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

Preliminary and Incomplete

ASSA 2021 Virtual Annual Meeting
January 5, 2021

Econometric Society Session
Market Power and Market Design

Disclaimer

- The views expressed are our own and not those of any organization we have been associated with
- Peter Cramton is an independent director of ERCOT, the Texas electricity system operator, and an academic advisor to CartaX, a private-equity marketplace
- Albert S. Kyle is an independent director of a U.S.-based asset management company which trades global equities

Motivation for a New Market Design

Description of current exchanges:

- Use mostly standard limit orders—a price, quantity, and direction: “Buy 1000 share of AAPL at $126.85 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
- 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 require placing and canceling thousands of orders as prices change
- 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

Flow Trading: Combination of Four Ideas

- Piecewise-linear, downward-sloping demand curves, continuous in price and quantity
  - Also reflects supply curves as demand curves for negative quantities
- Flow Orders: specify maximum rate of trade (one share per second), executed over time
- Frequent Batch Auctions: held at intervals such as once per second
- Orders for portfolios (linear combination of assets): Flow rate of trading depends on price of portfolio calculated from underlying asset prices

Benefits: A Language for Expressing Preferences

Flow trading implements chosen trading strategies easily

- Reduces the risk of resting limit orders being picked off
- Allows traders to trade gradually with just one order, mimicking Volume-Weighted-Average-Price (VWAP), achieving Time-Weighted-Average-Price (TWAP) exactly
- Makes it easier to implement pairs trades and portfolio orders, mimicking institutional basket trading and long-short relative value strategies
- Eliminates frictions associated with a minimum tick size, such as timing trades based on changes in bid and ask quantities

Flow trading is a language for implementing near-optimal strategies in the game traders are actually playing.
Flow trading is not designed to mitigate market failures related to market power or private information by placing restrictions on trading strategies.

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 are common issue in the market design literature: Lahaie and Parkes (2004); Sandhom and Boutilier (2006); Cramton (2017)
- Rostek and Yoon (2020a,b) discuss welfare implications of clearing assets jointly versus separately
- Growing literature on the financial market design: Duffie and Zhu (2017), Zhang (2020)

How Orders Work

An order specifies:

- Description of portfolio: List of securities plus list of asset weights describe a sparse vector of portfolio weights $\boldsymbol{w}^i$:
  - Individual asset: One nonzero weight to buy (positive) or sell (negative) one asset
  - Substitutes: One positive weight and one negative weight for a pairs trade
  - Complements: 500 positive index weights to buy the S&P 500
  - Market making orders buy and sell simultaneously
- Two limit prices for the portfolio ($p_H^i = \$50.40$ and $p_L^i = \$50.30$ per share)
  - Negative portfolio weight (-1 share) and negative portfolio limit prices $p_H^i = -\$50.30$ and $p_L^i = -\$50.40$ for sell order.
- Maximum execution rate ($q^i = 1.00$ portfolio unit per second)
- Cumulative quantity to be executed ($Q_{\text{max}}^i = 10\,000$ portfolio units)

How Orders Work

Order executes

- At zero rate (nonexecutable) if price above upper limit p_H^i
- At maximum rate (fully executable) q^i if price below lower limit p_L^i
- At linearly interpolated rate (partially executable) if price in $[p_L^i, p_H^i]$

Quantities in nano-shares, prices in micro-dollars: trade 0.123450000 shares during one second, at price $50.312345

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

Inflexibility: Orders cannot be arbitrary:

- Order quantity cannot depend on price of an asset not traded by the order (cannot buy a right shoe alone at a rate depending on price of a left shoe)
- Order cannot treat right shoes and left shoes as perfect substitutes

Math: One Portfolio Order

Let $\boldsymbol{p} = (p_1, \ldots p_N)$ denote vector N market-clearing asset prices Price of portfolio is weighted sum of asset prices:

$$p^i = \boldsymbol{p}^{\mathsf{T}} \boldsymbol{w}^i \tag{1}$$

Execution rate x given by

$$x^i = D^i(p^i) = q^i \cdot \text{trunc}\left(\frac{p_H^i - p^i}{p_H^i - p_L^i}\right), \qquad \text{where} \qquad \text{trunc}(x) := \begin{cases} 1, & \text{for } x \geq 1 \\ x, & \text{for } 0 < x < 1 \\ 0, & \text{for } x \leq 0 \end{cases} \tag{2}$$

