Explicit Methods for Portfolio Optimization with Realistic Costs and Constraints

GORDON RITTER*†

†Baruch College (CUNY), New York University, and Columbia University, New York, NY

(*submitted: August 20, 2025; revised: July 6, 2026*)

We present algorithms for mean-variance optimization with realistic costs, including borrow costs, bid-offer spread cost, and market impact according to the well-known square root law. The algorithm performs coordinate updates, each of which reduces to a univariate subproblem with a closed-form algebraic solution. The technique extends to problems with box constraints, and to certain kinds of multi-horizon objectives. We demonstrate an improvement in the computational complexity when the covariance matrix has a factor structure.

Keywords: Finance; Investment analysis; Portfolio optimization; Transaction costs; Closed-form solutions

JEL Classification: G11, C61, C63

1. Introduction

For institutional traders, one of the largest sources of cost is market impact. Tóth *et al.* (2011) present convincing empirical evidence that the price impact of large meta-orders has a square root dependence on the trade size as a fraction of volume. Researchers at Barra, especially Nicolò Torre and Mark Ferrari, identify a square root dependence in a dataset collected by Loeb (1983), and present the square root model for industrial use (Torre 1997). Grinold and Kahn (2000) cite the Barra handbook, and suggest a model including spread cost and square root impact in eq. (16.4) of their famous book. Kyle and Obizhaeva (2016, Section 6) show that both linear and square root impact models are consistent with market microstructure invariance, but the square root model has a simpler, more elegant coefficient structure. Almgren *et al.* (2005) fit an exponent of 0.6 to a particular dataset, confirming that the exponent is close to 0.5.

Zarinelli *et al.* (2015) find that the square root law is a good approximation over two orders of magnitude, beyond which a logarithmic form seems better. Bucci *et al.* (2019) document a cross-over from linear to square root impact. More recently, Sato and Kanazawa (2024) show, using a comprehensive microscopic dataset from the Tokyo Stock Exchange, that the exponent is equal to 1/2 within statistical errors at both the individual stock level and the individual trader level.

Given the resilience of the square root law across multiple studies, markets and asset classes, it is of clear interest to study realistic portfolio optimization methods which combine square root impact with spread cost, as suggested by Grinold and Kahn (2000) and Kyle *et al.* (2016). For

*Corresponding author. Email: ritter@post.harvard.edu

long-short equity managers, costs associated with stock loan financing, called *borrow costs*, are also an important source of cost.

In the present work, we show that mean-variance optimization with square root impact, spread cost, and borrow costs can be solved in *semi-closed form*. Precisely, the optimal trades can be found by repeatedly solving single-asset subproblems, where each subproblem has a fully closed-form solution. Hence, for this class of problems, complicated and potentially opaque general-purpose numerical solvers are completely unnecessary. This remains true even after including box constraints, and for a restricted class of multi-horizon problems. We also show that, when a factor structure is present, our technique takes full advantage of the reduction of dimensionality afforded by the factor model.

The inclusion of borrow cost is a novel element here, at least as far as closed-form solutions are concerned. The market microstructure literature does not usually concern itself with borrow costs; nevertheless, for a long-short equity manager, borrow costs are as important as market impact. Anecdotally, lending institutions charge in the range of 20–30 basis points (in annualized units) for stock loans under normal conditions and much higher rates when shares available for lending are in short supply. Unlike market impact or spread cost, which are purely functions of the traded quantity, borrow costs depend on the final position. Fortunately, the single-asset subproblem still admits a fully closed-form solution.

The rest of the paper is organized as follows. In Section 2, we derive the objective function and set up a coordinate descent method inspired by the Gauss–Seidel method and by Friedman *et al.* (2007). We then work out the precise form of the single-asset subproblem, and show that the calculation of subproblem parameters is much more efficient if the covariance matrix has a factor structure. Furthermore, we give a restricted class of multi-horizon problems that can be treated by the same algorithm. Section 3 calculates the subproblem optimum explicitly in closed form as an algebraic expression. We conclude with a discussion of box constraints.

We expect these results to be of interest to any portfolio manager who uses the square root law or short selling in some way, and who constructs portfolios using an optimizer. This includes, in particular, most quantitative or systematic equity strategies in the world today.

2. Reduction to the one-dimensional case

2.1. *Portfolio selection*

Let $w_t \in \mathbb{R}$ denote an investor's wealth at time t, with the current time at $t = 0$. We shall refer to the currency of w_t as *dollars*, but the framework applies to any currency.

The investor is concerned with optimizing expected utility of future wealth w_T at some future time $T > 0$, called the *investment horizon*. Let $h \in \mathbb{R}^n$ denote the investor's initial $t = 0$ holdings denominated in dollars. Share holdings are valued using prices that are observable at $t = 0$; for example, the most recent arrival price from the order book. The same prices that are used to value the positions will also be used to convert meta-orders between shares and dollars, and as a benchmark to calculate implementation shortfall; for this reason they are called *benchmark prices*. The sign of h_i, positive or negative, indicates a long or a short position in asset i.

Let $R \in \mathbb{R}^n$ denote the cross section of asset returns over the interval $[0, T]$, where the pricing methodology used to compute returns must agree with the benchmark pricing methodology. As viewed at $t = 0$, of course R is a random variable.

