Flow Trading

Eric Budish
University of Chicago

Peter Cramton
University of Cologne

Albert S. Kyle
University of Maryland

Mina Lee
Washington Univ. St. Louis

David Malec
University of Maryland

Annual Conference in Dynamic Structural Econometrics:
Market Design
Massachusetts Institute of Technology
August 19-20, 2022

Motivation for a New Market Design

Description of current exchanges:

- Use mostly standard limit orders—a price, quantity, and direction: "Buy 1000 shares of AAPL at $150.00 per share or better"
- Orders are typically for an individual asset rather than portfolios
- Orders are processed one-at-a-time continuously, with incoming "executable" orders matched with "resting" orders in limit order book
- Displayed bids and offers respect the minimum tick size of one cent and quantities respect the minimum lot size of one hundred shares

Motivation for a New Market Design

Current stock market design makes it costly for investors to implement trading strategies:

- Orders subject to immediate execution risk being picked off by high frequency traders when new information changes prices
- Discrete minimum tick size (one cent) induces queuing and race for time priority. Discrete minimum lot size widens bid-ask spread, encourages speed by enhancing value of time priority
- Institutional traders, who want to spread out their trade over time, must place and cancel thousands of small orders, requiring large resources
- Arbitrage trades between assets (pairs trades) or trading portfolios in general requires placing and canceling thousands of orders as prices change

Flow Trading: Combination of Three Ideas

- Flow Orders: Piecewise-linear, downward-sloping demand curves, continuous in price and quantity, with quantity expressed as “flows”
  - Buy a maximum of 1 share per second until 1000 shares are bought
  - Instead of “Buy 1000 shares right now”
- Frequent Batch Auctions: markets are cleared in discrete-time batch auctions, held at intervals such as once per second
  - Relative to status quo: time is discrete instead of continuous, and prices and quantities are continuous instead of discrete
- Orders for Portfolios: a portfolio is a user-defined linear combination of assets (arbitrary vector)
  - Both positive and negative weights allowed: buying and selling
  - Complements: same sign
  - Substitutes: opposite signs

Flow Trading vs. Traditional Exchange

Flow Trading	Traditional Exchange
Downward-sloping piecewise-linear supply and demand curves for flows	Discontinuous step functions for discrete quantities
Batch auctions once per second	Sequential matching one at a time
Orders for portfolios (linear combinations)	Orders for one asset

Table 1: Comparison of Flow Trading with Traditional Exchange

Benefits

Flow trading has four types of benefits relative to status quo:

- Investors can directly express many common trading demands:
  - Time-weighted average price (TWAP)
  - User-defined portfolios (customized ETFs)
  - Pairs trades: Buy A, Sell B
- Reduces the importance of speed
  - Reduces risk of resting limit orders being picked off
  - No race for queue position
- Easier to provide liquidity across related assets
  - Suppose A and B are highly correlated: can directly provide liquidity in Buy A, Sell B and Sell A, Buy B
  - Like a string that ties prices together. No correlation breakdown
- Transparency and fairness: all executable orders trade at the same price
  - Traders can verify appropriate execution from publicly announced market clearing prices exactly

Caveat: Flow trading is not designed to mitigate market failures related to market power or private information.

Market participants still must think strategically about how to trade on private information and manage their price impact, just as in the status quo.

Our Contribution

Combine the following well-understood concepts into a coherent and practical market design for trading stocks, bonds, futures contracts:

- Piecewise-linear downward sloping demand schedules
- Portfolios as linear combinations of assets
- General equilibrium theory
- Quadratic programming
- Batch auctions
- Reducing temporary price impact by trading slowly

Technical contributions:

- Existence and uniqueness results
- Computational proof-of-concept
- Microfoundations for portfolio orders

Literature

- Our market design combines orders for portfolios with ideas from Budish, Cramton, Shim (2015) and Kyle and Lee (2017)
- Flow orders are motivated by theoretical models (Vayanos (1999); Du and Zhu (2018); Kyle, Obizhaeva, Wang (2018)) as well as empirical evidence (popularity of TWAP and VWAP trading)
- Sophisticated expressions of preferences over multiple objects: Lahaie and Parkes (2004); Sandholm and Boutilier (2006); Milgrom (2009); Klemperer (2010); Cramton (2017); Budish et al (2017)
- Wittwer (2021) and Rostek and Yoon (2020a,b) discuss welfare implications of clearing assets jointly versus separately.
- Growing literature on financial market design: Duffie and Zhu (2017), Zhang (2020), Chen and Duffie (2020), Duffie and Dworczak (2021), Budish Lee and Shim (2021), many others
- Indivisible goods and existence of competitive eqm: Kelso and Crawford (1982), Hatfield and Milgrom (2005), Hatfield et al (2013, 2019), Baldwin and Klemperer (2019)

How Orders Work

There are N assets indexed n, I orders indexed i.
An order is specified by a tuple $(\boldsymbol{w}_i, p_i^L, p_i^H, q_i, Q_i^{\max})$

- Description of portfolio: vector of portfolio weights $\boldsymbol{w}_i \in \mathbb{R}^N$:
  - Individual asset: One nonzero weight to buy (+) or sell (-) one asset
  - Substitutes: One positive weight, one negative weight for pairs trade
  - Complements: 500 positive index weights to buy the S&P 500
  - Market making: pair of orders with weights $\boldsymbol{w}_i$ and $-\boldsymbol{w}_i$
- Two limit prices for the portfolio ($p_i^L = \$50.30$ and $p_i^H = \$50.40$ per share)
  - Negative portfolio weight (-1 share) and negative portfolio limit prices $p_i^L = -\$50.40$ and $p_i^H = -\$50.30$ for sell order. $p_i^L < p_i^H$ with both in $\mathbb{R}$
- Maximum execution rate ($q_i = 1.00$ portfolio unit per second)
- Cumulative quantity to be executed ($Q_i^{\max} = 10\,000$ portfolio units)