Math: Market Clearing

Market clears in assets, not portfolios

The exchange converts portfolio units to underlying assets (multiplying by weights $\boldsymbol{w}^i$), calculates net excess demand vector by summing demands for assets across orders with price vector $\boldsymbol{p}$:

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

The exchange seeks to find a market clearing price vector

$$D(\boldsymbol{p}) = \mathbf{0} \tag{4}$$

in which case each order executes at rate

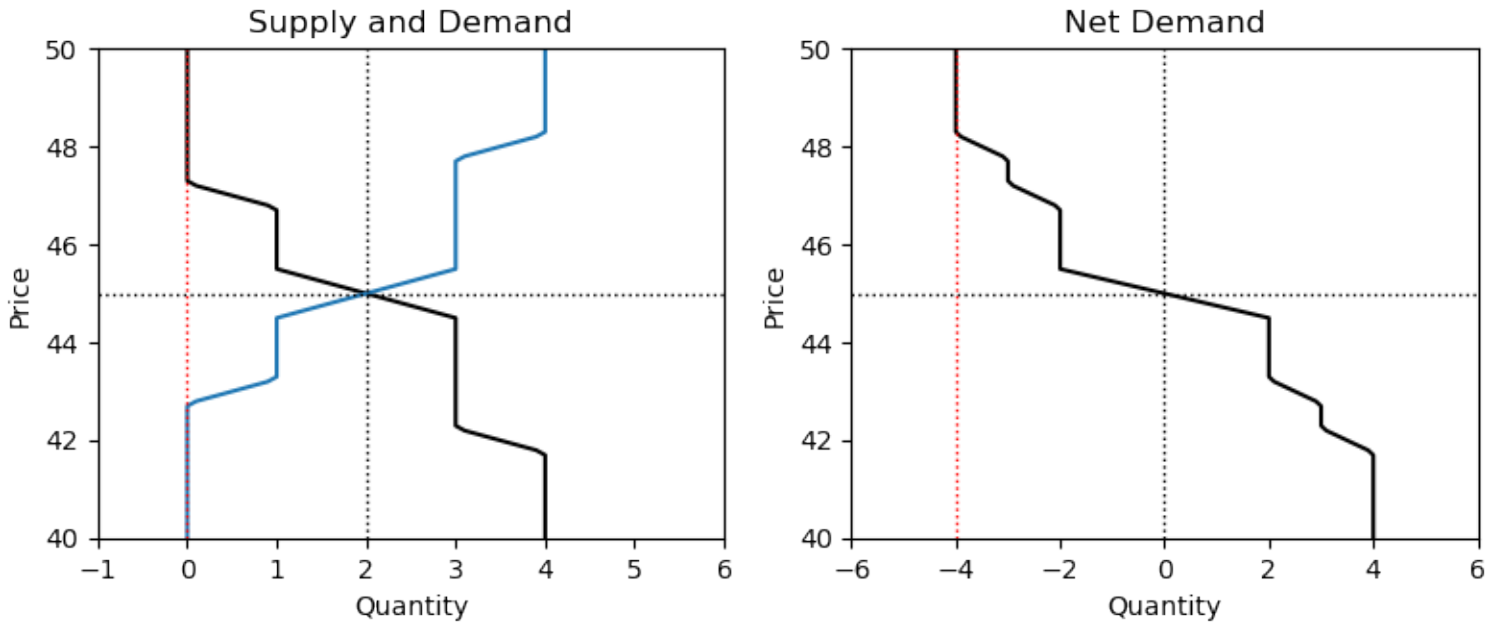
$$x^i = D^i(\boldsymbol{p}^\mathsf{T}\boldsymbol{w}^i) \tag{5}$$

With thousands of assets and hundreds of thousands of orders, the exchange faces a computational challenge of calculating market clearing prices in one second

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?
- Is it computationally feasible to calculate prices and quantities in one second?