The investor contemplates n meta-orders with dollar amounts $v_i, i = 1, \ldots, n$. Here $v_i > 0$ denotes a buy order for asset i, $v_i < 0$ denotes a sell order, $v_i = 0$ means no trade shall be ordered in asset i, and the orders are converted from shares to dollars using benchmark prices. If all orders could be executed at benchmark price, then

$$w_T = w_0 + (h+v)'R = w_0 + h'R + v'R \tag{1}$$

where prime denotes the transpose. In real markets, the orders cannot necessarily be filled at benchmark price and additional costs should be subtracted from (1) as detailed in the next section; for example (11) and (12).

Assume a filtered probability space with filtration $\mathcal{F}$ where $\mathcal{F}_t$ denotes the information set at time t, with $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$, and similarly for $\mathbb{V}_t$. Benveniste *et al.* (2024) show that, for a broad class of asset return models, there exists $\kappa > 0$ such that utility maximization has the same solution as the following mean-variance optimization (MVO) form:

$$\underset{v \in \mathbb{R}^n}{\operatorname{argmax}} \quad \mathbb{E}_0[w_T] - \frac{\kappa}{2}\mathbb{V}_0[w_T], \tag{2}$$

where w_T depends on v through (1). Further define

$$r := \mathbb{E}_0[R] \in \mathbb{R}^n, \quad \text{and} \quad \Omega := \mathbb{V}_0[R] \in \mathbb{R}^{n \times n} \tag{3}$$

where Ω is a strictly positive definite $n \times n$ covariance matrix. The MVO problem (2) is then, concretely,

$$\underset{v \in \mathbb{R}^n}{\operatorname{argmax}} \; F_{\mathrm{MV}}(v) \tag{4}$$

$$F_{\mathrm{MV}}(v) := (h+v)'r - \frac{\kappa}{2}(h+v)'\Omega(h+v) - \mathrm{tcost}(h,v) \tag{5}$$

where $\mathrm{tcost}(h,v)$ denotes the sum of all costs associated with starting from portfolio h, trading to portfolio $p = h + v$, and then holding p until time T. The problem (4) is, therefore, our primary objective, and the purpose of this paper is to solve this problem with realistic functions $\mathrm{tcost}(h,v)$, making use of explicit formulas rather than optimization heuristics.

The precise form of $\mathrm{tcost}(h,v)$ that is typically of interest to large institutional traders is the topic of the next section, but we posit two assumptions that govern its structure.

Assumption A1 The cost function is simply additive across assets,

$$\mathrm{tcost}(h,v) := \sum_{i=1}^{n} \mathrm{tcost}_i(h_i, v_i). \tag{6}$$

Assumption A2 The function tcost_i is strictly convex in its second argument, for all $i = 1, \ldots, n$.

It follows that F_{MV} is strictly concave, and therefore that minimizing $-F_{\mathrm{MV}}(v)$ is a strictly convex optimization.

2.2. *Transaction costs*

The square root law is defined as follows; we use notation from Kyle and Obizhaeva (2018). Let G denote the percentage cost of executing a meta-order of Q shares of stock with price P, expressed as a fraction of the value of the bet PQ. Let σ^2 denote the asset's return variance per day, and let

V denote the asset's trading volume in shares per day. Then

$$G \sim \sigma\Big(\frac{|Q|}{V}\Big)^{1/2}. \tag{7}$$

For example, if daily volatility is 2 percent then execution of 5 percent of daily trading volume is expected to cost a trader $0.02(0.05)^{1/2}$ times the dollar value traded, i.e. about 45 basis points.

The Kyle–Obizhaeva form (7) admits an equivalent representation in dollars. Indeed, we shall use (7) as a guide to obtain an approximation to the dollar cost $\mathrm{tcost}_i(h_i, v_i)$ required by our objective. In (6) we consider a trade v_i in dollars, so v_i represents PQ for the i-th asset. The inner ratio from (7) can be proxied by

$$\frac{|Q_i|}{V_i} \approx \frac{|v_i|}{\mathrm{adv}_i} \tag{8}$$

where adv_i denotes the average daily *dollar* volume of the i-th asset. Since G represents a fraction, to obtain the cost in dollars we must multiply (7) by the traded notional $|v_i|$.

Our version of (7) expresses the cost in dollars as

$$\begin{aligned} \mathrm{tcost}_i(h_i, v_i) &= \beta_1 \sigma_i \Big(\frac{|v_i|}{\mathrm{adv}_i}\Big)^{1/2} |v_i| & (9)\\ &= c_i |v_i|^{3/2}, \text{ where } \; c_i := \beta_1 \sigma_i \mathrm{adv}_i^{-1/2} > 0, & (10) \end{aligned}$$

and where β_1 is an overall constant that is fitted empirically, but which according to multiple sources including Grinold and Kahn (2000) and Kyle and Obizhaeva (2018), should be approximately 1.

Gatheral (2016) states that practitioners' models almost always incorporate bid-offer spread cost in addition to the square root law. An example is eq. (18) of Kyle *et al.* (2016). Therefore we modify (9) by adding an additional term,

$$\mathrm{tcost}_i(h_i, v_i) = c_i |v_i|^{3/2} + s_i |v_i| \tag{11}$$

where $s_i > 0$ denotes the predicted spread of the asset, scaled to the fraction of spread we expect to pay. For example, if we expect to pay a quarter of the full spread, then on a stock with spread 20bps, we would set $s_i = 5$bps.