How Orders Work

Order executes

- At maximum rate (fully executable) q_i if price weakly below p_i^L
- At zero rate (nonexecutable) if price weakly above p_i^H
- At linearly interpolated rate (partially executable) if price in $[p_i^L, p_i^H]$

Buying vs. Selling

- “Selling” an asset is buying a portfolio with negative weight on that asset at a negative price.
- Whether buying or selling, always $p_i^L < p_i^H$.
- Think of p_i^L as analogous to the limit price in a limit order, whether positive or negative: “trade my full quantity at any price weakly better than p_i^L”.
- Think of p_i^H as expressing the price at which indifferent between trading and not: “do not trade if the price is p_i^H or worse.”
- For portfolios with positive and negative weights, there may not be a natural buying versus selling direction to the order.

Math: One Portfolio Order

Flow order i is described by the tuple $(\boldsymbol{w}_i, p_i^L, p_i^H, q_i, Q_i^{\max})$

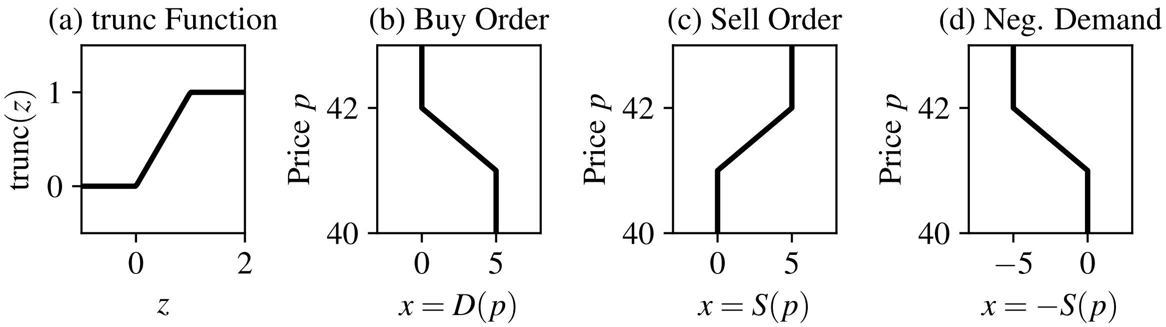
Let $\boldsymbol{\pi} = (\pi_1, \ldots \pi_N)$ denote a vector of N market-clearing asset prices. The price of the portfolio is the weighted sum of asset prices:

$$p_i = \boldsymbol{\pi}^{\mathsf{T}} \boldsymbol{w}_i \tag{1}$$

Assume the order's cumulative purchased quantity is not within q_i of $Q_i^{\max}$. The execution rate x_i of order i is given by:

$$x_i = D^i(p_i) = q_i \cdot \text{trunc}\left(\frac{p_i^H - p_i}{p_i^H - p_i^L}\right), \qquad \text{where} \qquad \text{trunc}(x) := \begin{cases} 1, & \text{for } x \geq 1 \\ x, & \text{for } 0 < x < 1 \\ 0, & \text{for } x \leq 0 \end{cases} \tag{2}$$

Illustration of Buying and Selling



For an order to sell 5 shares of a single asset n between \$41 and \$42:

- $\boldsymbol{w}_i$ is sparse vector with $w_{in} = -1$ share, zero otherwise
- $p_i^L = -42.00 < p_i^H = -41.00$ dollars per share
- $q_i = 5$ shares

Flexible, Limited Language for Preference Expression

Flexibility: Assets can be substitutes or complements (shoe analogy):

- Buy or sell a left shoe or right shoe separately
- Substitutes: Swap a left shoe for a right shoe (or vice versa)
- Complements: Buy left shoe and right shoe together
- Urgency expressed by maximum execution rate
- Arbitrary continuous downward-sloping portfolio demand function can be approximated with piecewise-linear orders

Some key financial use cases:

- Standard limit order (set $q_i = Q_i^{\max}$ and $p_i^H \to p_i^{L+}$)
- Time-weighted average price (TWAP) order (p_i^L sufficiently aggressive). Our analog of a market order
- Pairs trades ($(+, -, 0, \ldots, 0)$)
- Portfolio trades ($(+, +, \ldots)$ or $(-, -, \ldots)$)
- General long-short strategies
- Market making strategies (two orders with weights $\boldsymbol{w}_i$, $-\boldsymbol{w}_i$)

Flexible, Limited Language for Preference Expression

Limitations

- Trading demands are only defined at exactly the ratio of portfolio weights specified in the order. (Contrast to consumer theory)
- Trading demands are linear within each order.
- Language does not allow for indivisibilities (Ex: at least 100 shares, or none).
- Demand depends only on portfolio prices. Can't condition demand for asset A on realized price of asset B: "buy whichever of Left Shoe or Right Shoe gives me more surplus given my reported valuations". (Contrast: Klemperer (2010) or Milgrom (2009)).
- Relatedly, no in-order contingencies ("buy A if the price of B is high enough")

