COMPUTATIONAL FINANCE

OPTIMAL CONTROL OF EXECUTION COSTS FOR PORTFOLIOS

The authors apply stochastic dynamic programming to derive trading strategies that minimize the expected cost of executing a portfolio of securities over a fixed time period. They test their strategies using real-world stock data.

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 constitute a significant fraction of the security's average daily volume. Both active managers and passive indexers must frequently rebalance their portfolios, to include new stock picks, to 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 portfolio's risk/reward characteristics. The transaction costs associated with trading such "lists" of securities—often called *execution* costs—can be substantial.

Execution costs have 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 act of trading creates, not unlike the turbulence that a ship's wake generates. A security's seller 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 the trade affects the price. For portfolios that turn over frequently or have large positions to trade, these costs can significantly hinder the fund's overall performance.[1]

Recent studies show that institutional investors often break up their larger trades into smaller "packages" that they execute over the course of several days.[2–5] There is a compelling economic rationale for package trading. Trading is fundamentally a dynamic, path-dependent, stochastic problem. Trading takes time, and the act of trading affects price and price dynamics, which, in turn, affect execution costs. Controlling the execution costs of large blocks of stock must be accomplished by trading over a number of time periods. This was recognized by Dimitris Bertsimas and Andrew Lo, who used stochastic dynamic-programming techniques to

1521-9615/99/$10.00 © 1999 IEEE

DIMITRIS BERTSIMAS AND ANDREW W. LO
Massachusetts Institute of Technology
PAUL HUMMEL
Long Term Capital Management

This article was edited to *CiSE* house style and space requirements. For the full version, contact Andrew Lo at the address listed for him on p. 23.

derive optimal or *best-execution* strategies.[6]

In this article, we extend the single-asset framework Bertsimas and Lo outlined, to construct best-execution strategies for portfolio problems. We developed a specification for these problems that is both empirically plausible and computationally tractable to implement. The closed-form solution provides insight into the nature of trading portfolios. To quantify the potential cost savings of our strategies, we fit the parameters of our price-impact model using historical data on 25 large-cap New York Stock Exchange (NYSE) stocks.

The portfolio problem

Specifically, we solve this problem:

- given fixed blocks of shares in n stocks, $\bar{s} = [\bar{s}_1\ \bar{s}_2 \dots \bar{s}_n]'$, to be purchased within a fixed finite number of periods T, and
- given a set of price dynamics that capture price impact (that is, an individual trade's execution price 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.

Because, as is well-known, the short-term demand curves for even the most actively traded equities are not perfectly elastic,[7] a market order at date 0 for the entire block $\bar{s}$ is clearly not an optimal trading strategy.

(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. For simplicity and without loss in generality, we consider the case of purchasing $\bar{s}$ only. Selling $\bar{s}$ 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].)

Let $s_t = \{s_{1t}, s_{2t}, \dots, s_{nt}\}$ be the number of shares of each stock acquired in period t at prices $\mathbf{p}_t = \{p_{1t}, p_{2t}, \dots, p_{nt}\}$, where $t = 1, \dots, T$. We can express the investor's objective as

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

subject to

$$\mathbf{p}_t = f(\mathbf{p}_{t-1}, \mathbf{x}_t, \mathbf{s}_t, \epsilon_t)$$

$$\mathbf{x}_t = g(\mathbf{x}_{t-1}, \eta_t)$$

$$\sum_{t=1}^{T} \mathbf{s}_t = \bar{\mathbf{s}} \tag{2}$$

where $\mathbf{x}_t$ is a vector of state variables, ϵ_t is vector white noise, and $f(\cdot)$ and $g(\cdot)$ are the state equations or laws of motion that incorporate the price dynamics of $\mathbf{p}_t$, the price impact of trading $\mathbf{s}_t$, and the dynamics of the state variables. We might also wish to impose additional constraints—for example, a no-sales constraint, $\mathbf{s}_t \geq 0$—or other conditions that are placed on the portfolio manager by institutional restrictions, tax considerations, or other aspects of his or her investment process.

(If a portfolio manager is attempting to acquire a block of securities, selling the same securities during the acquisition period is difficult to justify [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 considered a violation of the fiduciary trust that portfolio managers have to act in the best interests of their investors.)

The portfolio case contains several interesting features that the single-stock analysis of Bertsimas and Lo and others did not capture. Perhaps the most important feature is the ability to capture cross-stock relations such as the cross-autocorrelations reported by Andrew Lo and Craig MacKinlay.[8] Price movements in one stock can induce similar movements in the price of another, because of either common factors driving both or linked trading strategies—for example, pairs trading, index arbitrage, and risk arbitrage. In such cases, the price impact of trading a portfolio might 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, the portfolio execution cost might be lower than the sum of the individual stocks' execution costs. This is because of a diversification effect in which 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, the portfolio setting clearly is considerably more complex than the single-stock case.

The state equations

We now present a specification for the state equations that incorporates a multivariate price-impact function with cross-stock interactions.

Let the execution price $\boldsymbol{p}_t$ be the sum of two components:

$$\boldsymbol{p}_t = \tilde{\boldsymbol{p}}_t + \boldsymbol{\delta}_t, \tag{3}$$

where $\tilde{\boldsymbol{p}}_t$ is a "no-impact" price—the price that would prevail in the absence of any market impact—and $\boldsymbol{\delta}_t$ is the impact. A plausible and observable proxy for the no-impact price is the midpoint of the bid/offer spread, although it can be arbitrary so long as the trade size $\boldsymbol{s}_t$ does not affect it. For convenience, and to ensure nonnegative prices, we model $\tilde{\boldsymbol{p}}_t$ as vector-geometric Brownian motion:

$$\tilde{\boldsymbol{p}}_t = \exp(\boldsymbol{Z}_t)\tilde{\boldsymbol{p}}_{t-1}, \tag{4}$$

where $\boldsymbol{Z}_t$ is a diagonal matrix whose diagonal is a normal random vector $\mathbf{z}_t$ with mean $\boldsymbol{\mu}_z$ and covariance matrix Σ_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 $\boldsymbol{\delta}_t$ we set

$$\boldsymbol{\delta}_t = \tilde{\boldsymbol{P}}_t(\boldsymbol{A}\tilde{\boldsymbol{P}}_t\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{x}_t), \tag{5}$$

where $\tilde{\boldsymbol{P}}_t = \text{diag}[\tilde{\boldsymbol{p}}_t]$ and $\text{diag}(\cdot)$ is the diagonalization operator that maps its vector argument into a diagonal matrix with the vector as the diagonal. This specification captures the impact of trading $\boldsymbol{s}_t$ shares on the transaction prices $\boldsymbol{p}_t$. It also 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 price impact's form (see Equation 5) 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.[6]

This specification of the dynamics of $\boldsymbol{p}_t$ has several advantages over other specifications (see the "Other specifications" sidebar). First, $\tilde{\boldsymbol{p}}_t$ is guaranteed to be nonnegative, and $\boldsymbol{p}_t$ is guaranteed to be nonnegative under mild restrictions on $\boldsymbol{\delta}_t$. Second, separating the transaction price $\boldsymbol{p}_t$ into the no-impact component $\tilde{\boldsymbol{p}}_t$ and impact component $\boldsymbol{\delta}_t$ makes the trade's price impact temporary. So, the impact affects the current transaction price but does not affect future prices. Third, the percentage price impact increases linearly with the trade size, which is empirically plausible.[1,9–11] Fourth, Equation 3 implies a natural decomposition of execution costs, decoupling market-microstructure effects from price dynamics, which is closely related to André Perold's notion of implementation shortfall.[12] Finally, we shall see in "The dynamic-programming solution" section that Equation 3 admits a closed-form solution in which the best-execution strategy is a simple linear function of the state variables and in which the optimal-value function is quadratic.

(Because Equation 3 implies that price impact is temporary, affecting only $\boldsymbol{p}_t$ and not $\tilde{\boldsymbol{p}}_t$, the objective function [see Equation 1] separates into two terms. The first is the no-impact cost of execution and the second is the total impact cost. This decomposition is precisely the one Perold proposed in his definition of implementation shortfall, but we apply it to executing $\bar{\boldsymbol{s}}$. In particular, the first term gives the "paper" return or execution cost, and the sum of the two terms gives the actual cost. So the second term is the implementation shortfall in executing $\bar{\boldsymbol{s}}$.)