We now define the model's borrow costs. When the final position $h_i + v_i$ is negative, it consists of borrowed shares, and the broker charges a fee equal in magnitude to the final position multiplied by a lending rate $\ell_i > 0$. The lending rate is specific to the stock and the time period in question, and is negotiated with the broker; when lendable shares are scarce, it can be orders of magnitude higher than normal. This suggests augmenting (11) as follows:

$$\mathrm{tcost}_i(h_i, v_i) = c_i |v_i|^{3/2} + s_i |v_i| - \ell_i \min(h_i + v_i, 0) \tag{12}$$

where c_i, s_i, and ℓ_i are all positive. Note that (9) and by extension (11) and (12) are strictly convex as functions of v_i, in line with Assumption A2.

2.3. *Coordinate descent*

The efficacy of coordinate descent methods is very well known in the statistics community. Friedman *et al.* (2007) show that coordinate descent is very competitive with the LARS procedure in large lasso problems, and can be applied to many other convex statistical problems such as the garrote and elastic net. Yen and Yen (2014) apply coordinate descent to standard Markowitz problems with L^1 and L^2 constraints, which works because those problems are mathematically equivalent to the elastic net.

As in statistics, coordinate descent is a standard technique for mean-variance optimization, especially in the multi-period case. Kolm and Ritter (2014) suggest it as a solution technique for a broad class of multi-period mean-variance models that are equivalent to Bayesian dynamic models. Xie *et al.* (2018) apply block coordinate ascent to mean-variance policy search in reinforcement learning, specifically in contexts where the variance of the reward is part of the objective.

Friedman *et al.* (2007) also note that the reason for the method's effectiveness is that each of the individual coordinate optimizations can be done very quickly; our main point here is to show that for realistic cost functions that portfolio managers care about, the individual coordinate optimizations can be done in closed form.

Algorithm 2.1
Inputs: *starting portfolio* $h \in \mathbb{R}^n$, *numerical tolerance* $\epsilon_{\text{tol}} > 0$, *objective function* F_{MV}.
Initialize: $v = \mathbf{0} \in \mathbb{R}^n$.

1. **for** $i = 1, \ldots, n$ **do**:
   (a) **let** $x^* \in \mathbb{R}$ *be the point that maximizes the real function*

$$x \longmapsto F_{\text{MV}}\,(v + x\,\mathbf{e}_i) \tag{13}$$

   where $\mathbf{e}_i$ *is the i-th standard basis vector in* $\mathbb{R}^n$.
   (b) **increment** $v_i \leftarrow v_i + x^*$.
2. *repeat step 1 if any of the values* $|x^*|$ *computed in step 1 exceeds* ϵ_{tol}.

Return: $v \in \mathbb{R}^n$ *as the optimal trade vector.*

Defining $p := h + v$, the variance term from $F_{\text{MV}}\,(v + x\,\mathbf{e}_i)$ can be written

$$\begin{aligned}(p + x\,\mathbf{e}_i)'\Omega(p + x\,\mathbf{e}_i) &= p'\Omega p + 2p'\Omega(x\,\mathbf{e}_i) + (x\,\mathbf{e}_i)'\Omega(x\,\mathbf{e}_i)\\ &\longrightarrow 2\,(p'\Omega\mathbf{e}_i)\,x + \Omega_{ii}x^2\end{aligned}$$

where the arrow indicates dropping terms not depending on x, such as $p'\Omega p$. The single-asset subproblem from step 1(a) then reduces to

$$\max_x[-a_i x^2 + b_i x - \text{tcost}_i(h_i, v_i + x)] \tag{14}$$

$$a_i := \frac{\kappa}{2}\Omega_{ii}, \quad b_i := r_i - \kappa\, p'\Omega\mathbf{e}_i\,. \tag{15}$$

Remark 2.1 The coefficients a_i do not change during the algorithm loop, so they can be precomputed for efficiency. However, in step 1(b), the value of v_i is updated and hence the value of $p := h + v$ must also update. Therefore all single-asset problems must recompute b_i inside the loop, leading to $O(n^2)$ complexity. We show in the next section that with additional model assumptions, this computational burden can be avoided, leading back to $O(n)$ complexity.

The remainder of this section proves that, for our class of problems, the above method converges to the global optimum. The mathematical heavy lifting was done in Tseng (1988) and Tseng (2001), but Tseng's results are extremely general and some work is required to translate them into our setting. Most of that work was done in the textbook of Beck (2017), and we do the rest here.

Proposition 2.2 *Let*

$$F(x) = f_0(x) + \sum_{k=1}^{n} f_k(x_k), \quad x \in \mathbb{R}^n, \tag{16}$$

where $f_0 \in C^\infty(\mathbb{R}^n)$ is strictly convex and each $f_k : \mathbb{R} \to \mathbb{R}$ is convex. Assume existence of $x^0 \in \mathbb{R}^n$ with compact sublevel set $\{x \in \mathbb{R}^n : F(x) \leq F(x^0)\}$. Let $\{x^r\}$ be the exact coordinate descent iterates starting from x^0. Then F has a unique global minimizer $x^\star$, and $x^r \to x^\star$.