Math: Market Clearing

Market clears in N assets

At price vector $\boldsymbol{\pi} \in \mathbb{R}^N$, the exchange converts each order i's demand for portfolio units to demand for underlying assets by multiplying by its portfolio weights $\boldsymbol{w}_i$. Summing over the I orders yields the excess demand vector:

$$\text{Excess Demand Vector} = D(\boldsymbol{\pi}) := \sum_{i=1}^{I} D^i \left(\boldsymbol{\pi}^\mathsf{T} \boldsymbol{w}_i\right) \cdot \boldsymbol{w}_i \tag{3}$$

The exchange seeks to find a market clearing price vector

$$D(\boldsymbol{\pi}) = \mathbf{0}, \qquad (N \text{ equations in } N \text{ unknowns}) \tag{4}$$

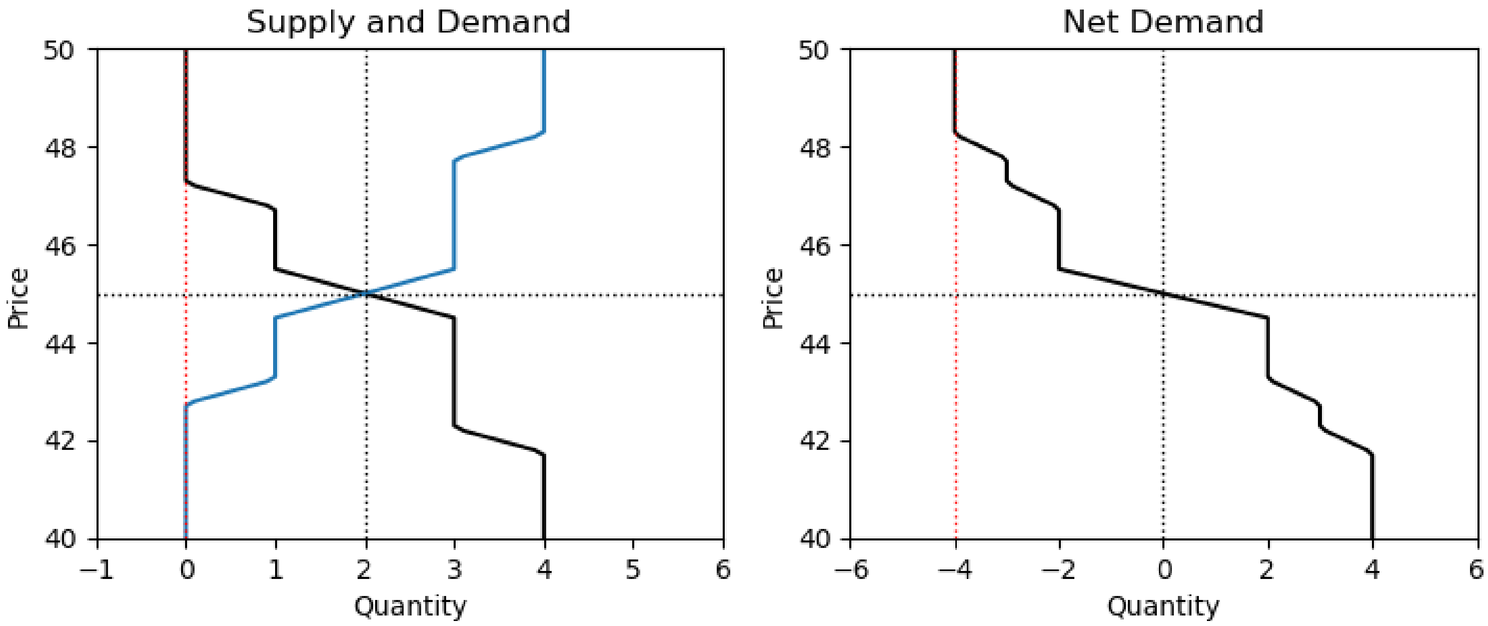
in which case each order executes at rate

$$x_i = D^i(\boldsymbol{\pi}^\mathsf{T} \boldsymbol{w}_i), \qquad (\text{scalar equation in portfolio units } x_i) \tag{5}$$

Illustration of Market Clearing

One asset, six orders (symmetric about $45.00 for buying and selling)

- One fully executable buy order and one fully executable sell order
- One non-executable buy order and one non-executable sell order
- One partially executable buy order and one partially executable sell order



Existence and Uniqueness

Questions:

- Do equilibrium prices and quantities exist?
- If they exist, are they unique?

Idea for proof:

- Treat orders as expressions of preferences, implying quasi-linear (dollar) quadratic utility for order i for quantities in range $[0, q_i]$
  - Note: order i's preferences are only defined exactly on the $\boldsymbol{w}_i$ vector
  - Note: order i's preferences are satiated at q_i
- Exchange solves the problem of maximizing sum of dollar utility across orders subject to market clearing constraint
- Interpret Lagrange multipliers for market clearing constraint as price vector $\boldsymbol{\pi}$

Setting up the Optimization Problem

Infer quadratic utility from “as-bid” linear portion of demand schedule

$$V_i(x) = p_i^H x - \frac{p_i^H - p_i^L}{2q_i} x^2 \tag{6}$$

Exchange solves the problem of finding quantities $\boldsymbol{x} = (x_1, \ldots, x_I)$ to solve

$$\max_{\boldsymbol{x}} \sum_{i=1}^{I} V_i\left(x_i\right) \qquad \text{subject to} \begin{cases} \sum_{i=0}^{I} \; x_i \, \boldsymbol{w}_i = \boldsymbol{0} & \text{(market clearing)} \\ 0 \leq x_i \leq q_i \text{ for all } i & \text{(order execution rate),} \end{cases} \tag{7}$$

