Optimal Control of Execution Costs for Portfolios*

Dimitris Bertsimas†, Paul Hummel††, and Andrew W. Lo‡

This Draft: September 12, 1999

Abstract

The dramatic growth in institutionally managed assets, coupled with the advent of internet trading and electronic brokerage for retail investors, has led to a surge in the size and volume of trading. At the same time, competition in the asset management industry has increased to the point where fractions of a percent in performance can separate the top funds from those in the next tier. In this environment, portfolio managers have begun to explore active management of trading costs as a means of boosting returns. Controlling execution cost can be viewed as a stochastic dynamic optimization problem because trading takes time, stock prices exhibit random fluctuations, and execution prices depend on trade size, order flow, and market conditions. In this paper, we apply stochastic dynamic programming to derive trading strategies that minimize the expected cost of executing a portfolio of securities over a fixed period of time, i.e., we derive the optimal *sequence* of trades as a function of prices, quantitites, and other market conditions. To illustrate the practical relevance of our methods, we apply them to a hypothetical portfolio of 25 stocks by estimating their price-impact functions using historical trade data from 1996 and deriving the optimal execution strategies. We also perform several Monte Carlo simulation experiments to compare the performance of the optimal strategy to several alternatives.

*This research is part of the MIT Laboratory for Financial Engineering's Transactions Costs Project. We are grateful to Investment Technology Group, the National Science Foundation (Grant Nos. DMI–9610486 and SBR–9709976), and the sponsors of the LFE for financial support. We thank Dave Cushing, Chris Darnell, Robert Ferstenberg, Rohit D'Souza, John Heaton, Leonid Kogan, Bruce Lehmann, Greg Peterson, Jiang Wang, and seminar participants at Columbia, Harvard, London Business School, MIT, Northwestern, the NYSE Conference on Best Execution, the Society of Quantitative Analysts, and Yale for helpful comments and discussion.

†Boeing Professor of Operations Research, MIT Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, E53-359, Cambridge, MA 02142–1347.

††Portfolio Strategist, Long Term Capital Management, One East Weaver St., Greenwich, CT 06831.

‡Harris & Harris Group Professor, MIT Sloan School of Management and Operations Research Center, 50 Memorial Drive, E52–432, Cambridge, MA 02142–1347.

Contents

1 Introduction

The rapid growth in equity investing, driven by the increasing popularity of mutual funds and defined-contribution retirement plans, has led to a rising concentration of assets among institutional money managers. A typical portfolio manager now oversees a large portfolio of several hundred securities, with individual positions that might comprise a significant fraction of the average daily volume of the security. Both active managers and passive indexers must frequently rebalance their portfolios, either to include new stock picks, sell stocks that are out of favor, or to improve the tracking of a given index or benchmark. This generates sizeable orders across many stocks that must be executed within a relatively short time horizon, and that must be executed together so as to maintain the risk/reward characteristics of the portfolio. The transactions costs associated with trading such "lists" of securities—often called *execution* costs—can be substantial.

Execution costs are comprised of several components: explicit costs such as commissions and bid/ask spreads, and costs that are harder to quantify such as the opportunity cost of waiting and the price impact from trading. Opportunity costs arise because market prices are moving constantly and can move favorably or unfavorably without warning, generating unexpected profits or lost opportunities while a portfolio manager hesitates. Price impact is the typically unfavorable effect on prices that the very act of trading creates, not unlike the turbulence that is generated by a ship's own wake. The seller of a security will, by the very act of selling, push down the security's price, yielding lower proceeds from the sale, and similarly for the buyer. Moreover, the larger the order, the more heavily is the price affected by the trade. For portfolios that turn over frequently or with large positions to trade, these costs can add a significant drag on the fund's overall performance (see,for example, Loeb (1983)).

Recent studies by Chan and Lakonishok (1995) and Keim and Madhavan (1995a–c) show that institutional investors often break up their larger trades into smaller "packages" that they execute over the course of several days. There is a compelling economic rationale for "package" trading. Trading is fundamentally a dynamic, path-dependent, stochastic problem insofar as it takes time to trade and the very act of trading affects price and price dynamics which, in turn, affect execution costs. Controlling execution costs of large blocks of stock must be accomplished by trading over a number of time periods, a fact recognized by Bertsimas and Lo (1998) who used stochastic dynamic programming techniques to derive optimal or *best-execution* strategies.