Proof. By Beck (2017, Theorem 14.9), the sequence $\{x^r\}$ is bounded and every limit point of $\{x^r\}$ is an optimal solution. Since F is strictly convex, that optimal solution is unique. If a sequence in $\mathbb{R}^n$ is bounded and has exactly one limit point $x^\star$, then the whole sequence converges to $x^\star$. (If $\{x^r\}$ did not converge to $x^\star$, then some subsequence would stay a positive distance away from $x^\star$; by Bolzano–Weierstrass, that subsequence would admit a convergent subsequence, whose limit would be different from $x^\star$, a contradiction). Hence $x^r \to x^\star$. □

Corollary 2.3 *The iterates in Algorithm 2.1, step 1, converge to the global optimum.*

Proof. The function $-F_{\mathrm{MV}}$ takes the form (16) with f_0 a strictly convex quadratic function given by -1 times the expected return and risk terms, and $f_i(x) = \mathrm{tcost}_i(h_i, x)$. The quadratic function f_0 is strictly convex because Ω is strictly positive definite, and f_i is convex because it is given by (12), which we have already noted is convex. The initial sublevel set for $v = \mathbf{0}$ is bounded because, as any trade vector tends to infinity, so do variance and costs. It is easy to see that the initial sublevel set contains all of its limit points, hence is closed and also compact by Heine–Borel. □

2.4. *Factor model covariance*

Instead of allowing very complex forms for the multivariate asset return distribution, representations based on arbitrage pricing theory (APT) constrain the distribution's complexity by assuming a linear structure; see Ross (1976), Roll and Ross (1980). Models of this form have been marketed by Axioma and Barra, extensively documented by their researchers (e.g. Menchero *et al.* (2008)), and have become ubiquitous in industry. One of the consequences of the multifactor linear structure is the following decomposition of the asset-level covariance matrix

$$\Omega = XFX' + D, \quad D = \mathrm{diag}(\omega_1^2, \ldots, \omega_n^2), \quad \omega_i > 0 \ \forall i \tag{17}$$

where X is an $n \times k$ matrix of factor loadings, $k \ll n$, and F is a $k \times k$ positive-definite factor covariance matrix. In typical applications to the US equity market, $k \approx 50$ and $n \approx 2{,}000$, so the reduction in dimensionality is extremely significant. The diagonal elements ω_i^2 are called *residual variance* or *idiosyncratic (idio) variance*.

Let $\mathrm{x}_i \in \mathbb{R}^k$ denote the i-th row of X, transposed to be a column vector. Plugging the factor model covariance (17) into the subproblem coefficients (15) yields

$$a_i = \frac{\kappa}{2}(\mathrm{x}_i' F \mathrm{x}_i + \omega_i^2) \tag{18}$$

$$b_i = r_i - \kappa(p' X F \mathrm{x}_i + p_i \omega_i^2)\,. \tag{19}$$

The form of (19) suggests a computational speedup: keep a running calculation of the factor exposures as a column vector $E := X'p \in \mathbb{R}^k$, initialized as $X'h$. Each time the single-asset optimizers (step 1 of the algorithm) select a new trade, E is incremented by the change in factor exposure that the new trade implies; this update is $O(k)$ where $k \ll n$.

Algorithm 2.2 (Fast Version)
Inputs: *starting portfolio* $h \in \mathbb{R}^n$; *expected returns* $r \in \mathbb{R}^n$; *factor matrices* X, F, D *as in* (17); *cost parameters* c_i, s_i, ℓ_i; *risk aversion* $\kappa > 0$; *numerical tolerance* $\epsilon_{\mathrm{tol}} > 0$.
Initialize: $v = \mathbf{0} \in \mathbb{R}^n$, $E = X'h \in \mathbb{R}^k$. *Precompute* a_i $(i = 1, \ldots, n)$ *using* (18).

1. **for** $i = 1, \ldots, n$ **do**:
   (a) **let** $x^* \in \mathbb{R}$ *be the point that maximizes the real function*

   $$-a_i x^2 + b_i x - \mathrm{tcost}_i(h_i, v_i + x)$$

   where $b_i = r_i - \kappa(E' F \mathrm{x}_i + (h_i + v_i)\omega_i^2)$. *Compute* x^* *using closed-form algebraic expressions of Section 3.*
   (b) **increment** $v_i \leftarrow v_i + x^*$ *and* $E \leftarrow E + x^* \mathrm{x}_i$.
2. *repeat step 1 if any of the values* $|x^*|$ *computed in step 1 exceeds* ϵ_{tol}.

Return: $v \in \mathbb{R}^n$ *as the optimal trade vector.*

Remark 2.4 The computations of b_i in step 1(a), and the updates in step 1(b), are $O(k)$, which explains Algorithm 2.2's superior performance over Algorithm 2.1. When $k \ll n$ this resolves the computational bottleneck mentioned in Remark 2.1.

Remark 2.5 We performed computational experiments with Algorithm 2.2 using a test portfolio of $n = 2{,}000$ assets and $k = 50$ factors, with $\epsilon_{\mathrm{tol}} = 0.01$. On our test hardware (PowerEdge R660 server with Intel Xeon Gold CPU, single-threaded execution), Algorithm 2.2 converged in 50 iterations (each iteration being one full sequence of steps 1 and 2) and took 200 milliseconds. One cannot compare to quasi-Newton or even gradient-based optimizers since the cost term is non-differentiable, so instead we compared to a Java implementation of Powell's conjugate direction method, which took several seconds to solve the same objective. The actual speedup is highly problem-dependent and hardware-dependent, but our numerical tests showed significant outperformance. One can also apply Brent's method to step 1(a), but the closed-form algebraic expressions of Section 3 are superior to Brent's method because they give the exact answer in one step.