This is a quadratic optimization problem with:

- N linear equality constraints enforcing market clearing of N assets,
- $2I$ linear inequality constraints enforcing no overfilling or underfilling of I orders

In matrix form: Let $\boldsymbol{D}$ denote the $I \times I$ positive definite diagonal matrix whose ith diagonal element is $(p_i^H - p_i^L)/q_i$.

$$\max_{\boldsymbol{x}} \left[\boldsymbol{x}^\mathsf{T}\boldsymbol{p}^H - \tfrac{1}{2}\,\boldsymbol{x}^\mathsf{T}\boldsymbol{D}\,\boldsymbol{x}\right] \qquad \text{subject to} \qquad \boldsymbol{W}\,\boldsymbol{x} = \boldsymbol{0} \qquad \text{and} \qquad \boldsymbol{0} \leq \boldsymbol{x} \leq \boldsymbol{q}. \tag{8}$$

Theorem: Existence and Uniqueness of Quantities

Theorem 1 (Existence and Uniqueness of Optimal Quantities)**.** *There exists a unique quantity vector* $\boldsymbol{x}^*$ *which solves the maximization problem* (7)

Proof.

- Compactness: Inequality constraints on quantities. Market clearing conditions are linear constraints, which defines the intersection of hyperplanes. Thus the set of vectors of trade rates $\boldsymbol{x}$ that satisfies all constraints is compact and convex.
- Feasibility: No trade ($\boldsymbol{x} = \mathbf{0}$) is feasible: satisfies constraints with finite value of objective.
- Strictly Concave Objective Function: Each function V_i is quadratic and therefore strictly concave. Since V is the sum of V_i across i, the function V is concave as well (on $\mathbb{R}^I$ and on the compact and convex subset defined by the constraints.)
- Well-known principle of convex analysis that a strictly concave objective function on a non-empty compact and convex set has a unique maximizing vector $\boldsymbol{x}^*$ (References: Boyd and Vandenberghe (2004); Bertsekas (2009)).

□

Why is Existence so Simple?

The language allows users to define arbitrary portfolios, including complements and substitutes.

In GE and Indivisible Goods literatures, complements especially make existence hard. (Arrow-Debreu-McKenzie, Starr 1969, Hatfield-Kominers-Westkamp 2021, Baldwin-Klemperer 2019)

Yet here the existence proof is simple. Why?

- Goods are infinitely divisible.
- Portfolio demand schedules are downward sloping.
- Utility for each order is defined only on the line segment associated with the portfolio weights (not defined off diagonal).
- Quantities are bounded. Zero trade is feasible.
- No in-order contingencies or linkages across multiple orders. (This limits the comps and subs)

A sweet spot? Expressive enough to be useful, and existence is guaranteed

Dual Problem: Prices

Define Lagrangian

$$L(\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := V(\boldsymbol{x}) - \sum_{i=1}^{N} (x_i \cdot \boldsymbol{w}_i)^{\mathsf{T}} \boldsymbol{\pi} + \boldsymbol{x}^{\mathsf{T}} \boldsymbol{\mu} + (\boldsymbol{q} - \boldsymbol{x})^{\mathsf{T}} \boldsymbol{\lambda} \tag{9}$$

- Prices $\boldsymbol{\pi}$ are positive or negative Lagrange multipliers enforcing market clearing
- $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are Lagrange multipliers enforcing order execution rate constraints (“Taxes” and “subsidies”)

The dual objective associated with the primal problem of solving for optimal quantities is

$$\hat{G}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \max_{\boldsymbol{x}} \; L(\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{for} \qquad \boldsymbol{\pi} \in \mathbb{R}^N, \qquad \boldsymbol{\mu} \geq \mathbf{0}, \qquad \boldsymbol{\lambda} \geq \mathbf{0} \tag{10}$$

The dual problem is

$$g^* := \inf_{\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\mu}} \hat{G}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{subject to} \qquad \boldsymbol{\pi} \in \mathbb{R}^N, \qquad \boldsymbol{\mu} \geq \mathbf{0}, \qquad \boldsymbol{\lambda} \geq \mathbf{0} \tag{11}$$

Result: Existence of Market Clearing Prices

Theorem 2 (Existence of Market Clearing Prices)**.** *There exists at least one optimal solution* $(\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *to the dual problem* (11)*. The solutions* $\boldsymbol{x}^*$ *and* $(\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *are a primal-dual pair which satisfies the strict duality relationship*

$$g^* = V(\boldsymbol{x}^*). \tag{12}$$

Proof. The quadratic program has these properties:

- Concavity: The objective function $V(\boldsymbol{x})$ is concave
- Finite solution: Sum of concave objectives bounded from above
- Feasibility: No trade ($\boldsymbol{x} = \mathbf{0}$) is feasible: clears markets, satisfies order execution rates
- Linear constraints: Market clearing and order execution rates

Standard result from convex programming that these conditions guarantee that a solution to the dual problem exists and has the same value as the supremum to the primal (Bertsekas 2015, Proposition 5.3.4). □

The set of market clearing prices is convex, but may be unbounded

Computation

Question: prices and quantities exist. Can we compute them?

- Many economic settings where prices are known to exist but hard to find (Scarf and Hansen, 1973)
- Many economic settings where prices are trivial to compute—one asset version of our problem is an easy example!
- Our problem lies in between
- Plan
  - Show that gradient method works. “Easier than Scarf’s problem”
  - Result: gradient method convergence slow (confirmed in simulations)
  - Add “exchange as market maker” which enables interior point methods
  - Result: faster in theory, and also in simulations.
  - Goal in mind: solve large problems in less than one second.