In this paper, we extend the single-asset framework outlined by Bertsimas and Lo (1998) to construct best-execution strategies for portfolio problems. Specifically, we solve the following problem: Given fixed blocks of shares in n stocks, $\overline{\boldsymbol{s}} = [\ \overline{s}_1 \ \ \overline{s}_2 \ \cdots \ \overline{s}_n \ ]'$,[1] to be purchased within a fixed finite number of periods T, and given a set of price dynamics that capture price impact, i.e., the execution price of an individual trade as a function of the shares traded and other "state" variables, find the optimal *sequence* of trades (as a function of the state variables) that minimizes the expected execution costs.[2] Since it is well-known that the short-term demand curves for even the most actively traded equities are not perfectly elastic (see, for example, Shleifer (1986)), a market order at date 0 for the entire block $\overline{\boldsymbol{s}}$ is clearly not an optimal trading strategy.

Denote by $\boldsymbol{s}_t = \{s_{1t}, s_{2t}, \ldots, s_{nt}\}$ the number of shares of each stock acquired in period t at prices $\boldsymbol{p}_t = \{p_{1t}, p_{2t}, \ldots, p_{nt}\}$, where $t = 1, \ldots, T$. Then the investor's objective may be expressed as:

$$\underset{\{\boldsymbol{s}_t\}}{\text{Min}} \ \mathrm{E}_1 \left[\ \sum_{t=1}^{T} \boldsymbol{p}_t' \boldsymbol{s}_t \ \right] \tag{1.1}$$

subject to:

$$\begin{aligned} \boldsymbol{p}_t &= f(\boldsymbol{p}_{t-1}, \boldsymbol{x}_t, \boldsymbol{s}_t, \boldsymbol{\epsilon}_t) && (1.2) \\ \boldsymbol{x}_t &= g(\boldsymbol{x}_{t-1}, \boldsymbol{\eta}_t) && (1.3) \\ \sum_{t=1}^{T} \boldsymbol{s}_t &= \overline{\boldsymbol{s}} && (1.4) \end{aligned}$$

where $\boldsymbol{x}_t$ is a vector of state variables, $\boldsymbol{\epsilon}_t$ is vector white noise, and $f(\cdot)$ and $g(\cdot)$ are the state equations or laws of motion that incorporates both price dynamics of $\boldsymbol{p}_t$, the price impact of

[1] We follow the common convention that all vectors are *column* vectors unless they are explicitly transposed, and boldface Roman and Greek letters denote vectors and matrices.

[2] For simplicity and without loss in generality, we consider the case of *purchasing* $\overline{\boldsymbol{s}}$ only. Selling $\overline{\boldsymbol{s}}$ and a combination of buying certain stocks and selling others can easily be accommodated with the appropriate sign conventions (positive numbers for purchases, negative for sales).

trading $\boldsymbol{s}_t$, and the dynamics of the state variables. We may also wish to impose additional constraints, e.g., a no-sales constraint, $\boldsymbol{s}_t \geq \mathbf{0}$,[3] or other conditions that are placed upon the portfolio manager by institutional restrictions, tax considerations, or other aspects of his investment process.

The portfolio case contains several interesting features not captured by the single-stock analysis of Bertsimas and Lo (1998) and others. Perhaps the most important feature is the ability to capture cross-stock relations such as the cross-autocorrelations reported in Lo and MacKinlay (1990). Price movements in one stock can induce similar movements in the price of another, either because of common factors driving both, or because of linked trading strategies, e.g., pairs trading, index arbitrage, risk arbitrage, etc. In such cases, the price impact of trading a portfolio may be larger than the sum of the price impact of trading the individual stocks separately. Alternatively, if some stocks are negatively correlated (perhaps because of portfolio substitution effects), or if the portfolio to be executed includes both purchases and sales, then the portfolio execution cost may be *lower* than the sum of the individual stocks' execution costs due to a kind of diversification effect in which the trades of one stock lower the price impact of trades in another. Whether execution costs are magnified or mollified in the portfolio case is, of course, an empirical issue that turns on the law of motion for the vector of prices and state variables. In either case, it is clear that the portfolio setting is considerably more complex than the single-stock case.

In Section 2 specify the state equations and explore in greater detail one particular specification that captures cross-stock interactions for a portfolio of stocks, allows for predictability in asset returns, and incorporates a "temporary" price-impact effect. The best-execution strategy is derived in closed form in Section **??** using stochastic dynamic programming, and in Section **??** we propose an approximation algorithm that provides a near-optimal execution strategy for the case with additional constraints. In Section **??** we evaluate the practical relevance of our approach by applying it to a portfolio of 25 stocks using historical transactions date from 1996 and estimating the execution costs for such a portfolio under several

[3]If a portfolio manager is attempting to acquire a block of securities, it is difficult to justify selling the same securities during the acquisition period (unless, of course, the manager has extremely accurate negative information regarding the security's price, which is somewhat inconsistent with the original premise that he is a buyer). Indeed, in many cases, it is illegal because it is viewed as a violation of the fiduciary trust that portfolio managers have to act in the best interests of their investors.

different trading scenarios. We conclude in Section ??.