The presence of the vector $\boldsymbol{x}_t$ in Equation 5 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 $\boldsymbol{x}_t$. For simplicity, we let

$$\boldsymbol{x}_t = \boldsymbol{C}\boldsymbol{x}_{t-1} + \eta_t, \tag{6}$$

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

The dynamic-programming solution

We use a stochastic dynamic-programming algorithm to solve the optimal-execution problem (see Equation 1). We denote by w_t the vector of shares remaining to be bought (or sold) at time t:

$$w_t = w_{t-1} - s_{t-1}$$

$$w_1 = \bar{s}$$

$$w_{T+1} = \mathbf{0}$$

The condition $w_{T+1} = \mathbf{0}$ ensures that all $\bar{s}$ shares are executed by time T. The complete statement of the problem is then

$$\underset{\{s_t\}}{\text{Min}}\, \mathrm{E}_1\left[\sum_{t=1}^{T} p_t' s_t\right],$$

subject to

$$p_t = f(p_{t-1}, x_t, s_t, \epsilon_t)$$

$$x_t = g(x_{t-1}, \eta_t)$$

$$w_t = w_{t-1} - s_{t-1}$$

$$w_1 = \bar{s}$$

$$w_{T+1} = \mathbf{0}$$

Linear-percentage price impact

As with all dynamic-programming solutions, we begin at the end. V_T is the optimal value function at the end of our trading horizon, period T. By definition,

$$V_T(\tilde{p}_T, x_T, w_T) = \underset{\{s_T\}}{\text{Min}}\, \mathrm{E}_T[p_T' s_T] = \mathrm{E}_T[p_T' w_T]$$

$$= \left[\tilde{P}_T(e_n + A\tilde{P}_T w_T + Bx_T)\right]' w_T \tag{7}$$

Because this is the last period and w_{T+1} must be set to zero, the remaining order w_T must execute. So, the optimal trade size $s_T^* = w_T$. Because $\tilde{P}_T = \tilde{P}'_T$, we can reexpress Equation 7 as

Other specifications

The specification for the state equation proposed in "The state equations" in the main article is only one of many possible specifications. For example, Dimitris Bertsimas and Andrew Lo[1] propose a natural multivariate extension of their "linear price impact" specification in which the state equation is

$$p_t = p_{t-1} + As_t + Bx_t + \epsilon_t, \tag{A}$$

where A is a positive definite $n \times n$ matrix, B is an arbitrary $n \times m$ matrix, x_t is an m vector of information variables, and ϵ_t is n-vector white noise with mean $\mathbf{0}$ and covariance matrix Σ_ϵ. As before, we assume that x_t follows a stationary AR(1) process. So,

$$x_t = Cx_{t-1} + \eta_t, \tag{B}$$

where C is an $m \times m$ matrix with eigenvalues all less than unity in modulus, and η_t is m-vector white noise with mean $\mathbf{0}$ and covariance matrix Σ_η, and which is independent of ϵ_t.

This specification differs significantly from the linear-percentage price impact in three basic respects. Perhaps the most important difference is that Equation A implies that the *price impact has a "permanent" effect on prices*, because of the random-walk nature of the price dynamics. Also, the price impact is linear, implying that a 10,000-share trade will have the same dollar impact on a \$1 stock as it will on a \$100 stock. Finally, unless some rather unnatural restrictions are placed on ϵ_t in Equation A, p_t could take on negative values—clearly an unrealistic prospect.

Of course, Bertsimas and Lo consider Equation A primarily for analytic tractability, not because of supporting empirical evidence.[1] Nevertheless, it is instructive to compare this specification with the linear-percentage specification 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.[2-8] However, given the ever-changing nature of financial markets, it is crucial to reestimate the state equation for each application using the most recent data sets available.

References

1. D. Bertsimas and A. Lo, "Optimal Control of Execution Costs: A Study of Government Bonds with the Same Maturity Date," *J. Financial Markets*, Vol. 1, No. 1, 30 Apr. 1998, pp. 1–50.
2. M. Barclay and R. Litzenberger, "Announcement Effects of New Equity Issues and the Use of Intraday Price Data," *J. Financial Economics*, Vol. 21, 1988, pp. 71–100.
3. M. Barclay and J. Warner, "Stealth Trading and Volatility: Which Trades Move Prices?" *J. Financial Economics*, Vol. 34, 1993, pp. 281–306.
4. L. Chan and J. Lakonishok, "Institutional Trades and Intra-Day Stock Price Behavior," *J. Financial Economics*, Vol. 33, 1993, pp. 173–199.
5. L. Chan and J. Lakonishok, "The Behavior of Stock Prices around Institutional Trades," *J. Finance*, Vol. 50, 1995, pp. 1147–1174.
6. R. Holthausen, R. Leftwich, and D. Mayers, "The Effect of Large Block Transactions on Security Prices: A Cross-Sectional Analysis," *J. Financial Economics*, Vol. 19, 1987, pp. 237–267.
7. R. Holthausen, R. Leftwich, and D. Mayers, "Large Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects," *J. Financial Economics*, Vol. 26, 1990, pp. 71–95.
8. A. Kraus and H. Stoll, "Price Impacts of Block Trading on the New York Stock Exchange," *J. Finance*, Vol. 27, 1972, pp. 569–588.

$$V_T(\tilde{\boldsymbol{p}}_T, \boldsymbol{x}_T, \boldsymbol{w}_T) = \boldsymbol{e}_n' \tilde{\boldsymbol{P}}_T \boldsymbol{w}_T + \boldsymbol{w}_T' \tilde{\boldsymbol{P}}_T \boldsymbol{A}' \tilde{\boldsymbol{P}}_T \boldsymbol{w}_T + \boldsymbol{x}_T' \boldsymbol{B}' \tilde{\boldsymbol{P}}_T \boldsymbol{w}_T,$$

which shows that the optimal value function is linear in $\boldsymbol{x}_T$ and linear-quadratic in $\boldsymbol{w}_T$. By continuing recursively in this fashion and applying Bellman's principle of optimality,[13] we find that the optimal value function V_{t-k} is

$$\begin{aligned} V_{T-k} &= \underset{\{s_{T-k}\}}{\text{Min}} \text{E}_{T-k}[\boldsymbol{p}_{T-k}' \boldsymbol{s}_{T-k} + V_{T-k+1} \\ &\qquad (\tilde{\boldsymbol{p}}_{T-k+1}, \boldsymbol{x}_{T-k+1}, \boldsymbol{w}_{T-k+1})] \\ &= \boldsymbol{e}_n' \boldsymbol{D}_{n,k} \boldsymbol{e}_n + \boldsymbol{e}_r' \boldsymbol{D}_{r,k} \boldsymbol{e}_r + \boldsymbol{e}_n' \boldsymbol{E}_k \boldsymbol{x}_{T-k} \\ &\quad + \boldsymbol{x}_{T-k}' \boldsymbol{F}_k \boldsymbol{e}_n + \boldsymbol{x}_{T-k}' \boldsymbol{G}_k \boldsymbol{x}_{T-k} + \boldsymbol{x}_{T-k}' \boldsymbol{H}_k \boldsymbol{w}_{T-k} \\ &\quad + \boldsymbol{w}_{T-k}' \boldsymbol{J}_{T-1} \boldsymbol{x}_{T-k} + \boldsymbol{e}_n' \boldsymbol{K}_k \boldsymbol{w}_{T-k}' + \tilde{\boldsymbol{w}}_{T-k}' \boldsymbol{L}_k \boldsymbol{e}_n \\ &\quad + \boldsymbol{w}_{T-k}' \boldsymbol{N}_k \boldsymbol{w}_{T-k} \end{aligned} \tag{8}$$

This yields the best-execution strategy

$$\tilde{\boldsymbol{s}}_{T-k}^* = \Delta_{x,k} \boldsymbol{x}_{T-k} + \Delta_{w,k} \boldsymbol{w}_{T-k} + \Delta_{1,k} \boldsymbol{e}_n \tag{9}$$

(For explicit expressions for $\boldsymbol{D}_{n,k}$, $\boldsymbol{D}_{r,k}$, $\boldsymbol{E}_k$, $\boldsymbol{F}_k$, $\boldsymbol{G}_k$, $\boldsymbol{H}_k$, $\boldsymbol{J}_k$, $\boldsymbol{K}_k$, $\boldsymbol{L}_k$, and $\boldsymbol{N}_k$, and for $\Delta_{x,k}$, $\Delta_{w,k}$, and $\Delta_{1,k}$, see the Appendix posted at http://computer.org/cise.) The recursion (see Equation 8) and best-execution strategy (see Equation 9) completely characterize the solution to our original problem, and yield the expected best-execution cost, V_{T-k}, as a by-product.