Gains Function

Theorem 3 (Gains function)**.** *Define the gains function as*

$$G(\boldsymbol{\pi}) := \min_{\boldsymbol{\lambda},\boldsymbol{\mu}} \hat{G}(\boldsymbol{\pi},\boldsymbol{\lambda},\boldsymbol{\mu}) \qquad \textit{subject to} \qquad \boldsymbol{\lambda} \geq \mathbf{0}, \qquad \boldsymbol{\mu} \geq \mathbf{0}. \tag{13}$$

Every market clearing price vector $\boldsymbol{\pi}^*$ *satisfies*

$$\boldsymbol{\pi}^* = \arg\min_{\boldsymbol{\pi}\in\mathbb{R}^N} G(\boldsymbol{\pi}). \tag{14}$$

The set of market clearing prices is a nonempty, closed convex set which may be unbounded.

- Economic interpretation: The gains function minimizes the sum of "consumer surplus" across orders. Analogous to minimizing expenditure function to maximize utility
- Intuition: optimizing against non-market clearing prices allows market participants to achieve greater surplus than trading at market clearing prices, but trading at nonclearing prices costs exchange more
- Derivative of the gains function is minus the market demand function
- Second derivative of the gains function is a negative semi-definite matrix

Computation: Tatonnement and Gradient Method

Economist intuition: Use Walrasian tatonnement on dual problem (prices):

- Tatonnement is equivalent to gradient method of optimization on gains function since gradient search direction is negative of derivative, which is excess demand search direction of tatonnement.
- Since gains function is variation on dual problem, we know clearing prices exist.
- Bad news: Although convergence can be guaranteed, with a bound on convergence rate, convergence is too slow for solving problem in one second.

Theorem 4. *Let G be a convex with continuously differentiable gradient satisfying a Lipschitz condition with constant L. Using step size $1/L$, the the error after k iterations of the gradient method $G(\boldsymbol{\pi}_k) - G(\boldsymbol{\pi}^*)$ is related to the error of the initial guess $\boldsymbol{\pi}_0 - \boldsymbol{\pi}^*$ by*

$$G(\boldsymbol{\pi}_k) - G(\boldsymbol{\pi}^*) \leq \frac{2L\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}^*\|^2}{k+4}. \tag{15}$$

(Nesterov 2004, Corollary 2.1.2)

Exchange as Market Maker

We now add the exchange as a small market maker:

- Generates good tie-breaker rule when clearing prices nonunique and even unbounded.
- Allows interior point method to be used to calculate prices by guaranteeing that initial allocations $\alpha \boldsymbol{x}$, $0 < \alpha < 1$ are on interior of feasible set, since exchange takes other side of otherwise uncleared quantities.
  - Natural starting point of no trade is not an interior point.

Exchange places small linear demand for each asset:

$$y_n = \epsilon_n(\pi_{0n} - \pi_n), \tag{16}$$

where $\epsilon_n > 0$ is small, π_{0n} is exchange's best guess at clearing price (e.g., last price or from some initial computation)

In vector notation, demand function and utility function of exchange are

$$\boldsymbol{y} = \boldsymbol{\epsilon}(\boldsymbol{\pi}_0 - \boldsymbol{\pi}), \qquad \boldsymbol{y}^{\mathsf{T}}\boldsymbol{\pi}_0 - \tfrac{1}{2}\,\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\epsilon}^{-1} \tag{17}$$

How Interior Point Method Works

Adding exchange as a trader, modified primal problem is

$$\max_{\boldsymbol{x},\boldsymbol{y}} \left[\boldsymbol{x}^{\mathsf{T}}\boldsymbol{p}^{H} - \tfrac{1}{2}\,\boldsymbol{x}^{\mathsf{T}}\boldsymbol{D}\,\boldsymbol{x} + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{\pi}_0 - \tfrac{1}{2}\,\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\epsilon}^{-1}\,\boldsymbol{y}\right] \quad \text{subject to} \quad \boldsymbol{W}\boldsymbol{x}+\boldsymbol{y} = \mathbf{0}, \quad \mathbf{0} \leq \boldsymbol{x} \leq \boldsymbol{q}. \tag{18}$$

Interior point method changes problem by replacing inequality constriants with penalty functions:

$$\max_{\boldsymbol{x},\boldsymbol{y}} \left[\boldsymbol{x}^{\mathsf{T}}\boldsymbol{p}^{H} - \tfrac{1}{2}\,\boldsymbol{x}^{\mathsf{T}}\boldsymbol{D}\,\boldsymbol{x} + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{\pi}_0 - \tfrac{1}{2}\,\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\epsilon}^{-1}\,\boldsymbol{y} + \bar{\nu}\log(\boldsymbol{x})^{\mathsf{T}}\mathbf{1} + \bar{\nu}\log(\boldsymbol{q} - \boldsymbol{x})^{\mathsf{T}}\mathbf{1}\right], \quad \boldsymbol{W}\boldsymbol{x}+\boldsymbol{y} = \mathbf{0}. \tag{19}$$

Solution of modified problem with inequality constraints converges to solution of modified problem without inequality constraint in the limit $\bar{\nu} \longrightarrow 0$.

