant in US equity markets.

Almgren et al. [2005] calibrate this model to a large sample of US equity trades using a two-step procedure. First, the permanent impact term $I$ is calibrated, testing the hypothesis of linear impact and estimating the value of $\gamma$. Next, the temporary impact term $J$ is calibrated, determining values for the exponent $\beta$ and the coefficient $\eta$. A key aspect of the verification is characterization of the error terms to verify that volatility is an adequate explanation for the residuals. The result has $\beta \approx 0.6$, which is roughly compatible with earlier square-root models [Bar, 1997], as well as values for the coefficients $\gamma$ and $\eta$. For trades that are a few percent of daily volume executed across several hours, the predicted impact costs are tens of basis points.

A number of factors make such models at best very approximate. First is the extremely low values of $R^2$ in the regression, typically around a few percent at best. That is, the ability of the model to predict the cost of any single trade is very poor, because market volatility due to other trading activity is usually very large. Second is the difficulty of the model in distinguishing market impact from alpha: did the price move up because the buy program impacted the price, or was the trade executed because the

manager correctly anticipated a price rise? A third difficulty is different behavior between small and large trades: although the model claims to be universally valid, in practice a model developed for small trades gives poor results on large trades.

In any particular application, the model should be critically evaluated by the user and recalibrated and extended as necessary. Despite its intrinsic difficulties and limitations, this model and extensions are extremely useful to give approximate anticipated cost values for pre-trade planning, optimal trade scheduling, and post-trade evaluation.

## References

A. Alford, R. Jones, and T. Lim. Equity portfolio managment. In B. Litterman, editor, *Modern Investment Management: An Equilibrium Approach*. Wiley, 2003.

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

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

*Market Impact Model Handbook*. Barra, 1997.

J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart. Fluctuations and response in financial markets: The subtle nature of 'random' price changes. *Quant. Finance*, 4 (2):176–190, 2004.

R. Engle and R. Ferstenberg. Execution risk: It's the same as investment risk. *J. Portfolio Management*, 33(2):34–44, 2007.

R. C. Grinold and R. N. Kahn. *Active Portfolio Management*. McGraw-Hill, 2nd edition, 1999.

J. Hasbrouck. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, 2007.

F. Lillo, J. D. Farmer, and R. N. Mantegna. Master curve for price-impact function. *Nature*, 421:129–130, 2003.

A. Madhavan. Market microstructure: A survey. *J. Financial Markets*, 3:205–258, 2000.

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

T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Notices Amer. Math. Soc.*, 50(5): 537–544, 2003.

ROBERT ALMGREN