Linear price impact

Under the law of motion (see Equations A and B in the sidebar), Bertsimas and Lo[6] show that the portfolio problem (see Equation 1) can be solved with Bellman's equation, which yields the following best-execution strategy,

$$\boldsymbol{s}_{T-k}^* = (\boldsymbol{I} - \frac{1}{2} \boldsymbol{A}_{k-1}^{-1} \boldsymbol{A}') \boldsymbol{w}_{T-k} + \frac{1}{2} \boldsymbol{A}_{k-1}^{-1} \boldsymbol{B}_{k-1}' \boldsymbol{C} \boldsymbol{x}_{T-k}, \tag{10}$$

and optimal-value function,

$$\begin{aligned} V_{T-k}(\boldsymbol{p}_{T-k-1}, \boldsymbol{x}_{T-k}, \boldsymbol{w}_{T-k}) &= \boldsymbol{p}_{T-k-1}' \boldsymbol{w}_{T-k} + \boldsymbol{w}_{T-k}' \boldsymbol{A}_k \boldsymbol{w}_{T-k} \\ &\quad + \boldsymbol{x}_{T-k}' \boldsymbol{B}_k \boldsymbol{w}_{T-k} + \boldsymbol{x}_{T-k}' \boldsymbol{C}_k \boldsymbol{x}_{T-k} + d_k \end{aligned}$$

for $k = 0, \ldots, T-1$, where

$$\boldsymbol{A}_k = \boldsymbol{A} - \frac{1}{4} \boldsymbol{A} \boldsymbol{A}_{k-1}^{-1} \boldsymbol{A}', \qquad \boldsymbol{A}_0 = \boldsymbol{A}$$

$$\boldsymbol{B}_k = \frac{1}{2} \boldsymbol{C}' \boldsymbol{B}_{k-1} (\boldsymbol{A}_{k-1}')^{-1} \boldsymbol{A}' + \boldsymbol{B}', \qquad \boldsymbol{B}_0 = \boldsymbol{B}'$$

$$\boldsymbol{C}_k = \boldsymbol{C}' \boldsymbol{C}_{k-1} \boldsymbol{C} - \frac{1}{4} \boldsymbol{C}' \boldsymbol{B}_{k-1} (\boldsymbol{A}_{k-1}')^{-1} \boldsymbol{B}_{k-1} \boldsymbol{C}, \ \boldsymbol{C}_0 = \boldsymbol{0}$$

$$d_k = d_{k-1} + \text{E}[\eta_{T-k}' \boldsymbol{C}_{k-1} \eta_{T-k}], \qquad d_0 = 0$$

The best-execution strategy (see Equation 10) is qualitatively similar to the optimal single-stock strategy of Bertsimas and Lo.[6] However, it has one key difference: in the portfolio case, unless the matrix $\boldsymbol{A}$ is diagonal, the best-execution strategy for one stock will depend on the parameters and state variables of all the other stocks. To see this, observe that the matrix coefficient

$$\left(\boldsymbol{I} - \frac{1}{2} \boldsymbol{A}_{k-1}^{-1} \boldsymbol{A}' \right)$$

multiplying $\boldsymbol{w}_{T-k}$ in Equation 10 will generally not be a diagonal matrix unless $\boldsymbol{A}$ is diagonal. Of course, if $\boldsymbol{A}$ is diagonal, trading in one stock has no price impact on any other stocks (see Equation A in the sidebar). So, the portfolio problem essentially reduces to n independent single-stock problems.

So, whether the portfolio best-execution cost is greater or less than the sum of the individual stocks' best-execution costs depends wholly on the values in $\boldsymbol{A}$. This is an empirical issue that we consider in detail later.

Imposing constraints

Most practical applications will have constraints on the kind of execution strategies that institutional investors can follow. For example, if a block of shares is to be purchased within T periods, selling the stock during these T periods is very difficult to justify even if such sales are warranted by the best-execution strategy. (Other common constraints include sector-balance, turnover, tax-motivated, and, in the portfolio case, dollar-balance constraints. This last type of constraint—the portfolio's dollar value at the end of trading must lie within some fixed interval—is one of the most difficult to impose. This is because the constraint is a function of the entire vector of prices, which is stochastic. Bertsimas and Lo devised a probabilistic method of imposing such constraints.[6]) So, in practice, buy programs will almost always have nonnegativity constraints, and sell programs will almost always have nonpositivity constraints. Such constraints are often binding for best-execution strategies, particularly when the information variable has a large effect on price impact.

Why constraints are problematic

Although there are well-known techniques for solving constrained-optimization problems in a static setting, corresponding techniques for dynamic-optimization problems have not yet been developed. To see why this task is difficult, consider the simplest case of imposing nonnegativity constraints $s_t \geq \mathbf{0}$ in the linear percentage price-impact model with only one asset (scalar equations). Without any constraints, the optimal-value function V_{T-k} is quadratic in the state variable w_{T-k}, so the Bellman equation can be easily solved in closed form. But if nonnegativity constraints are imposed, V_{T-k} becomes a piecewise-quadratic function, with 3^k pieces.

To see how this arises, observe that for $k = 0$, the optimal control is $s_T^* = w_T$ and V_T is a quadratic function of w_T. In the next stage, $k = 1$, we calculate the optimal control s_{T-1}^* by minimizing a quadratic function of s_{T-1} subject to the constraints $0 \leq s_{T-1} \leq w_{T-1}$. The solution is

$$s_{T-1}^* = \begin{cases} 0 & \text{if } s_{u,T-1} < 0 \\ s_{u,T-1} & \text{if } 0 < s_{u,T-1} < w_{T-1} \\ w_{T-1} & \text{if } s_{u,T-1} > w_{T-1} \end{cases} ,$$

where

$$s_{u,T-1} = \frac{1}{\tilde{p}_{T-1}} \left[\Delta_{w,1} \tilde{p}_{T-1} w_{T-1} + \Delta_{x,1} x_{T-1} + \Delta. \right.$$

This partitions the range of w_{T-1} into three intervals; a different optimal control s_{T-1}^* is over each interval, and a continuous quadratic function of w_{T-1} is within each interval V_{T-1}.

The next stage, $k = 2$, partitions each of these three intervals into another three intervals, each with a different optimal control s_{T-2}^*, and so on. The number of intervals grows exponentially with k. Therefore, even in this simple case, calculating s_{T-k}^* and V_{T-k} exactly is only feasible for a very small number of periods T. (For example, when $T = 20$ there are $3^{20} = 3{,}486{,}784{,}401$ intervals at the last stage of the dynamic program!)