Complexity of Interior Point Method

Nice accessible discussion by Gondzio(2012)

Let ϵ denote size of required error (percentage reduction)

- Complexity of interior point method: $O(\log(1/\epsilon))$
- Complexity of gradient method: $O(1/\epsilon)$ or $O(1/\epsilon^2)$ or $O(\log(1/\epsilon))$

Characterization: Karush–Kuhn–Tucker Conditions

Theorem 5 (Necessary and Sufficient Conditions)**.** *The vector of quantities* $\boldsymbol{x}^*$ *is the unique primal solution and a vector of multipliers* $(\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *is a dual solution if and only if the following conditions hold:*

$$\sum_{i=0}^{I} x_i \, \boldsymbol{w}_i = \mathbf{0}, \quad \mathbf{0} \le \boldsymbol{x} \le \boldsymbol{q}, \qquad \textit{(Primal Feasibility)}, \tag{20}$$

$$\boldsymbol{\pi} \in \mathbb{R}^N, \qquad \boldsymbol{\lambda} \ge \mathbf{0}, \qquad \boldsymbol{\mu} \ge \mathbf{0}, \qquad \textit{(Dual Feasibility)} \tag{21}$$

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathbb{R}^I}{argmax} \;\; L(\boldsymbol{x}, \boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*), \qquad \textit{(Primal Optimality)} \tag{22}$$

$$\boldsymbol{\lambda}^* \cdot (\boldsymbol{q} - \boldsymbol{x}^*) = \mathbf{0}, \qquad \boldsymbol{\mu}^* \cdot \boldsymbol{x}^* = \mathbf{0} \qquad \textit{(Complementary Slackness)} \tag{23}$$

In the above theorem, maximizing the Lagrangian can be replaced by the first-order conditions for $\boldsymbol{x}$

Interior point methods and KKT conditions

KKT conditions with interior point method:

- Add the exchange's demand to the first-order condition.
- Add the exchange's quantities traded to the market-clearing condition.
- Replace the complementary slackness condition $\boldsymbol{\mu}^* \cdot \boldsymbol{x}^* = \mathbf{0}$ with $\boldsymbol{\mu}^* \cdot \boldsymbol{x}^* = \bar{\nu} \cdot \mathbf{1}$, and then let $\bar{\nu} \to 0$.

The modified KKT conditions are

$$\boldsymbol{W}\ \boldsymbol{x}^* + \boldsymbol{y}^* = \mathbf{0}, \qquad \mathbf{0} \leq \boldsymbol{x}^* \leq \boldsymbol{q}, \qquad \boldsymbol{y}^* \in \mathbb{R}^N \qquad \text{(Primal Feasibility)}, \tag{24}$$

$$\boldsymbol{\pi}^* \in \mathbb{R}^N, \qquad \boldsymbol{\lambda}^* > \mathbf{0} \qquad \boldsymbol{\mu}^* > \mathbf{0}, \qquad \text{(Dual Feasibility)} \tag{25}$$

$$\boldsymbol{p}^H - \boldsymbol{D}\boldsymbol{x}^* - \boldsymbol{\epsilon}^{-1}\boldsymbol{y}^* - \boldsymbol{W}^{\mathsf{T}}\boldsymbol{\pi}^* + \boldsymbol{\mu}^* - \boldsymbol{\lambda}^* = \mathbf{0} \qquad \text{(Primal Optimality)} \tag{26}$$

$$\boldsymbol{\lambda}^* \cdot (\boldsymbol{q} - \boldsymbol{x}^*) = \bar{\nu} \cdot \mathbf{1}, \qquad \boldsymbol{\mu}^* \cdot \boldsymbol{x}^* = \bar{\nu} \cdot \mathbf{1}, \qquad \bar{\nu} > 0, \qquad \bar{\nu} \to 0 \qquad \text{(Complementary Slackness)} \tag{27}$$

Intuition: Solve revised problem for finite $\bar{\nu} > 0$ while at the same time pushing $\bar{\nu}$ closer and closer to zero, keeping updated guesses interior points. Result: quick convergence to the solution of our original KKT conditions

Our Simulations

Our own implementation of CVXOPT algorithm using Python. Strategy:

- Linearize KKT conditions ($3I + N$ equations).
- Solve linearized system with $\bar{\nu} = 0$.
- Take step which keeps guessed quantities $\boldsymbol{x}$ interior point, keeping $\bar{\nu} > 0$ but making it smaller.
- Solve equations by expressing multipliers as functions of $\boldsymbol{x}$, express $\boldsymbol{x}$ as function of prices $\boldsymbol{\pi}$, then solve much smaller $N \times N$ system for $\boldsymbol{\pi}$ using Cholesky decomposition.
- Need new Cholesky decomposition at each iteration to incorporate new information from "deep in the order book."

Simulation Assumptions

Try to pose a difficult problem:

- 500 assets and 100,000 orders in base case scenario.
- Huge variation in liquidity across assets (Invariance assumptions).
- Correlated index orders (value-weighted and equally-weighted).
- Small (one basis point) difference between p_i^H and p_i^L on average (step function approximation)
- Exchange provides very little liquidity.

Simulation Results

- Panel A varies orders: $t \approx 1$ sec for 500 assets and 2,000,000 orders.
- Panel B varies assets: $t \approx 1$ sec for 2000 assets and 100,000 orders.