2.5. *Multi-horizon problems*

Consider the setup of Section 2.1, which ends with target positions $p \in \mathbb{R}^n$ at time T. We augment this setup by assuming further that for $t > T$, the investor plans a continuous-time trading path $\{x_t\}$ starting from $x_T = p$, which optimizes an integrated mean-quadratic-variation objective; the precise form of this objective is given in (20). A dot over a variable denotes the time derivative (Newton notation), and prime denotes transpose. Following Baldacci *et al.* (2022), $\mathcal{A}$ denotes the space of $\mathbb{R}^n$-valued adapted processes that are almost surely differentiable with square-integrable derivative, Λ is a symmetric positive-definite $n \times n$ market-impact coefficient matrix, and μ_t is an adapted process representing expected returns.

The problem to be solved, then, is the following augmentation of (4):

$$\underset{p \in \mathbb{R}^n}{\operatorname{argmax}} \quad \Big(F_{\mathrm{MV}}(p-h) + \sup_{\substack{x \in \mathcal{A} \\ x_T = p}} \mathbb{E}_0 \int_T^\infty L(x_t, \dot{x}_t)\, dt \Big), \tag{20}$$

where $p = h + v \in \mathbb{R}^n$ denotes the target positions at time T, and L denotes the running objective along the trading path x_t, combining expected return, a quadratic-variation penalty, and instantaneous impact cost:

$$L(x_t, \dot{x}_t) := x_t' \mu_t - \frac{\kappa}{2} x_t' \Omega x_t - \frac{1}{2} \dot{x}_t' \Lambda \dot{x}_t \,.$$

The second term in (20), considered as a function of p, is an example of a value function in stochastic optimal control theory (Bellman 1957).

The results of Baldacci *et al.* (2022) imply an explicit form for the value function in precisely this case:

$$\sup_{x \in \mathcal{A};\, x_T = p} \mathbb{E}_T \int_T^\infty L(x_t, \dot{x}_t)\, dt = -\frac{1}{2} p' \Lambda \Gamma p + p' \Lambda b_T + C_T \,, \tag{21}$$

$$\text{where} \quad \Gamma := \Lambda^{-1/2} \left(\kappa \Lambda^{-1/2} \Omega \Lambda^{-1/2} \right)^{1/2} \Lambda^{1/2}, \tag{22}$$

$$b_t := \int_t^\infty e^{-\Gamma(s-t)} \Lambda^{-1} \mathbb{E}_t[\mu_s]\, ds \,, \tag{23}$$

and where C_T does not depend on p. The expression for Γ in (22) is the principal square root of $\kappa \Lambda^{-1} \Omega$. Indeed, Baldacci *et al.* (2022, Theorem 1) show that the optimal trajectory solves $\dot{x}_t = -\Gamma x_t + b_t$; matching this to the optimal control implied by a quadratic value function yields the quadratic coefficients in (21). Note that b_T is $\mathcal{F}_T$-measurable, hence random as viewed from the decision time $t = 0$. Taking $\mathbb{E}_0$ of (21) and using the tower property of conditional expectation, the second term of (20) becomes (by the dynamic programming principle, and dropping the constant $\mathbb{E}_0[C_T]$) the deterministic quadratic function

$$-\frac{1}{2} p' \Lambda \Gamma p + p' \Lambda \bar{b}_T, \quad \text{where} \;\; \bar{b}_T := \int_T^\infty e^{-\Gamma(s-T)} \Lambda^{-1} \mathbb{E}_0[\mu_s]\, ds \,. \tag{24}$$

Plugging the quadratic function (24) back into (20), noting that $\Lambda\Gamma$ is symmetric and positive definite, we find that the objective function in (20) has the mathematical form of a concave quadratic minus trading costs (6). In other words, it has the same functional form as (4) with different input parameters; any objective of that form can be optimized by Algorithm 2.1. A major advantage of this approach is that (20) still involves only n decision variables – the components of p. The disadvantage is that beyond the horizon T, trading costs are approximated using linear instantaneous impact, but it is precisely this approximation which renders the infinite-horizon problem tractable and leads to the closed form (24).

3. Exact solutions for the one-dimensional case

We derive closed-form solutions to the single-asset problems in Algorithms 2.1 and 2.2,

$$\max_x [-a_i x^2 + b_i x - \mathrm{tcost}_i(h_i, v_i + x)] \,.$$

It is more convenient to solve for the full second argument $v_i + x$ of tcost_i, because then expressions involving the square root law are centered at zero trade, which has zero impact. We therefore make the simple affine change of variables $u = v_i + x$ and solve for u instead of x. Dropping constant terms, the objective becomes

$$-a_i u^2 + (2a_i v_i + b_i)u - \mathrm{tcost}_i(h_i, u) \,. \tag{25}$$

So the shift merely adjusts the linear term. Once we have found the optimal point u^*, we shift back, getting $x^* = u^* - v_i$ for use in step 1(a).

The rest of this section is devoted to explicit closed-form solutions to (25) for the three main types of trading costs (impact, spread, borrow). For that we will assume a single asset i is fixed, and drop the subscripts i. To keep notation compact, in the rest of Section 3 we let b denote a generic linear-term coefficient, but in application to (25), for example b refers to $2a_i v_i + b_i$.