Idea for proof: Mimic the economics of general equilibrium theory:

- Treat orders as expressions of preferences, implying quasi-linear (dollar) quadratic utility for quantities in range $[0, q^i]$
- 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 $\boldsymbol{p}$

Assumptions Related to Existence and Uniqueness

Standard, well-studied problem: quadratic objective, linear equality and inequality constraints

Existence of market-clearing quantities (primal problem):

- Continuous (or concave) objective function
- Compact space of quantities traded
- Feasibility of no-trade (allowed by limit order, market clears)

Existence of market-clearing prices (dual problem):

- Concave (quadratic) objective bounded from above
- Feasibility of no trade
- Linear equality constraints (market clearing) and inequality constraints (order execution rates)

Uniqueness of quantities:

- Strictly concave (quadratic) objective function

Prices may be non-unique, even unbounded, but belong to convex set

Implementation

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

$$V^i(x) = p_H^i x - \frac{p_H^i - p_L^i}{2q^i} x^2 \tag{6}$$

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

$$\max_{\boldsymbol{x}} \sum_{i=1}^{I} V^i\left(x^i\right) \qquad \text{subject to} \begin{cases} \sum_{i=0}^{I} x^i \boldsymbol{w}^i = \mathbf{0} & \text{(market clearing)} \\ 0 \le x^i \le q^i \text{ for all } i & \text{(order execution rate)}, \end{cases} \tag{7}$$

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

Since this is a quadratic optimization problem with linear equality and inequality constraints, it has a nice structure both for proving existence and uniqueness and for computation

Theorem: Existence and Uniqueness of Quantities

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

Proof.

- Compactness: Inequality constraints on quantities
- Feasibility: No trade is feasible (satisfies constraints with finite value of objective)
- Strictly Concave Objective Function: Quadratic function is strictly concave

□

Dual Problem: Prices

Define Lagrangian

$$L(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := V(\boldsymbol{x}) - \sum_{i=1}^{N} (x^i \cdot \boldsymbol{w}^i)^{\mathsf{T}} \boldsymbol{p} + \boldsymbol{x}^{\mathsf{T}} \boldsymbol{\mu} + (\boldsymbol{q} - \boldsymbol{x})^{\mathsf{T}} \boldsymbol{\lambda} \tag{8}$$

- Prices $\boldsymbol{p}$ are positive or negative Lagrange multipliers enforcing market clearing
- "Taxes" and "subsidies" $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are Lagrange multipliers enforcing order execution rate constraints

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

$$\hat{G}(\boldsymbol{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \max_{\boldsymbol{x}} \; L(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{for} \qquad \boldsymbol{p} \in \mathbb{R}^N, \qquad \boldsymbol{\mu} \geq \mathbf{0}, \qquad \boldsymbol{\lambda} \geq \mathbf{0} \tag{9}$$

The dual problem is

$$g^* := \inf_{\boldsymbol{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}} \hat{G}(\boldsymbol{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{subject to} \qquad \boldsymbol{p} \in \mathbb{R}^N, \qquad \boldsymbol{\mu} \geq \mathbf{0}, \qquad \boldsymbol{\lambda} \geq \mathbf{0} \tag{10}$$

Result: Existence of Market Clearing Prices

Theorem 2 (Existence of Market Clearing Prices)**.** *There exists at least one optimal solution* $(\boldsymbol{p}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *to the dual problem* (10)*. The solutions* $\boldsymbol{x}^*$ *and* $(\boldsymbol{p}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *are a primal-dual pair which satisfies the strict duality relationship*

$$g^* = V(\boldsymbol{x}^*). \tag{11}$$

The prices $\boldsymbol{p}$ *clear markets.*

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 ($\boldsymbol{x} = \boldsymbol{0}$) is feasible: clears markets, satisfies order execution rates
- Linear constraints: Market clearing and order execution rates

Implies strict duality: primal and dual have same solution. Prices clear markets because exchange could lower losses by changing prices if markets did not clear (Bertsekas 2015, Proposition 5.3.4, p. 173) □

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

Duality and General Equilibrium Theory

Primal-dual problem is like a zero-sum game in which which the exchange tries to minimize its losses from trading at non-market clearing prices (Von Neumann)
While our approach has the flavor of general equilibrium theory, our implementation differs in some ways