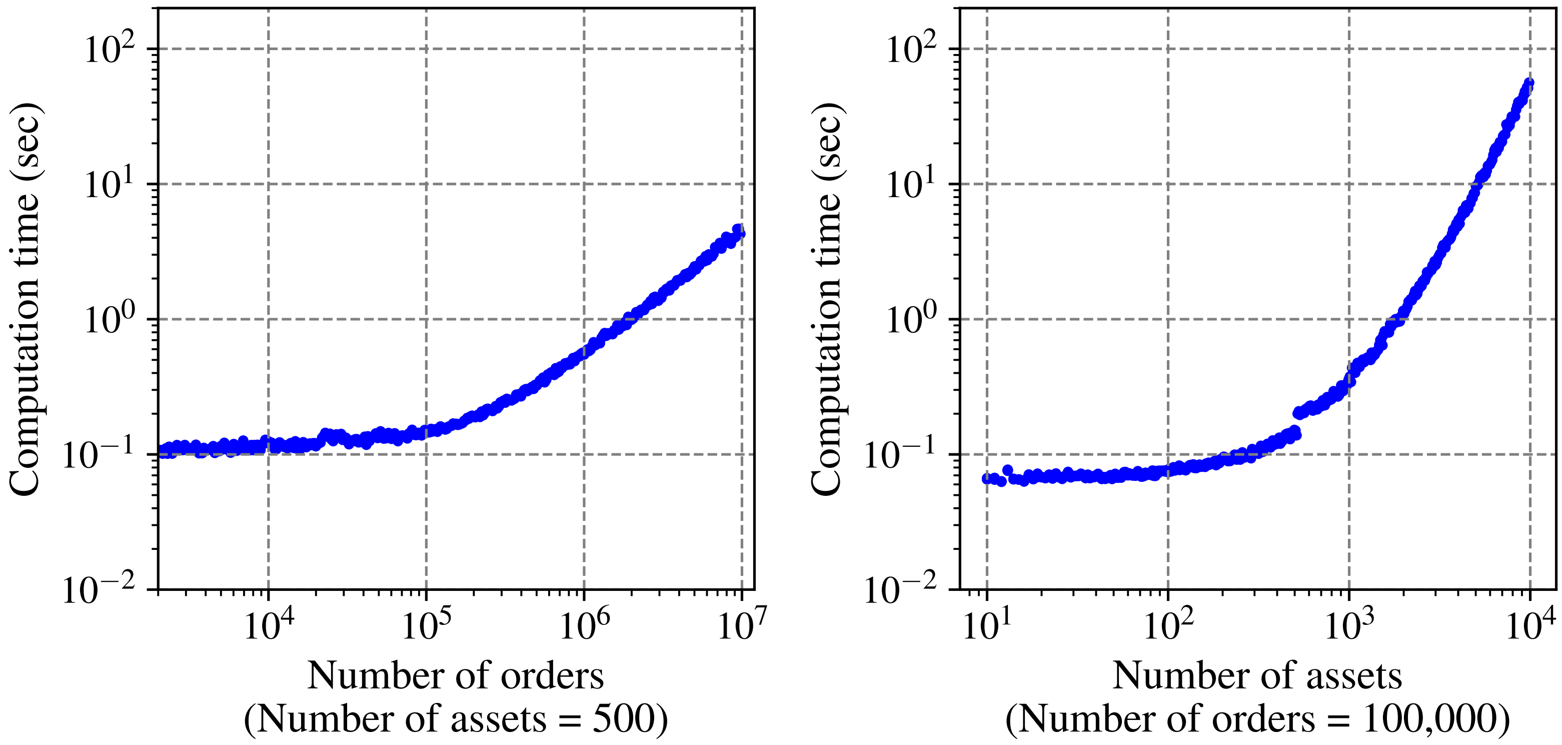


Figure 1: Execution time for 2×500 simulated market outcomes.

Importance of Exchange Trading

- In simulation results reported above, exchange share of trading volume is less than 10^{-7} when there are tens of thousands of orders.
- Exchange also picks up much smaller uncleared quantities due to imperfect algorithm convergence (numerical error).
- This modest exchange trading seems to make algorithm more numerically stable (as expected).
- Exchange trading also provides a tie-breaker rule, keeping non-unique prices near exchange's no-trade price

Microfoundation for Portfolio Orders

Portfolio orders are more general than traditional limit orders but the language is still restrictive. In general, asset demands should depend on the complete vector of asset prices, not prices of user-defined portfolios. We provide a modest microfoundation for our approach.

- Assume trader has CARA (exponential) utility with risk aversion A.
  - Other utility functions might generate wealth effects which make demand slope upward, not downward (margin calls).
- Assume trader has subjective beliefs that liquidation values $\boldsymbol{v}$ are jointly normally distributed with subjective mean $\boldsymbol{m}$ and subjective variance $\boldsymbol{\Sigma}$.

Standard finance theory implies demand is

$$\boldsymbol{\omega}^* = (A\boldsymbol{\Sigma})^{-1}(\boldsymbol{m} - \boldsymbol{\pi}). \tag{28}$$

Quantity of asset n, given by ω_n is a seemingly arbitrary linear function of

entire price vector $\boldsymbol{\pi}$, not just price of asset n itself.

Implementing the Optimum with Portfolio Orders

If the prices $\boldsymbol{\pi}$ are known and fixed, the trader can implement their optimum with a single portfolio order.

- Set weights $\boldsymbol{w}_i$ and quantity $Q_i^{\max}$ such that $\boldsymbol{w}_i Q_i^{\max} = \boldsymbol{\omega}^*$
- Set $p_i^L > \boldsymbol{\pi}\boldsymbol{w}_i$ so that the order is fully executable at the portfolio price

What if the trader does not know the asset prices? This might capture that prices are changing over time and traders trade gradually.

We will show that the trader can implement their optimum without any knowledge of prices, using a set of portfolio orders.