A static-approximation method

Faced with these difficulties, we propose an approximation method to address the optimal control problem with constraints. The dynamic-optimization algorithm we presented for the case without nonnegativity constraints gives the best-execution strategy s_{T-k}^* (see Equation 9) as a function of the state vector $(x_{T-k}, w_{T-k}, \tilde{p}_{T-k})$ at time $T - k$. At time $t = 1$, the expected execution cost V_1 is

$$V_1 = \mathrm{E}\left[\sum_{k=1}^{T} p_k' s_k \right]$$

$$= \mathrm{E}\left[\sum_{k=1}^{T} [\tilde{P}_k e_n + \tilde{P}_k (A \tilde{P}_k s_k + B x_k)]' s_k \right]$$

$$\begin{aligned} = \mathrm{E}\Big[& e_n' \tilde{P}_1 s_1 + s_1' \tilde{P}_1 A' \tilde{P}_1 s_1 + x_1' B' \tilde{P}_1 s_1 + e_n' \tilde{Z}_2 \tilde{P}_1 s_2 \\ & + s_2' \tilde{P}_1 \tilde{Z}_2 A' \tilde{Z}_2 \tilde{P}_1 s_2 + (C x_1 + \eta_2)' B' \tilde{Z}_2 \tilde{P}_1 s_2 \\ & + e_n' \tilde{Z}_2 \tilde{Z}_3 \tilde{P}_1 s_3 + s_3' \tilde{P}_1 \tilde{Z}_3 \tilde{Z}_2 A' \tilde{Z}_2 \tilde{Z}_3 \tilde{P}_1 s_3 \\ & + \left[C(C x_1 + \eta_2) + \eta_3 \right]' B' \tilde{Z}_2 \tilde{Z}_3 \tilde{P}_1 s_3 \\ & + \mathrm{L} \\ & + e_n' \tilde{Z}_2 \tilde{Z}_3 \mathrm{L}\ \tilde{Z}_T \tilde{P}_1 s_T \\ & + s_T' \tilde{P}_1 \tilde{Z}_T \tilde{Z}_{T-1} \mathrm{L}\ \tilde{Z}_2 A' \tilde{Z}_2 \tilde{Z}_3 \mathrm{L}\ \tilde{Z}_T \tilde{P}_1 s_T \\ & + \left[C^T x_1 + C^{T-1} \eta_2 + \mathrm{L} + C \eta_T + \eta_T \right]' \\ & \times B' \tilde{Z}_2 \tilde{Z}_3 \mathrm{L}\ \tilde{Z}_T \tilde{P}_1 s_T \Big] \end{aligned}$$

where $\tilde{P}_k = \mathrm{diag}[p_k]$ and $\tilde{Z}_k = \exp(Z_k)$. Taking the expectation of the cost function yields

$$\begin{aligned} V_1 = {} & e_n' \tilde{P}_1 s_1 + s_1' A' \tilde{P}_1 s_1 + x_1' B' \tilde{P}_1 s_1 \\ & + e_n' Q \tilde{P}_1 s_2 + s_2' (A' \bullet R) \tilde{P}_1 s_2 + x_1' C' B' Q \tilde{P}_1 s_2 \\ & + e_n' Q^2 \tilde{P}_1 s_3 + s_3' (A' \bullet R \bullet R) \tilde{P}_1 s_3 + x_1' (C^2)' B' Q^2 \tilde{P}_1 s_3 \\ & + \mathrm{L} \\ & + e_n' Q^T \tilde{P}_1 s_T + s_T' (A' \bullet R \bullet R \bullet \mathrm{L} \bullet R \bullet R) \tilde{P}_1 s_T \\ & + x_1' (C^T)' B' Q^T \tilde{P}_1 s_T \end{aligned} \qquad (11)$$

where Q is a $(n \times n)$ diagonal matrix with entries

$$q_i = \exp\left(\mu_{z,i} + \frac{1}{2} \Sigma_{z,ii} \right),$$

R is an $(n \times n)$ symmetric matrix with elements

$$r_{ij} = \exp\left[\mu_{z,i} + \mu_{z,j} + \frac{1}{2} \left(\Sigma_{z,ii} + \Sigma_{z,jj} + 2\Sigma_{z,ij} \right) \right],$$

and the matrix dot operator "•" denotes an element-wise matrix multiplication—that is, $A \bullet B \equiv [a_{ij} b_{ij}]$.

Equation 11 depends on the entire sequence of controls, $\{s_1, s_2, \ldots, s_T\}$, and the observed states at time $t = 1$, $\tilde{p}_1$ and x_1. In general, each control variable s_t depends on the state at time t. Under the *static-approximation method*, we will restrict the class of controls to those that s_t depend *only* on the state at time $t = 1$. That is, they depend only on prices $\tilde{p}_1$ and information vector x_1.

Table 1. Ticker symbols, CUSIPs, company names, and closing prices on a randomly selected day in 1996 for 25 stocks that constitute the sample portfolio for the empirical implementation of the best-execution strategy.

Ticker	CUSIP	Company name	Closing price
AHP	02660910	AMER HOME PRODS	64.0625
AN	03190510	AMOCO	70.5000
BLS	07986010	BELLSOUTH	37.2500
CHV	16675110	CHEVRON	62.6250
DD	26353410	DUPONT	88.9375
DIS	25468710	WALT DISNEY	63.4375
DOW	26054310	DOW CHEMICAL	80.6875
F	34537010	FORD MOTOR	31.3125
FNM	31358610	FANNIE MAE	35.0625
GE	36960410	GENERAL ELECTRIC	90.9375
GM	37044210	GM	48.1250
HWP	42823610	HEWLETT PACKARD	48.8125
IBM	45920010	IBM	25.8750
JNJ	47816010	JOHNSON & JOHNSON	51.4375
KO	19121610	COCA COLA	50.8125
MCD	58013510	MCDONALDS	47.7500
MO	71815410	PHILIP MORRIS	90.1875
MOB	60705910	MOBIL	15.9375
MRK	58933110	MERCK & CO	70.1250
PEP	71344810	PEPSICO	28.3750
PG	74271810	PROCTER & GAMBLE	97.4375
S	81238710	SEARS ROEBUCK	44.8750
T	00195710	AT&T	51.9375
WMT	93114210	WAL MART STORES	26.3125
XON	30229010	EXXON	83.5625

Under this approximation, the problem reduces to this quadratic-optimization problem:

$$
\begin{aligned}
\text{Minimize}\quad & e_n'\tilde{P}_1 s_1 + s_1' A'\tilde{P}_1 s_1 + x_1' B'\tilde{P}_1 s_1 + e_n' Q\tilde{P}_1 s_2 \\
& + s_2'(A' \bullet R)\tilde{P}_1 s_2 + x_1' C' B' Q\tilde{P}_1 s_2 + e_n' Q^2 \tilde{P}_1 s_3 \\
& + s_3'(A' \bullet R \bullet R)\tilde{P}_1 s_3 + x_1'(C^2)' B' Q^2 \tilde{P}_1 s_3 \\
& + \cdots \\
& + e_n' Q^T \tilde{P}_1 s_T + s_T'(A' \bullet R \bullet R \bullet \cdots \bullet R \bullet R)\tilde{P}_1 s_T \\
& + x_1'(C^T)' B' Q^T \tilde{P}_1 s_T \\
\text{subject to}\quad & \bar{s} = \sum_{t=1}^{T} s_t \\
& 0 \le s_t,\ t = 1, \ldots, T
\end{aligned}
\tag{12}
$$

We solve Equation 12 at time $t = 1$ and find the "optimal" controls $s_1^1, \ldots, s_T^1$, where the superscript indicates that this is the period-1 solution of Equation 12. However, we only implement the control s_1^1. After we observe the state vector in period $t = 2$, we re-solve Equation 12 for time $t = 2$ and find a new set of controls $s_2^2, \ldots, s_T^2$, but only implement the control s_2^2. We continue in this fashion, at each step solving a convex quadratic optimization problem that can be handled efficiently using commercially available packages—for example, C-Plex or Minos.

The static-approximation method might not yield adequate approximations in all cases. However, in many of the examples we explored, the technique performs admirably (for example, the empirical analysis in the next section). Of course, deriving accurate bounds on the approximation error in the most interesting cases is difficult because the optimal solutions are unknown for these cases. We hope to explore the theoretical properties of the static-approximation method in future research.

An empirical example

We now implement the best-execution strategies for a hypothetical stock portfolio. Specifically, we estimate the parameters of the linear-percentage model in "The state equations" section for each stock. We then construct several portfolio-rebalancing scenarios and compare the best-execution strategy with a "naive" strategy of trading equal-size lots in each time period.

The data

Our empirical analysis draws on three data sources. The primary source is a proprietary record of trades performed over the NYSE DOT system by the trading desk at Investment Technologies Group (ITG) on every trading day between 2 January and 31 December 1996. Each trade is cataloged with this information: order-submission date and time, order execution date and time, whether it is a buy or sell order, size in shares, execution price, and order type (for example, market order or limit order). We chose the 25 stocks that had the greatest number of market orders over the year-long interval (see Table 1).