- General equilibrium theory: Price space is made compact by focusing on relative prices. Existence of prices derived using Kakutani or Brouwer fixed point theorems. Market demand curves may be badly behaved, nonexistent or multiple prices. Uses gross substitutes assumption to make problem well-behaved. Reliance on non-empty interior assumption. Obtains welfare results (Pareto optimality).
- Our approach: Constraints on quantities imply compactness. Quadratic problem simplifies with linear constraints. Does not require non-empty interior assumption. Allow both substitutes and complements (based on same or opposite signs on portfolio weights). Our “welfare” results are based on “as-bid” preferences

Characterization: Karush–Kuhn–Tucker Conditions

Theorem 3 (Necessary and Sufficient Conditions)**.** *The vector of quantities $\boldsymbol{x}^*$ is the unique primal solution and a vector of multipliers $(\boldsymbol{p}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a dual solution if and only if the following conditions hold:*

$$\sum_{i=0}^{I} x^i \, \boldsymbol{w}^i = \mathbf{0}, \quad \mathbf{0} \leq \boldsymbol{x} \leq \boldsymbol{q}, \qquad \textit{(Primal Feasibility)}, \tag{12}$$

$$\boldsymbol{p} \in \mathbb{R}^N, \qquad \boldsymbol{\lambda} \geq \mathbf{0}, \qquad \boldsymbol{\mu} \geq \mathbf{0}, \qquad \textit{(Dual Feasibility)} \tag{13}$$

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathbb{R}^I}{\operatorname{argmax}} \; L(\boldsymbol{x}, \boldsymbol{p}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*), \qquad \textit{(Primal Optimality)} \tag{14}$$

$$\boldsymbol{\lambda}^* \cdot (\boldsymbol{q} - \boldsymbol{x}^*) = \mathbf{0}, \qquad \boldsymbol{\mu}^* \cdot \boldsymbol{x}^* = \mathbf{0} \qquad \textit{(Complementary Slackness)} \tag{15}$$

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

Gains Function

Theorem 4 (Gains function)**.** *Define the gains function as*

$$G(\boldsymbol{p}) := \min_{\boldsymbol{\lambda},\boldsymbol{\mu}} \ \hat{G}(\boldsymbol{p},\boldsymbol{\lambda},\boldsymbol{\mu}) \qquad \textit{subject to} \qquad \boldsymbol{\lambda} \geq \mathbf{0}, \qquad \boldsymbol{\mu} \geq \mathbf{0}. \tag{16}$$

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

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}\in\mathbb{R}^N} \ G(\boldsymbol{p}). \tag{17}$$

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
- Intuition: optimizing against non-market clearing prices allows market participants to achieve greater surplus than trading at market clearing prices
- The derivative of the gains function is minus the market demand function
- The second derivative of the gains function is a negative semi-definite matrix

Computation

The exchange can choose from multiple quadratic optimizers to calculate market clearing prices:

- Solve for quantities: hundreds of thousands of orders, linear constraints
- Minimize gains function to solve for prices: Maybe 1000 prices, no explicit constraints

Quadratic optimizers generally solve for prices and quantities at the same time using KKT conditions

- Simulations using commercial optimizer (Gurobi) or open source optimizer (OSQP) suggest problem with 100 000 orders and 500 assets can be solved to reasonable tolerance in about 10 seconds on a workstation with 18 cores
- Since processing of orders can be parallelized within an iteration, more powerful workstations should be able to calculate prices and quantities in less than one second

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

Additional Issues (continued)

- More complete preference expression: We do not allow all concave utility functions to be represented, such as an order which treats assets as perfect substitutes. Might add additional order types at later stage

Conclusion

- Simple yet powerful method to express preferences—piecewise-linear demand for portfolio flows
- Providing liquidity over time reduces incentives for arms race for speed
- Accommodates various types of financial trading such as pairs trades and index orders executed gradually
- Market clearing quantities and prices exist and are unique with tie-breaking rule
- Outcome maximizes as-bid social welfare
- Outcome is as-bid envy free (given prices, everyone gets their favorite bundle)
- Computationally tractable when scaled to many assets and orders