2 The State Equations

In this section we present a specification for the state equations (1.2)–(1.4) that incorporates a multivariate price-impact function with cross-stock interactions (Section 2.1), and discuss other possible specifications afterwards (Section 2.2).

2.1 Linear Percentage Price-Impact

Let the execution price p_t be the sum of two components:

$$\boldsymbol{p}_t \;=\; \tilde{\boldsymbol{p}}_t \;+\; \boldsymbol{\delta}_t \tag{2.1}$$

where $\tilde{\boldsymbol{p}}_t$ may be viewed as a "no-impact" price—the price that would prevail in the absence of any market impact—and the impact $\boldsymbol{\delta}_t$. A plausible and observable proxy for the no-impact price is the midpoint of the bid/offer spread, although it can be arbitrary as long as it is unaffected by the trade size $\boldsymbol{s}_t$. For convenience, and to ensure non-negative prices, we model $\tilde{\boldsymbol{p}}_t$ as vector-geometric-Brownian-motion:

$$\tilde{\boldsymbol{p}}_t \;=\; \exp(\boldsymbol{Z}_t)\tilde{\boldsymbol{p}}_{t-1} \tag{2.2}$$

where $\boldsymbol{Z}_t$ is a diagonal matrix whose diagonal is a normal random vector $\boldsymbol{z}_t$ with mean $\boldsymbol{\mu_z}$ and covariance matrix $\boldsymbol{\Sigma}_z$. The $\exp(\cdot)$ operator denotes the matrix exponential which, in this case, reduces to the element-wise exponential of the diagonal entries in $\boldsymbol{Z}_t$.

For the price-impact term, $\boldsymbol{\delta}_t$, we set:

$$\boldsymbol{\delta}_t \;=\; \tilde{\boldsymbol{P}}_t(\boldsymbol{A}\tilde{\boldsymbol{P}}_t\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{x}_t) \tag{2.3}$$

where $\tilde{\boldsymbol{P}}_t = \text{diag}[\tilde{\boldsymbol{p}}_t]$ and $\text{diag}(\cdot)$ is the diagonalization operator which maps its vector argument into a diagonal matrix with the vector as its diagonal. This specification captures the impact of trading $\boldsymbol{s}_t$ shares on the transaction prices $\boldsymbol{p}_t$, and implies that as a percentage of the no-impact price, $\tilde{\boldsymbol{p}}_t$, the price impact is a linear function of the dollar-value of the trade and other state variables $\boldsymbol{x}_t$. The form of the price impact (2.3) differs from the single-stock case in that the percentage price-impact function for each stock i is a linear function, $\boldsymbol{A}\tilde{\boldsymbol{P}}_t\boldsymbol{s}_t$, of the dollar-values of the trades of *all* n stocks, not just of stock i. In the special case where $\boldsymbol{A}$ is diagonal, the portfolio problem reduces to n independent single-stock problems solved by Bertsimas and Lo (1998).

This specification of the dynamics of $\boldsymbol{p}_t$ has a number of advantages over other specifications (see Section 2.2). First, $\tilde{\boldsymbol{p}}_t$ is guaranteed to be non-negative, and $\boldsymbol{p}_t$ is also guaranteed to be non-negative under mild restrictions on $\boldsymbol{\delta}_t$. Second, by separating the transaction price $\boldsymbol{p}_t$ into a no-impact component $\tilde{\boldsymbol{p}}_t$ and the impact component $\boldsymbol{\delta}_t$, the price impact of a trade is temporary, affecting the current transaction price but having no affect on future prices. Third, the percentage price impact increases linearly with the trade size, which is empirically plausible (see, for example, Loeb (1983, Table II), Birinyi (1995), or Leinweber (1993, 1994)). Fourth, (2.1) implies a natural decomposition of execution costs, decoupling market-microstructure effects from price dynamics, which is closely related to Pérold's (1988) notion of implementation shortfall.[4] And finally, we shall see in Section **??** that (2.1) admits a closed-form solution in which the best-execution strategy is a simple linear function of the state variables and the optimal-value function is quadratic.