3.1. *The continuously-differentiable case*

Define

$$f(u) := -au^2 + bu - c|u|^{3/2} \quad \text{where } a > 0 \text{ and } c > 0 \tag{26}$$

where positivity of a, c is required by the problem setup in (10) and in (15); variance, daily volume and volatility must all be positive for the problem to make sense.

Theorem 3.1 *Assume* (26)*; then f is concave, bounded above, continuously differentiable, and has a unique global maximum; the optimal point is given by*

$$u_*(b) = \mathrm{sgn}(b)\, \frac{8b^2}{16a|b| + 9c^2 + 3c\sqrt{9c^2 + 32a|b|}} \,, \tag{27}$$

where $\mathrm{sgn}(b)$ *is* $-1, 0,$ *or 1 depending on whether b is negative, zero, or positive.*

Proof. Under these conditions, f is concave and bounded above. Hence f' is monotone decreasing. Also, f is differentiable at $u = 0$ and $f'(0) = b$. Since $a > 0, c > 0$, it follows that f increases monotonically to its maximum u_*, and then decreases monotonically for $u > u_*$. We can infer whether the maximum is on $\mathbb{R}_+$ or $\mathbb{R}_-$ based on the sign of $f'(0) = b$. If $b > 0$ then f is still increasing at $u = 0$ and the maximum is in $\mathbb{R}_+$, while if $b < 0$ then f is decreasing at $u = 0$ and the maximum is in $\mathbb{R}_-$. If $b = 0$ then the maximum is easily seen to be at $u_* = 0$. Let f_+ and f_- denote the function f on the positive and negative real axis, respectively:

$$f_+(u) = -au^2 + bu - cu^{3/2}, \quad u \geq 0$$
$$f_-(u) = -au^2 + bu - c(-u)^{3/2}, \quad u < 0.$$

If $b > 0$ then the maximum satisfies the first-order condition for f_+, i.e. $-2au + b - \frac{3}{2}c\sqrt{u} = 0$

while if $b < 0$ then the maximum satisfies the first-order condition for f_-. Solving the two first-order conditions gives $\mathrm{sgn}(b)(16a|b| + 9c^2 - 3c\sqrt{9c^2 + 32a|b|})/(32a^2)$, which is correct but not numerically stable when $32a|b| \ll 9c^2$, due to catastrophic cancellation. Multiplying numerator and denominator by the conjugate yields the numerically stable form given in (27). $\square$

3.2. *Including bid-offer spread cost*

Now consider the inclusion of a scaled absolute-value term. This allows inclusion of bid-offer spread costs in addition to market impact costs. We continue to use the function u_* that was found in (27).

THEOREM 3.2 *Under the same assumptions and notation of Theorem 3.1, and for $s > 0$, the optimal point of the concave objective function*

$$g(u) := f(u) - s|u|$$

is located at $u_*(b - s)$ if $b > s$, and $u_*(b + s)$ if $b < -s$, and 0 if $b \in [-s, s]$.

Proof. The derivative is $g'(u) = f'(u) - \mathrm{sgn}(u)s$, where

$$f'(u) = -2au + b - \mathrm{sgn}(u)\frac{3}{2}c\sqrt{|u|} \tag{28}$$

If $b \in [-s, s]$, then $g'(u) > 0$ for negative u and $g'(u) < 0$ for positive u, so the maximum of g in this case is at 0. Otherwise, if $b \notin [-s, s]$ then $g'(u)$ does not change sign at $u = 0$. If $b > s$ then by the proof of Theorem 3.1, the optimal point of f lies in $\mathbb{R}_+$. Let $\tilde{u}$ denote the optimal point for g. Suppose $\tilde{u} < 0$; then $g(\tilde{u}) = f(\tilde{u}) + s\tilde{u} > g(-\tilde{u}) = f(-\tilde{u}) + s\tilde{u}$, which implies $f(\tilde{u}) > f(-\tilde{u})$. However, the even terms of f cancel in the difference, leaving $f(-\tilde{u}) - f(\tilde{u}) = -2b\tilde{u} > 0$ since $b > 0$ and $\tilde{u} < 0$; a contradiction, hence $\tilde{u} \geq 0$. On the positive half-line the function $g(u)$ is given by the same formula as f_+ but with b replaced by $b - s$. As long as $b - s > 0$, the location of the critical point of g is (27), after making the substitution $b \to b - s$. Hence the critical point is at $u_*(b - s)$ when $b > s$. The proof for the case $b < -s$ is the mirror image. $\square$

3.3. *Including spread and borrow costs*

We now consider adding borrow costs to the model, which already included spread costs. Given an initial position h and trade u, the final position is $h + u$. The lending rate ℓ will be charged on $|h + u|$ but only when $h + u < 0$. Define the objective function

$$\phi(u) = f(u) + \phi_1(u) \tag{29}$$

$$\phi_1(u) := -s|u| + \ell \min(h + u, 0)\,. \tag{30}$$

Globally, ϕ is increasing up to its global maximum, and decreasing thereafter; see Figure 1. The derivative $\phi'(u)$ is globally decreasing. Also, $\phi'(u)$ is possibly discontinuous only at the left and right cutpoints

$$c_L := \min(-h, 0), \quad c_R := \max(-h, 0) \tag{31}$$

and continuous elsewhere.