Because of our selection rule, our sample consists of companies with large market capitalizations. This ensures that we will have enough data to fit the model and arrive at reasonably accurate estimates of the parameters. But such a sample tends to exhibit a lower-than-average price impact because stocks that trade very frequently are, by definition, very liquid and have much smaller price impact. Such a bias in our sampling procedure by no means invalidates our example's relevance. If we can demonstrate that our best-execution strategy is beneficial for highly liquid stocks, our approach's value is likely to be even greater for less liquid stocks, where price impact is significantly higher.

The ITG database provides valuable trade information, but we must augment our analysis with NYSE TAQ data to extract quotes prevailing at the time of ITG trades. The TAQ database is a complete history of all trades and quotes on the NYSE, AMEX, and Nasdaq exchanges.

Finally, we use S&P 500 tick data provided by Tick Data Inc. to get intraday levels for the S&P 500 index during 1996.

The estimation procedure

Our estimation procedure consists of three steps. First, we estimate the parameters μ_z and Σ_z of the no-impact price dynamics (see Equation 4) for each stock. Given the geometric-Brownian-motion specification, we know that the continuously compounded returns z_{it} are independently and identically (IID) normal random variates:

$$z_{i,t} = \log\left(\frac{\tilde{p}_{i,t}}{\tilde{p}_{i,t-1}}\right) \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad (13)$$

for each stock i, where $i = 1, \ldots, 25$ and $\mathcal{N}(\mu_i, \sigma_i^2)$ is the normal distribution with mean μ_i and variance σ_i^2. The no-impact price is taken to be the midpoint of the prevailing bid and offer prices at time t (hence the need for quotes):

$$\tilde{p}_{it} = \frac{\tilde{p}^b_{it} + \tilde{p}^a_{it}}{2} \quad (14)$$

where $\tilde{p}^b_{it}$ and $\tilde{p}^a_{it}$ are the bid and ask prices for stock i at time t. For each of our 25 stocks, we collect TAQ quotes at every half hour over the course of the 1996 trading year and calculate the midpoint to construct the no-impact prices, $\tilde{p}_t$. Thus, the time index, t, ranges over half hours, $t = 1, 2, \ldots, N_h$, where N_h is the total number of half hours in the 1996 trading year (approximately 250 days times 13 periods per day).

We then form log returns according to Equation 13 and discard any overnight returns to eliminate nonsynchronous trading effects. This gives us a sample of 2,069 observations of z_t during the 1996 calendar year from which we can estimate μ_z and Σ_z in the standard way. Table 2 summarizes the results (to conserve space, we report estimates only for the first five stocks of Table 1). The drift and volatility are expressed in percent per year; we scale them up from the half-hourly units by assuming each of the 250 trading days per year consists of 13 half-hour trading intervals. The drift and volatility estimates are consistent with intuition and agree reasonably well with other data sources such as BARRA.

Our second task is to estimate the parameters of the market-information process in Equation 6. The variable x_t captures the potential impact of changing market conditions or private information about the security. For example, we could construct a short-term excess-returns model for this purpose. In our example, x_t denotes the half-hourly returns on the S&P 500 index, a common factor that influences the prices of most securities. For this specification, the AR(1) coefficients, C, and the covariance matrix of the noise, Σ_η, reduce to scalar quantities, C and σ_η. Using the S&P 500 tick data from 2 January to 31 December 1996, we construct the returns x_t, where t is the same time index used previously. We rescale the returns by subtracting out the mean and dividing by standard deviation. This leaves us with a zero-mean, unit-standard-deviation process:

$$\tilde{x}_t = \frac{x_t - \mu_x}{\sigma_x}. \quad (15)$$

Assuming $|C| < 1$, we can rewrite Equation 15 as

$$\tilde{x}_t = C\tilde{x}_{t-1} + \eta_t\,.$$

The maximum likelihood estimator of the AR(1) coefficient C is

Table 2. Parameter estimates and correlations for the no-impact price process $\tilde{p}_t$ for five stocks, using 2,069 half-hourly observations from 2 January to 31 December 1996. The first and second rows give the annual drift and volatility parameters (percent/year) scaled up from half-hourly estimates by assuming 250 trading days with 13 half-hour periods per day. The last five rows report the correlation coefficients for the half-hourly returns of the five stocks.

	AHP	AN	BLS	CHV	DD
$\hat{\mu}$	0.200	−0.103	−0.387	−0.167	0.283
$\hat{\sigma}$	0.268	0.196	0.283	0.222	0.225
AHP	1.000	0.196	0.284	0.226	0.345
AN	0.196	1.000	0.173	0.408	0.283
BLS	0.284	0.173	1.000	0.259	0.328
CHV	0.226	0.408	0.259	1.000	0.314
DD	0.345	0.283	0.328	0.314	1.000

$$\hat{C} = \frac{\frac{1}{T_1}\sum_{t=2}^{T_1} \tilde{x}_t \tilde{x}_{t-1}}{\frac{1}{T_2}\sum_{t=1}^{T_2} \tilde{x}_t^2}. \tag{16}$$

To avoid nonsynchronous trading effects, we discard all products $\tilde{x}_t \tilde{x}_{t-1}$ that straddle an overnight period in Equation 16's numerator. Similarly, we exclude overnight-return terms from Equation 16's denominator. The constants, T_1 and T_2, are the number of terms that are included in calculating the numerator and denominator and are 1,902 and 2,078. Our estimate of $\hat{C}$ is 0.0354. (Not surprisingly, the level of serial correlation in the S&P 500 index is quite low. If not, profitable trading strategies would be possible that would quickly drive the predictability of index returns back to a low level.)

Given $\hat{C}$, the maximum-likelihood estimator for the standard deviation of η_t is

$$\sigma_\eta = \sqrt{1 - \hat{C}^2}.$$

Our estimate is 0.999. The parameters $\hat{C}$ and $\hat{\sigma}_\eta$ fully characterize the AR(1) process that describes the S&P 500 returns.

Our final task is to estimate the parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ of the price-impact equation (see Equation 5). We can recast the vector equation as 25 separate linear regressions by rearranging terms:

$$\frac{p_{it} - \tilde{p}_{it}}{\tilde{p}_{it}} = \tilde{p}_{it} s_t \boldsymbol{a}_i' + \boldsymbol{x}_i \boldsymbol{b}_i',$$

where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are the ith rows of $\boldsymbol{A}$ and $\boldsymbol{B}$. This expression shows that the percentage price impact to the ith security is a linear function of the dollar volume we intend to trade in the ith security, the dollar volumes that we and others are currently trading in the other 24 stocks, and the S&P 500 return over the preceding half hour.

Obviously, trading in stock i should have a price impact on p_{it}. But less obvious is the role that trading in other stocks might play in determining the price impact on p_{it}. Such cross-effects have several economic sources. One stock might be a close substitute for another, so a high price impact for one would imply the same for the other. Another motivation is that, in a market with sharply rising (or falling) prices and high volume, the overall market impact will likely increase as liquidity providers demand higher premiums above posted quotes for large market orders.

To estimate $\boldsymbol{A}$ and $\boldsymbol{B}$, we use a combination of ITG proprietary data, TAQ data, and SPX tick data. For each executed market order in a given stock i, the ITG database gives complete information about the market order except for the prevailing quote. We search the TAQ database to find the quote. As before, we form the no-impact price, $\tilde{p}_{it}$, as the average of the bid and offer (see Equation 14) and then construct the dependent variables,

$$\frac{p_{it} - \tilde{p}_{it}}{\tilde{p}_{it}},$$

for each trade. The ITG database provides one independent variable—namely, the dollar volume of stock i: $\tilde{p}_{it} s_{it}$.