The presence of the vector $\boldsymbol{x}_t$ in (2.3) captures the potential influence of changing market conditions or private information about the securities on the price impact $\boldsymbol{\delta}_t$. For example, $\boldsymbol{x}_t$ might be the return on the S&P 500 index, a common component in the price of most equities. We model $\boldsymbol{x}_t$ as a vector with r elements, allowing for multiple sources of information to influence execution prices (or several lags of a single state variable). To complete our specification of the state equation, we must specify the dynamics of $\boldsymbol{x}_t$, and for simplicity we let:

$$\boldsymbol{x}_t \quad = \quad \boldsymbol{C}\boldsymbol{x}_{t-1} \ + \ \boldsymbol{\eta_t} \tag{2.4}$$

[4]Because (2.1) implies that price impact is a temporary phenomenon, affecting only the current trade price $\boldsymbol{p}_t$ and not the no-impact price level $\tilde{\boldsymbol{p}}_t$, the objective function (1.1) separates into two terms. The first term is the no-impact cost of execution and the second term is the total impact cost. This decomposition is precisely the one proposed by Pérold (1988) in his definition of implementation shortfall, but now applied to executing $\overline{\boldsymbol{s}}$. In particular, the "paper" return or execution cost is given by the first term, the actual cost is given by the sum of the two terms, hence the second term is the implementation shortfall in executing $\overline{\boldsymbol{s}}$.

where $\boldsymbol{\eta_t}$ is vector white noise with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\eta$. The fact that $\boldsymbol{x}_t$ is a vector autoregressive process with one lag (an "AR(1)") allows us to capture varying degrees of predictability in information or market conditions. The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ measure the sensitivity of price impact to trade size and market conditions. Matrix $\boldsymbol{A}$ must be positive definite, $\boldsymbol{B}$ is arbitrary, and must $\boldsymbol{C}$ have eigenvalues less than unity in modulus (to ensure stationary of $\boldsymbol{x}_t$).

2.2 Other Specifications

The specification for the state equation proposed in Section 2.1 is only one of many possible specifications. For example, Bertsimas and Lo (1998) propose a natural multivariate extension of their "linear price impact" specification in which the state equation is given by:

$$\boldsymbol{p}_t \;=\; \boldsymbol{p}_{t-1} + \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t \tag{2.5}$$

where $\boldsymbol{A}$ is a positive definite $(n \times n)$-matrix, $\boldsymbol{B}$ is an arbitrary $(n \times m)$-matrix, $\boldsymbol{x}_t$ is an m-vector of information variables, and $\boldsymbol{\epsilon}_t$ is n-vector white noise with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\epsilon$. As before, we assume that $\boldsymbol{x}_t$ follows a stationary AR(1) process, hence:

$$\boldsymbol{x}_t \;=\; \boldsymbol{C}\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t \tag{2.6}$$

where $\boldsymbol{C}$ is an $(m \times m)$-matrix with eigenvalues all less than unity in modulus, and $\boldsymbol{\eta}_t$ is m-vector white noise with mean $\mathbf{0}$, covariance matrix $\boldsymbol{\Sigma}_\eta$, and which is independent of $\boldsymbol{\epsilon}_t$.

This specification differs significantly from the linear-percentage price impact of Section 2.1 in a number of respects. Perhaps the most important difference is the fact that (2.5) implies that the price impact has a "permanent" effect on prices, due to the random-walk nature of the price dynamics. Also, the price impact is linear in price, implying that a 10,000-share trade will have the same dollar impact on \$1-stock as it will on a \$100-stock. Finally, unless some rather unnatural restrictions are placed on $\boldsymbol{\epsilon}_t$ in (2.5), there is a positive probability that $\boldsymbol{p}_t$ can take on negative values, clearly an unrealistic prospect.

Of course, Bertsimas and Lo (1998) consider (2.5) primarily for analytic tractability, not because of supporting empirical evidence. Nevertheless, it is instructive to compare this specification with the linear-percentage specification of Section 2.1 to develop some intuition for the more practical issues in specifying the state equation. In practice, the state equation must be estimated empirically. For example, several empirical studies seem to point to both permanent and temporary price-impact in US equity data, e.g., Kraus and Stoll (1972), Holthausen, Leftwich, and Mayers (1987, 1990), Barclay and Litzenberger (1988), Barclay and Warner (1993), and Chan and Lakonishok (1993, 1995). However, given the ever-changing nature of financial markets, it is crucial to re-estimate the state equation for each application using the most recent datasets available.