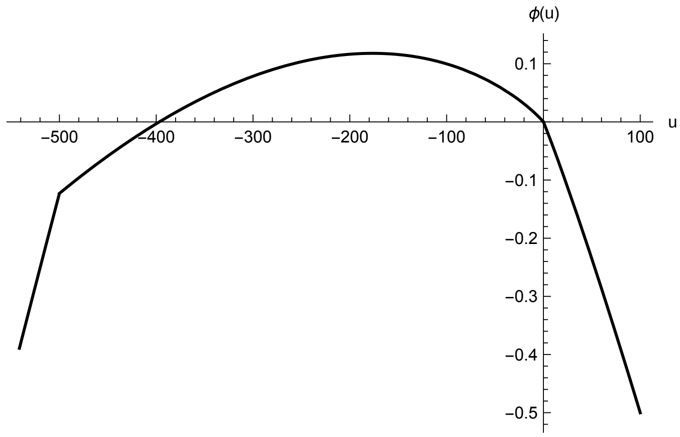

Figure 1. Example objective function ϕ with parameters $a = 2 \times 10^{-8}$; $b = -0.003$; $c = 0.0001$; $s = 0.001$; $\ell = 0.005$; $h = 500$.

The optimum of ϕ is of course located at the unique point where $\phi'(u)$ changes sign from positive to negative. To determine whether the slope changes sign at one of the cutpoints, let d_1, d_2 be the left-derivative and right-derivative of ϕ at the left cutpoint c_L. Similarly, let d_3, d_4 be the left-derivative and right-derivative of ϕ at the right cutpoint c_R. We calculate the precise values of $\{d_i\}$ in the following lemmas.

Lemma 3.3 *If $h > 0$, in which case $c_L = -h$ and $c_R = 0$, then*

$$\begin{aligned}
d_1 &= \lim_{\epsilon \to 0+} \phi'(-h-\epsilon) = 2ah + \frac{3c\sqrt{h}}{2} + b + s + \ell \\
d_2 &= \lim_{\epsilon \to 0+} \phi'(-h+\epsilon) = d_1 - \ell \\
d_3 &= \lim_{\epsilon \to 0+} \phi'(-\epsilon) = b + s \\
d_4 &= \lim_{\epsilon \to 0+} \phi'(\epsilon) = b - s
\end{aligned}$$

Lemma 3.4 *If $h < 0$, in which case $c_L = 0$ and $c_R = -h$, then*

$$\begin{aligned}
d_1 &= \lim_{\epsilon \to 0+} \phi'(-\epsilon) = b + s + \ell \\
d_2 &= \lim_{\epsilon \to 0+} \phi'(\epsilon) = b - s + \ell \\
d_3 &= \lim_{\epsilon \to 0+} \phi'(-h-\epsilon) = 2ah - \frac{3c\sqrt{-h}}{2} + b - s + \ell \\
d_4 &= \lim_{\epsilon \to 0+} \phi'(-h+\epsilon) = d_3 - \ell
\end{aligned}$$

Proof. For convenience define $m_h := s\,\mathbb{I}[h \geq 0] + (\ell - s)\,\mathbb{I}[h < 0]$, where $\mathbb{I}$ is the indicator function. Hence m_h is either s or $\ell - s$ depending on $\operatorname{sgn}(h)$. Then

$$\phi_1'(u) = \begin{cases} s + \ell & u < c_L \\ m_h & u > c_L \ \ \& \ \ u < c_R \\ -s & u > c_R \end{cases}. \tag{32}$$

The full derivative $\phi'(u)$ is computed by adding $f'(u)$, computed in (28), to $\phi_1'(u)$, computed in (32). This gives the values reported in the lemmas. □

We are now prepared to determine the optimum of ϕ in closed form.

THEOREM 3.5 *The objective function ϕ defined in* (29) *is continuous, concave and bounded above with a unique global maximum. The derivative of ϕ is monotone decreasing and is continuous except possibly at the cutpoints* (31)*. If $h \neq 0$, then $d_1 \geq d_2 \geq d_3 \geq d_4$, and the following cases are exhaustive and determine the optimal point:*

conditions	optimal point
$d_1 < 0$	$u_*(b + s + \ell)$
$d_1 \geq 0$ and $d_2 \leq 0$	c_L
$d_2 > 0$ and $d_3 < 0$	$u_*(b + m_h)$
$d_3 \geq 0$ and $d_4 \leq 0$	c_R
all $d_i > 0$	$u_*(b - s)$

If $h = 0$, the cutpoints coincide at zero, $d_1 = d_3 \geq d_2 = d_4$, and the following cases are exhaustive and determine the optimal point:

conditions	optimal point
$d_1 < 0$	$u_*(b + s + \ell)$
$d_1 \geq 0$ and $d_4 \leq 0$	0
$d_4 > 0$	$u_*(b - s)$

Proof (sketch). The optimum of ϕ is located where $\phi'(u)$ changes sign from positive to negative. Since the values of $\{d_i\}$ collectively give the left-derivatives and right-derivatives at both cutpoints, we can easily determine whether the sign change occurs at a cutpoint. If so, that cutpoint is the optimal point. Otherwise, when the optimal point is not at a cutpoint, the optimal point may be found by applying the function u_* to $b + \phi_1'(u)$ where ϕ_1' is calculated as in (32), and where u_* was defined previously in (27). □