A difficulty arises in constructing other dollar-volume-related independent variables (that is, $\tilde{p}_{jt} s_{jt}$ for $i \neq j$). The ITG data is too sparse to find nearby trades in time, so we must turn to the TAQ data to resolve this observability problem. One possible solution is to use the nearest TAQ trade that occurs before an ITG market order. Unfortunately, the time alignment of the TAQ and ITG data sets can be imprecise because of recording lags by either party. To reduce the impact of this type of error, we define a proxy for the closest trade by forming a 30-second window before each market order and computing an average dollar volume within it. Although this averaging procedure tends to smooth the data and reduces its information content, it ensures that temporal sequencing is not violated.

Specifically, we find all N_k trades in stock j that occur within that window. Each trade is executed at price p_{ik}, where $k = 1, \ldots, N_k$. Trades that are executed above or at the midpoint of their quotes are classified as buys, and the rest as sells. We then compute an average dollar volume within the window for stock j as

$$\tilde{p}_{jt}s_{jt} = \frac{1}{N_k}\sum_{k=1}^{N_k}\tilde{p}_{jk}s_{jk} .$$

Finally, for the S&P 500 return, x_t, we split the trading day into 13 half-hour intervals and compute the return in the half hour before the trade.

We now have a complete set of data with which to estimate the parameters of our model's price-impact portion. We performed the regressions in SAS; they contained no intercept term because the price impact should be zero if no stocks are being traded. Table 3 summarizes the first five of the 25 regressions. For each regression, the table reports the parameter estimates for the 26 regressors—the lagged returns for the 25 stocks and the S&P 500 lagged return—and their t-statistics. R^2 and the sample size appear at the bottom of each column. (We performed diagnostics on the residuals to test for the presence of heteroskedasticity and autocorrelation. The Durbin-Watson test indicated low levels of positive serial correlation, with statistics ranging from 1.12 to 1.69 for the 25 regressions. The test of first and second moment specification indicated a very weak presence of heteroskedasticity, because the p-values were, in general, very low.)

To develop some intuition for the coefficients, consider the estimated price impact for American Home Products in $\hat{A}$ caused by trading in AHP, which is 4.97×10^{-10}, according to Table 3. If we traded a 100,000-share block of AHP at its beginning-of-year price of \$64.0625 with no impact, our total cost would be 100,000 × \$64.0625 = \$6,406,250. But according to Table 3, the full-impact cost would be

$$100{,}000 \times (\tilde{p}_t + \delta_t) = 100{,}000 \times (\$64.0625 + \delta_t)$$

$$\delta_t = \tilde{p}_t\left(4.97 \times 10^{-10} \times \tilde{p}_t \times 100{,}000\right) = 0.203969$$

$$100{,}000 \times (\tilde{p}_t + \delta_t) = \$6{,}426{,}647$$

which implies a price impact of approximately 20 cents per share (ignoring the other factors in the regression). This estimated price impact is unacceptably high; no professional trader would submit such a large order except in the most desperate circumstances.

Further inspection of the regression diagnostics shows that R^2 ranges from 0.052 to 0.440 for the 25 regressions, indicating that the regressions have varying degrees of explanatory power. How-

Table 3. Coefficients of the unconstrained price-impact regressions for five stocks, based on market orders from 2 January to 31 December 1996. All coefficients have been multiplied by 10^{10} except the SPX coefficients, which have been multiplied by 10^5. The last two rows contain the sample size T and R^2 coefficients; the t-statistics appear in parentheses below the coefficients.

Variable	AHP	AN	BLS	CHV	DD
SPX	3.74	2.29	0.27	11.73	2.38
	(0.92)	(0.43)	(0.04)	(1.94)	(0.56)
AHP	4.97	–2.04	–1.89	0.80	–1.78
	(3.36)	(–2.11)	(–1.12)	(0.69)	(–1.65)
AN	–1.49	5.86	–2.27	1.38	–1.89
	(–1.62)	(5.96)	(–1.49)	(1.14)	(–1.81)
BLS	4.46	0.31	0.85	4.37	–1.84
	(2.05)	(0.18)	(0.29)	(2.20)	(–1.07)
CHV	2.02	–0.25	4.90	8.72	3.24
	(1.38)	(–0.32)	(2.54)	(2.84)	(3.30)
DD	–1.08	–0.70	5.31	1.29	7.21
	(–0.81)	(–0.62)	(3.17)	(0.89)	(5.53)
DIS	2.35	3.04	0.01	–1.54	1.17
	(2.26)	(3.23)	(0.03)	(–1.28)	(1.35)
DOW	–1.12	0.12	–3.53	–2.43	–2.52
	(–1.12)	(0.27)	(–2.04)	(–1.89)	(–3.44)
F	1.75	0.19	0.01	–1.01	2.19
	(0.70)	(0.15)	(0.02)	(–0.47)	(1.16)
FNM	–5.10	–4.87	4.59	2.33	2.81
	(–1.44)	(–2.63)	(1.19)	(0.95)	(1.23)
GE	0.14	0.78	0.11	0.96	–0.29
	(0.12)	(1.30)	(0.12)	(1.13)	(–0.67)
GM	–3.74	–0.67	–0.30	–0.16	–3.30
	(–2.32)	(–0.99)	(–0.50)	(–0.21)	(–4.80)
HWP	–1.20	1.45	–5.55	3.78	3.26
	(–0.74)	(1.28)	(–2.32)	(2.95)	(4.30)
IBM	–0.64	0.51	–0.01	–0.17	–0.50
	(–0.68)	(0.80)	(–0.05)	(–0.16)	(–1.80)
JNJ	0.81	1.60	3.88	1.37	1.65
	(0.40)	(2.20)	(2.15)	(1.43)	(1.26)
KO	0.58	–0.20	2.43	–0.36	–0.30
	(0.53)	(–0.28)	(1.46)	(–0.52)	(–0.56)
MCD	2.56	–5.83	5.80	–3.34	–4.00
	(1.20)	(–3.75)	(1.56)	(–2.96)	(–2.82)
MO	0.17	–1.78	0.69	3.06	–0.01
	(0.28)	(–1.53)	(0.77)	(1.93)	(–0.05)
MOB	0.70	1.36	–2.02	–0.48	–0.32
	(0.69)	(1.32)	(–1.16)	(–0.56)	(–0.40)
MRK	–1.20	0.65	0.31	–3.23	1.03
	(–2.33)	(0.70)	(0.24)	(–2.06)	(1.37)
PEP	–0.55	5.71	2.01	–0.37	1.12
	(–0.32)	(5.82)	(0.83)	(–0.17)	(1.62)
PG	–0.28	–2.15	0.91	2.92	2.85
	(–0.24)	(–1.99)	(0.50)	(2.19)	(2.47)
S	–0.01	–2.69	–9.44	0.21	1.43
	(–0.04)	(–1.95)	(–2.80)	(0.08)	(0.96)
T	1.78	0.47	–0.49	–1.49	0.45
	(2.63)	(0.05)	(–0.32)	(–2.82)	(0.82)
WMT	1.58	–0.41	3.89	–8.31	3.69
	(0.89)	(–0.24)	(1.17)	(–2.35)	(1.83)
XON	1.66	–1.02	–0.35	0.20	1.04
	(2.56)	(–1.98)	(–0.28)	(0.38)	(2.02)
T	726	887	768	791	759
R^2	0.11	0.23	0.08	0.09	0.21

ever, 30% (193 of 625) of the t-statistics are significant at the 5% level, implying the importance of factors other than own-stock trading in determining price impact. Also, 18 of the 25 own-stock price-impact terms—that is, $\hat{a}_{i,i}$ (the ith diagonal entry of $\hat{\boldsymbol{A}}$)—are statistically significant. These terms should be the most dominant in determining price impact, and our regression confirms this conjecture. Nevertheless, cross-stock effects are significant.

Consider, for example, the AHP regression: while the own-price effect has a coefficient of 4.97, coefficients for BLS and FNM are 4.46 and –5.10. That these two cross-stock coefficients have opposite signs underscores the portfolio approach's importance for minimizing execution costs. Because of significant cross-stock price-impact effects, the expected cost of executing a portfolio is not simply the sum of the expected values of executing each security in isolation.

Although some regressions have low explanatory power, recall that we have proposed a rather naive specification for these regressions, omitting many other variables that proprietary traders and other professional portfolio managers have at their disposal. But even with our naive specification, we still achieve R^2's as high as 0.440 (for Merck, not shown in Table 3), which is quite substantial, considering the data's high frequency.