Singular Value Decomposition

Since the covariance matrix $\boldsymbol{\Sigma}$ is positive semidefinite, its SVD has the form

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^{\mathsf{T}} \tag{29}$$

where $\boldsymbol{U}$ is orthonormal and $\boldsymbol{\Delta}$ is diagonal with nonnegative elements. Let K denote the rank of $\boldsymbol{\Sigma}$. We can rewrite as:

$$\boldsymbol{\Sigma}^{-1} = \sum_{k=1}^{K} \frac{1}{\delta_k} \boldsymbol{u}_k \boldsymbol{u}_k{}^{\mathsf{T}}, \tag{30}$$

Using this, we can express the optimal portfolio $\boldsymbol{\omega}^*$ from above as:

$$\boldsymbol{\omega}^* = \sum_{k=1}^{K} \left(\frac{\boldsymbol{u}_k{}^{\mathsf{T}}\boldsymbol{m} - \boldsymbol{u}_k{}^{\mathsf{T}}\boldsymbol{\pi}}{A\ \delta_k} \right) \boldsymbol{u}_k. \tag{31}$$

This is a linear combination of downward sloping linear demands for portfolios $\boldsymbol{u}_k$ as functions of the portfolio price $\boldsymbol{u}_k{}^{\mathsf{T}}\boldsymbol{\pi}$.

Theorem

Theorem 6. Consider a static CARA-normal framework in which a trader believes that the variance-covariance matrix of the asset payoffs has rank K. Then the trader's optimal portfolio (equation (28)) can be represented as the sum of K downward-sloping demand schedules for portfolios, each of which depends only on that portfolio's price (equation (31)).

The same logic applies to the model of strategic trading if the price impact matrix, denoted by $\mathbf{\Lambda}$, is positive semidefinite: Let K' denote the rank of $A\mathbf{\Sigma} + \mathbf{\Lambda}$. Then a strategic trader's optimal portfolio can be represented as the sum of K' downward-sloping demand schedules for portfolios, each of which depends only on that portfolio's price.

Limitation: The logic does not extend to demand schedules with wealth effects or to demand schedules with learning from price because demand schedules may not be downward sloping.

Practical Implementation

- With some algebra, we can express the dollar value of expected utility at prices $\boldsymbol{\pi}$ as the sum of squared portfolio-Sharpe-ratios:

$$\sum_{i=1}^{K} \frac{1}{2A} \left(\frac{\boldsymbol{u}_i{}^{\mathsf{T}} \boldsymbol{m} - \boldsymbol{u}_i{}^{\mathsf{T}} \boldsymbol{\pi}}{\sqrt{\delta_i}} \right)^2, \tag{32}$$

- In practice, traders may select a few portfolios which they perceive to have sufficiently attractive Sharpe ratios, rather than submitting demands for all K portfolios to implement the theoretical optimum.

Policy Discussion

- Efficiency: Our proposal dramatically reduces market interface costs for users, market makers, and other intermediaries
  - Efficiency based on “as-bid” strategically expressed preferences rather than unknown true preferences
- Competition: Allows traders to focus on alpha models, market impact models, and risk models, not speed, bandwidth, and complexity of order handling systems
- Fairness: Levels the technological playing field
- Transparency: All orders receive the same prices at the same time. Executable TWAP orders automatically achieves TWAP price
- Trust: Proper order execution can be verified from history of market-clearing prices

Additional Issues

- Tie-breaking: If prices not unique, minimize distance to prior price
- Exchange as liquidity provider: If exchange places a linear order in each asset, prices are unique and computation is faster (geometric convergence based on eigenvalue ratio)
- Backup plan: If exchange cannot compute prices in one second, allow the exchange to trade small quantities to clear markets. The alternative is to ration orders, like “fast market conditions” suspending traders usual expectations of order execution quality
- Post-trade transparency: At a minimum, exchange publishes prices and market volume each second.
- Pre-trade transparency: A large trader can estimate temporary price impact by canceling order execution for one second, see how far the price moves. To avoid such price blips, the exchange might publish information about depth of book for some assets and portfolios

Conclusion (1)

- This paper has introduced a new market design for trading financial assets, such as stocks, bonds, futures, currencies
- Three elements: flow orders from Kyle and Lee (2017), frequent batch auctions from Budish, Cramton and Shim (2015), and a novel language for trading portfolios
- Technical foundations: existence and uniqueness results; computational results; microfoundations for portfolio orders
- Benefits of combining KL+BCS: a market design in which time is discrete and prices and quantities are continuous
  - Status quo market design has these reversed.
  - Continuous time and discrete prices/quantities are the cause of significant complexity, inefficiency, and rent seeking
  - Policy debates that relate: arms race for trading speed; complex order types; proprietary market data and access; internalization of retail investors' order flow

Conclusion (2)

- Benefits of novel language for portfolio orders:
  - Rich enough to directly express many important kinds of trading demands
  - While also allowing for guaranteed existence and fast computation
  - A "point on the frontier of language design" — tradeoff between expressiveness and computability
- Main open topic for future research: efficiency consequences of portfolio trading. We conjecture two main benefits:
  - Reduce costs and complexity: traders can directly express many important trading demands, reducing need for costly/complex intermediation
  - More efficient to link liquidity and price discovery across correlated assets. Ex: if A and B are perfect substitutes, prices can be perfectly correlated with Bertrand competition on cost of (Buy A, Sell B)
- Also important to study: strategic issues that arise under portfolio trading around bid-shading and managing price impact