3.4. *Box constraints*

It is very well known that quadratic programming with arbitrary linear inequalities does not admit a closed-form solution. Hence there is little hope for a closed-form solution to (4) with general inequality constraints. Nevertheless, there is a restricted class of constrained problems that can still be solved. All of the single-asset subproblems

$$\phi(u) = -au^2 + bu - c|u|^{3/2} - s|u| + \ell \min(h + u, 0)$$

are concave maximization problems with unique global maxima. Consider maximizing $\phi(u)$ subject to $L \leq u \leq U$, which is called a *box constraint.* Let u^*_{free} be the unconstrained optimum of ϕ, given in closed form by Theorem 3.5. The location of the optimal point for the constrained problem is given by

$$u^*_{\text{constr}} = \min(U, \max(L, u^*_{\text{free}})). \tag{33}$$

Convergence of Algorithms 2.1 and 2.2 is preserved because a box constraint corresponds to adding a convex indicator function to the separable term f_i in (16), a case covered by Tseng (2001).

4. Conclusions

Typical approaches to portfolio optimization with realistic transaction costs rely on relatively complicated general-purpose numerical optimization software. We give an algorithm, inspired by the Gauss–Seidel method, in which mean-variance optimization with square root impact, spread cost, and borrow costs can be solved in *semi-closed form*. The method repeatedly solves single-asset subproblems, where each subproblem admits a fully closed-form solution as an algebraic expression. This remains true even after including box constraints. When a factor structure is present, our technique takes full advantage of the reduction of dimensionality afforded by the factor model. Indeed, as we see in (19), when a factor model is present, one never needs to compute the full $n \times n$ covariance matrix and the inner loop is $O(k)$ where $k \ll n$ is the number of factors. Finally, the method immediately generalizes to a certain class of multi-horizon problems.

Disclosure of interest

There are no interests to declare.

Funding

No funding was received.

References

Almgren, R., Thum, C., Hauptmann, E. and Li, H., Direct estimation of equity market impact. *Risk*, 2005, **18**, 58–62.

Baldacci, B., Benveniste, J. and Ritter, G., Optimal turnover, liquidity, and autocorrelation. *Risk.net and arXiv:2110.03810*, 2022.

Beck, A., *First-order methods in optimization*, 2017 (SIAM: Philadelphia, PA).

Bellman, R.E., *Dynamic Programming*, 1957 (Princeton University Press: Princeton, NJ).

Benveniste, J., Kolm, P.N. and Ritter, G., Untangling universality and dispelling myths in mean-variance optimization. *The Journal of Portfolio Management*, 2024, **50**, 90–116.

Bucci, F., Benzaquen, M., Lillo, F. and Bouchaud, J.P., Crossover from linear to square-root market impact. *Physical Review Letters*, 2019, **122**, 108302.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. *et al.*, Pathwise coordinate optimization. *The Annals of Applied Statistics*, 2007, **1**, 302–332.

Gatheral, J., Three models of market impact. *Market Microstructure and High Frequency Data*, 2016.

Grinold, R.C. and Kahn, R.N., *Active portfolio management*, 2000 (McGraw-Hill: New York, NY).

Kolm, P.N. and Ritter, G., Multiperiod portfolio selection and Bayesian dynamic models. *Risk*, 2014, **28**, 50–54.

Kyle, A.S. and Obizhaeva, A.A., Market microstructure invariance: Empirical hypotheses. *Econometrica*, 2016, **84**, 1345–1404.

Kyle, A.S. and Obizhaeva, A.A., The market impact puzzle. *Working paper*, 2018.

Kyle, A.S., Obizhaeva, A.A. and Kritzman, M., A practitioner's guide to market microstructure invariance. *Journal of Portfolio Management*, 2016, **43**, 43–53.

Loeb, T.F., Trading cost: the critical link between investment information and results. *Financial Analysts Journal*, 1983, **39**, 39–44.

Menchero, J., Morozov, A. and Shepard, P., The Barra Global Equity Model (GEM2). *MSCI Barra Research Notes*, 2008, p. 53.

Roll, R. and Ross, S.A., An empirical investigation of the arbitrage pricing theory. *The Journal of Finance*, 1980, **35**, 1073–1103.

Ross, S.A., The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 1976, **13**, 341–360.

Sato, Y. and Kanazawa, K., Does the square-root price impact law belong to the strict universal scalings? Quantitative support by a complete survey of the Tokyo Stock Exchange market. *arXiv preprint arXiv:2411.13965*, 2024.

Torre, N., Barra market impact model handbook. *BARRA Inc., Berkeley*, 1997.

Tóth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J. and Bouchaud, J.P., Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 2011, **1**, 021006.

Tseng, P., *Coordinate ascent for maximizing nondifferentiable concave functions*, 1988 (Massachusetts Institute of Technology, Laboratory for Information and Decision Systems: Cambridge, MA).

Tseng, P., Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 2001, **109**, 475–494.

Xie, T., Liu, B., Xu, Y., Ghavamzadeh, M., Chow, Y., Lyu, D. and Yoon, D., A block coordinate ascent algorithm for mean-variance optimization. *Advances in Neural Information Processing Systems*, 2018, **31**.

Yen, Y.M. and Yen, T.J., Solving norm constrained portfolio optimization via coordinate-wise descent algorithms. *Computational Statistics & Data Analysis*, 2014, **76**, 737–759.

Zarinelli, E., Treccani, M., Farmer, J.D. and Lillo, F., Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate. *Market Microstructure and Liquidity*, 2015, **1**, 1550004.