No-arbitrage constraints

One additional aspect of the estimation procedure must be considered: whether or not the parameter estimates yield a well-posed optimization problem (see Equations 1 and 2). In particular, for certain parameter values, the optimization problem is not convex, so the objective function can be made arbitrarily negative. The economic interpretation for such circumstances is an *arbitrage opportunity* (also known as a "free lunch"), a situation in which riskless profits can be manufactured out of thin air. Ordinarily, this would be a welcome state of affairs for investment professionals. In this case, the arbitrage is more likely a spurious side effect of sampling variation in our parameter estimates.

To avoid these false-arbitrage opportunities, a no-arbitrage restriction should be imposed on the estimation procedure. For the linear-percentage price-impact model, we accomplish this by constraining both $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{A}} \cdot \boldsymbol{R}$ to be positive definite matrices. This, in turn, involves estimating a constrained linear-regression model. Table 4 reports the results of such a procedure. The two most significant differences between Tables 3 and 4 are that the latter shows lower R^2's and the higher significance of the own-stock coefficients. The former is not surprising, because any constraint is bound to decrease the regression's explanatory power, although the decline is rather small for AN and DD. The higher significance of own-stock coefficients follows from the definition of positive definiteness—that is, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ for all vectors $\boldsymbol{x}$. The diagonal elements of $\boldsymbol{A}$ are coefficients of squared terms of the $\boldsymbol{x}$ values in the matrix product $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$. Therefore, by making the squared terms sufficiently large relative to the cross-terms, we arrive at a positive definite matrix.

As Table 4 shows, the cross-effects are also affected by the no-arbitrage constraint, highlighting its significance in the portfolio context. For example, in AHP's case, the coefficients of BLS and FNM are now smaller (3.49 and 0.01) than in Table 3 (4.46 and –5.10), where the no-arbitrage constraint has not been imposed. However, the coefficients of MCD and WMT become larger, increasing to 3.38 and 3.70 from 2.56 and 1.58 in Table 3.

Table 5 shows the ratio of the total sum of squared errors of the constrained regression to the unconstrained regression for all 25 stocks. The increase in squared errors is approximately 5% overall, a rather modest increase that provides some support for imposing the restriction. More important, if the no-arbitrage condition were not imposed, the dynamic-optimization algorithms described earlier might yield nonsensical results.

The empirical results of Tables 3 to 5 suggest that the state equations necessary for our dynamic-optimization algorithm can be estimated reasonably accurately and that a portfolio approach to execution-cost control has significant benefits.

Monte Carlo analysis

Having calibrated the state equation (see "The estimation procedure" section) for the linear percentage case (see "The state equations" section), we now investigate the performance of the best-execution strategy through Monte Carlo simulations. Specifically, we consider minimizing the execution costs of purchasing $\bar{s}$ shares of each of the 25 stocks in Table 1 over T periods. This occurs under the price dynamics (see Equations 3 to 6) where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ from the constrained regression (see the previous section) and $\boldsymbol{C}$, μ_z, and Σ_η are as we estimated in "The estimation procedure" section. We assume that the baseline covariance matrix of the no-impact price is Σ_z and that the initial no-impact prices are the prices in Table 1 (closing prices se-

Table 4. Coefficients of the constrained price-impact regressions for five stocks, based on market orders from 2 January to 31 December 1996. All coefficients have been multiplied by 10^{10} except for the SPX coefficients, which have been multiplied by 10^5. The last two rows contain sample size T and R^2 coefficients; t-statistics are in parentheses below the coefficients.

Variable	AHP	AN	BLS	CHV	DD
SPX	3.74	2.28	0.26	11.70	2.38
	(0.90)	(0.41)	(0.04)	(1.89)	(0.54)
AHP	12.40	–1.69	–1.99	1.04	–1.07
	(8.17)	(–1.70)	(–1.15)	(0.88)	(–0.96)
AN	–1.32	10.10	–1.96	1.09	–1.63
	(–1.40)	(9.98)	(–1.25)	(0.88)	(–1.53)
BLS	3.49	0.83	14.40	2.26	–2.45
	(1.56)	(0.48)	(4.82)	(1.11)	(–1.39)
CHV	2.09	–0.17	2.34	21.20	2.84
	(1.39)	(–0.22)	(1.19)	(6.72)	(2.83)
DD	–0.93	0.66	6.18	1.55	11.70
	(–0.68)	(0.57)	(3.60)	(1.05)	(8.72)
DIS	1.98	2.73	–0.30	–2.59	0.92
	(1.85)	(2.84)	(–0.09)	(–2.09)	(1.04)
DOW	–0.79	0.18	–3.88	–0.23	–1.59
	(–0.77)	(0.40)	(–2.19)	(–0.18)	(–2.13)
F	–0.23	0.66	3.29	–0.65	4.24
	(–0.09)	(0.50)	(0.82)	(–0.29)	(2.19)
FNM	0.01	–0.92	–0.43	3.27	0.39
	(0.00)	(–0.48)	(–0.11)	(1.31)	(0.17)
GE	0.77	–0.45	0.53	0.03	–0.41
	(0.66)	(–0.73)	(0.59)	(0.04)	(–0.92)
GM	–2.83	–0.86	0.23	0.28	–2.99
	(–1.72)	(–1.23)	(0.37)	(0.37)	(–4.25)
HWP	–1.66	0.53	–2.73	0.73	2.35
	(–1.00)	(0.46)	(–1.11)	(0.55)	(3.03)
IBM	–0.09	1.00	–0.77	–1.33	–0.24
	(–0.09)	(1.53)	(–0.47)	(–1.28)	(–0.84)
JNJ	–1.39	1.26	0.13	2.07	2.40
	(–0.67)	(1.70)	(0.07)	(2.11)	(1.78)
KO	1.11	0.61	0.69	–0.42	1.44
	(0.99)	(0.83)	(0.40)	(–0.60)	(2.58)
MCD	3.38	–3.73	–0.24	–2.48	–1.08
	(1.55)	(–2.34)	(–0.06)	(–2.15)	(–0.74)
MO	–0.63	–0.84	1.87	–1.75	0.25
	(–1.02)	(–0.71)	(2.05)	(–1.08)	(0.22)
MOB	0.36	–0.12	–0.98	–0.92	0.41
	(0.34)	(–0.11)	(–0.55)	(–1.03)	(0.48)
MRK	–0.75	0.55	–0.29	–0.18	1.30
	(–1.41)	(0.58)	(–0.22)	(–0.11)	(1.67)
PEP	–0.85	1.75	0.53	–1.93	–0.28
	(–0.48)	(1.74)	(0.21)	(–0.90)	(–0.39)
PG	–1.48	–1.94	0.89	3.51	2.70
	(–1.25)	(–1.75)	(0.47)	(2.56)	(2.29)
S	0.01	–1.76	–6.90	1.13	–0.30
	(0.00)	(–1.25)	(–2.00)	(0.43)	(–0.20)
T	1.84	0.70	–1.18	–0.20	0.59
	(2.67)	(0.77)	(–0.74)	(–0.36)	(1.04)
WMT	3.70	2.01	2.36	–3.65	0.96
	(2.02)	(1.15)	(0.69)	(–1.01)	(0.46)
XON	1.45	–0.14	0.77	–0.16	0.09
	(2.18)	(–0.26)	(0.61)	(–0.29)	(0.17)
T	726	887	768	791	759
R^2	0.06	0.19	0.03	0.04	0.17

Table 5. Ratios of the sum of squared residuals of the unconstrained and constrained price-impact regressions for 25 stocks, based on market orders from 2 January to 31 December 1996.

Ticker	Ratio	Ticker	Ratio	Ticker	Ratio	Ticker	Ratio	Ticker	Ratio
AHP	1.0501	DIS	1.0501	GM	1.0501	MCD	1.0501	PG	1.0500
AN	1.0498	DOW	1.0507	HWP	1.0416	MO	1.0499	S	1.0500
BLS	1.0500	F	1.0500	IBM	1.0508	MOB	1.0479	T	1.0281
CHV	1.0501	FNM	1.0497	JNJ	1.0501	MRK	1.0496	WMT	1.0500
DD	1.0501	GE	1.0500	KO	1.0494	PEP	1.0500	XON	1.0516

Table 6. Monte Carlo simulations of optimal execution strategies with short-sales constraints for purchasing $\bar{s}$ shares of all 25 stocks in the portfolio over T periods. Each row corresponds to an independent simulation experiment consisting of 10,000 IID replications. Costs are in cents per share. T is the number of execution periods. $\bar{s}$ is the amount of shares to purchase. k_v is the scaling of the volatility. s^* is the average execution cost for the optimal strategy with no short-sale constraints. s_c^* is the average execution cost for the optimal strategy with short-sale constraints. $\bar{s}/T$ is the naive strategy's average cost. Standard errors are in parentheses below the main entries.

T	$\bar{s}$	k_v	s^*	s_c^*	$\bar{s}/T$
20	10,000	1	−8.05	0.87	1.53
			(0.99)	(0.78)	(0.47)
20	20,000	1	−2.83	0.21	1.86
			(0.59)	(0.45)	(0.47)
20	30,000	1	−0.87	0.44	2.19
			(0.50)	(0.41)	(0.47)
20	50,000	1	1.09	1.35	2.85
			(0.46)	(0.42)	(0.47)
20	100,000	1	3.71	3.69	4.50
			(0.46)	(0.45)	(0.47)
20	200,000	1	7.50	7.48	7.80
			(0.46)	(0.46)	(0.47)
20	100,000	0.5	3.04	3.04	3.26
			(0.32)	(0.32)	(0.33)
20	100,000	2	3.79	3.75	3.45
			(0.62)	(0.61)	(0.66)
20	100,000	4	4.79	4.77	4.05
			(0.84)	(0.84)	(0.93)
5	100,000	1	13.72	13.55	13.63
			(0.26)	(0.26)	(0.26)
10	100,000	1	7.08	6.91	7.19
			(0.34)	(0.34)	(0.34)
15	100,000	1	4.95	4.81	5.20
			(0.41)	(0.40)	(0.41)
20	100,000	1	3.71	3.69	4.50
			(0.46)	(0.45)	(0.47)

lected from a random trading day in 1996).

To gauge the sensitivity of execution costs to the model's parameters, we vary the time horizon, T, the number of shares traded, $\bar{s}$ (assumed the same for each stock), and the no-impact price volatility. We modify the price volatility by scaling the variances by a constant while keeping the correlation structure fixed. The results for 10,000 replications are in Table 6, which reports the expected execution cost in cents per share. The table also lists the standard error of the estimate for the best-execution strategy (s^*), the strategy under no-sales constraints computed through the static-approximation method (s_c^*), and the naive strategy ($\bar{s}/T$).

Some general patterns emerge from the simulations. First, as T increases, execution costs fall. Because we can spread the trading over more time periods, and because we have the flexibility to be more patient and wait for particularly opportune times to trade, expected costs decline. Second, as $\bar{s}$ decreases, execution costs also decrease. With small-enough trade sizes, the expected price impact is negative! This is because the price impact consists of two terms: the impact of shares traded, s_t, which is quadratic in the share size, and the impact of information, which is linear in the information variable, x_t. When we trade small-enough quantities, the quadratic term is negligible and the information term dominates. Our strategy optimally uses information so as to trade when trading is least expensive. For sufficiently significant pieces of information, trading can be quite profitable (not a new insight to proprietary traders). Finally, increasing volatility seems to increase execution costs slightly.

In all but two cases, the optimal strategy outperforms the naive on average. In the two anomalous cases, the confidence interval of the difference between the two strategies is so wide that this outcome could easily have occurred purely by chance. If we increased the number of replications to 100,000, these two anomalies would no doubt disappear.

Another anomalous result is that for some simulations, the constrained strategy's execution cost is less than the unconstrained strategy's cost. Although the point estimates are indeed reversed in these cases, the sampling variation is so great (consider their standard errors) that making accurate inferences about their relative magnitudes is difficult. Indeed, the differences are not statistically significant. For the cases we consider, the no-sales constraint seems to have relatively little impact on the best-execution strategy's performance, except in cases with negative execution costs. To achieve negative execution costs, the no-sales constraints must be violated, so imposing them increases the costs dramatically.

Of course, these conclusions are highly portfolio- and time-period-specific. Similar analyses should be conducted case by case to determine the value added by the best-execution strategy in a given context.

The remaining challenge is to integrate these best-execution strategies directly into the investment process, which requires solving the portfolio optimization problem subject to transactions costs. This is a formidable challenge that is both theoretically and computationally intensive, and we plan to turn to these problems in future research.

Acknowledgments

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 (Grants 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, the London Business School, MIT, Northwestern, the NYSE Conference on Best Execution, the Society of Quantitative Analysts, and Yale for helpful comments and discussion.

References

1. T. Loeb, "Trading Cost: The Critical Link between Investment Information and Results," *Financial Analysts J.*, Vol. 39, 1983, pp. 39–44.
2. L. Chan and J. Lakonishok, "The Behavior of Stock Prices around Institutional Trades," *J. Finance*, Vol. 50, 1995, pp. 1147–1174.
3. D. Keim and A. Madhavan, "Anatomy of the Trading Process: Empirical Evidence on the Behavior of Institutional Traders," *J. Financial Economics*, Vol. 37, No. 3, Mar. 1995, pp. 371–398.
4. D. Keim and A. Madhavan, "The Upstairs Market for Large-Block Transactions: Analysis and Measurement of Price Effects," *Rev. Financial Studies*, Vol. 9, No. 1, Spring 1996, pp. 1–36.
5. D. Keim and A. Madhavan, "Execution Costs and Investment Performance: An Empirical Analysis of Institutional Equity Trades," working paper, School of Business Administration, Univ. of Southern California, Los Angeles, 1995.
6. D. Bertsimas and A. Lo, "Optimal Control of Execution Costs," *J. Financial Markets*, Vol. 1, No. 1, 30 Apr. 1998, pp. 1–50.
7. A. Shleifer, "Do Demand Curves for Stocks Slope Down?" *J. Finance*, Vol. 41, 1986, pp. 579–590.
8. A. Lo and C. MacKinlay, "When Are Contrarian Profits Due to Stock Market Overreaction?" *Rev. Financial Studies*, Vol. 3, No. 2, Summer 1990, pp. 175–205.
9. L. Birinyi, *What Does Institutional Trading Cost?* Birinyi Associates, Greenwich, Conn., 1995.
10. D. Leinweber, "Using Information from Trading in Trading and Portfolio Management," *Execution Techniques, True Trading Costs, and the Microstructure of Markets*, K. Sherrerd, ed., Assoc. for Investment Management and Research, Charlottesville, Va., 1993.
11. D. Leinweber, "Careful Structuring Reins in Transaction Costs," *Pensions and Investments*, 25 July 1994, p. 19.
12. A. Perold, "The Implementation Shortfall: Paper versus Reality," *J. Portfolio Management*, Vol. 14, No. 3, Spring 1988, pp. 4–9.
13. D. Bertsekas, *Dynamic Programming and Optimal Control, Vol. I*, Athena Scientific, Belmont, Mass., 1995.

Dimitris Bertsimas is the Boeing Professor of Operations Research at the MIT Sloan School of Management. He studies the theory and practice of optimization, as well as performance, analysis, and control of large-scale stochastic systems. He received his PhD in operations research and applied mathematics from MIT. Contact him at the MIT Sloan School of Management and Operations Research Center, Massachusetts Inst. of Technology, E53-363, Cambridge, MA 02142-1347; dbertsim@mit.edu.

Paul Hummel is a portfolio strategist at Long Term Capital Management. His responsibilities include trading in risk arbitrage and special situations; designing automated spread-trading, volume-weighted-average-price, and options-hedging strategies; and research support for short-term, high-frequency trading strategies. He received his MS in mechanical engineering from Cornell University and his MBA from the MIT Sloan School of Management. Contact him at Portfolio Strategist, Long Term Capital Management, One E. Weaver St., Greenwich, CT 06831; prhummel@aol.com.

Andrew W. Lo's biography appears on p. 23.