Optimal Design of a Financial Exchange

Daniel T. Chen[*]

June 17, 2022

Abstract

We consider the design of a market for a single asset where a finite number of risk averse traders may trade to share risk from asset endowments. We derive the direct mechanisms that maximize a linear combination of expected revenue and allocative efficiency. We find that the first best allocation is Bayesian-Nash implementable with ex-ante budget balance if and only if the expectations of traders' endowments are proportional to their risk capacities. We show that an optimal direct mechanism has an indirect implementation by a double auction with side payments. Thus there may be cause for regulation of side payments and potential to use them as effective policy tools.

Keywords: Double Auction, Market Microstructure, Mechanism Design, Make-Take Fees

1. **Introduction**

In modern financial markets, trade takes place according to a panopoly of different protocols. Examples include the double auction, central limit-order book, dark pool, bilateral negotiation in over-the-counter markets, among many others. The variety of different mechanisms is striking. There are two central questions motivating this paper: What are the optimal trading mechanisms that achieve the frontier of attainable combinations of revenue and allocative efficiency? Do these mechanisms resemble any that we see in practice?

We answer these questions for a model environment that is now a workhorse in research on financial market microstructure. There are a finite number of traders who may trade a

[*] dtchen@stanford.edu; Graduate School of Business, Stanford University.

I thank Darrell Duffie, Piotr Dworczak, Yuliy Sannikov, Andrzej Skrzypacz, Robert Wilson, Xinkai Wu, and especially Anthony Zhang for their comments and suggestions. This research is supported by the Bradley Graduate Fellowship through a grant to the Stanford Institute for Economic Policy Research.

single asset. Each trader is endowed with a random quantity of the asset which is realized prior to trade and which is his private information. On occasion, we will refer to the trader's endowment as his type. He incurs a quadratic holding cost over his post trade asset position and his utility is quasilinear in payments. Variants of this framework appear in Du and Zhu (2017b); Chen and Duffie (2021); Rostek and Yoon (2021); Antill and Duffie (2021); Zhang (2020); Wittwer (2021); Sannikov and Skrzypacz (2016) among many others.

Most existing work in this setup fixes a trading mechanism and then analyzes the revenue and allocative efficiency properties of the resulting equilibrium. In this paper, we solve for the trading mechanisms that maximize linear combinations of allocative efficiency and revenue across all feasible trading mechanisms. These are mechanisms such that 1. the designer does not absorb any quantity of the asset and 2. the budget is balanced ex-ante in that the designer's expected revenue is nonnegative. Consequently, the results reveal which inefficiencies are unavoidable features of the primitive economic environment and which are instead the symptoms of a flawed design.

We now summarize the main findings. The first result is the characterization of optimal trading mechanisms. For any weight in the objective, we show that an optimal allocation rule exists, is unique, and can be computed as the solution of a system of equations. The allocation depends only on traders' virtual types. These are similar to the virtual types in Myerson (1981) but differ in that they depend on the shadow cost of binding participation constraints and are therefore endogenous. In the optimal allocation, each trader unloads his virtual type and absorbs a fraction of the aggregate virtual type of the other traders. This fraction is proportional to his risk capacity, the reciprocal of the coefficient on his quadratic holding cost. Thus there is a common component to traders' allocations, which, in the sense of Malamud and Rostek (2017), we can interpret as "aggregate risk". Traders with more risk capacity have more exposure to aggregate risk.

We find that in the optimal allocation, the types of traders whose participation constraints bind are those who expect to trade zero. These are the types whose virtual type is equal in expectation to the aggregate virtual type (weighted by risk capacity). In general, there is a positive measure of these types which form an interval in type space. All types in this interval have the same virtual type. When traders are ex-ante symmetric, there is a positive probability that all traders have types in this interval in which case there is essentially no trade. We show, in an analytical example where traders' endowments are uniformly distributed, the probability that there is no trade decreases with the weight placed on allocative efficiency in the objective and with the number of traders in the market.

It is useful to compare the optimal mechanisms to the first best (ex-post efficient) allocation and to the double auction (without side payments). In optimal mechanisms the distortion of the allocation from first best is relatively higher for intermediate types than for extreme types. First, this reduces the information rents accruing to intermediate types, so that more surplus can be extracted from extreme types. Second, due to convexity of holding costs, there is more potential surplus to extract from extreme types in the first place.

In contrast, for the double auction, the distortion is relatively higher for extreme types than intermediate types. This is suboptimal as extreme types have the most urgent need to unload their positions. Even when allowing the designer to set fixed participation fees, the double auction does not lie on the efficiency-revenue frontier. We show, in an example with $N = 5$ traders and uniformly distributed endowments, that there is a point on the frontier with twice as much revenue and the same level of efficiency as a double auction with participation fees (set at the maximal level such that each type of each trader finds it individually rational to participate).

After characterizing optimal mechanisms, we present a second result which provides a condition for when first best is achievable with ex-ante budget balance. In contrast with Myerson and Satterthwaite (1983), there are some instances when this is possible. The intuition for this is as in Cramton, Gibbons, and Klemperer (1987). Because, traders can be either buyers or sellers, the utility gain from trade of the worst off type is positive. As a result, fees can be charged while satisfying individual rationality constraints, which can then be used to subsidize trade. We show that first best is achievable if and only if the vector of expectations of traders' endowments and the vector of their risk capacities are linearly dependent. In this case, no trader is in expectation a buyer or seller ex-ante. Thus enough fees can be extracted to subsidize efficient trade. McAfee (1991) prove this result for the special case when traders are symmetric. The result in this paper extends that to show there are instances when traders are asymmetric such that efficient trade is possible.

The final result in this paper shows how optimal trading mechanisms can be implemented in practice. We prove that an optimal mechanism can be implemented by a double auction with appropriately calibrated side payments.[1] Thus, while it is not possible to augment a double auction with participation fees and achieve the frontier, it is possible with more general side payments. The side payment of a trader is shown to depend only the clearing price in the double auction and the quantity purchased by that trader at that price. This result holds for an number of traders, for any weight in the objective on revenue relative to efficiency. Thus the presence of side payments can lead to a wide range of outcomes and can be used to target any point on the frontier. A designer can without loss fix the double auction mechanism, and instead focus on calibrating the side payments.

[1] In the double auction, each trader submits a demand schedule specifying the quantity that he will purchase at each realization of the market clearing price.

In practice, there are trading mechanisms which resemble double auctions with side payments. For example, make-take and take-make fees are widely used by exchanges for trade of equities. If we view the double auction as an idealization of the limit-order book, then the implementation result offers a microfoundation for these mechanisms. An implication of the result is that the revenue and efficiency properties of these mechanisms may be sensitive to the side payments used. Thus there may be cause for regulation of them, or on the flipside, opportunities to use them as regulatory tools as well.

The rest of this paper proceeds as follows. Section 2 reviews the related literature. Section 3 presents the model. Section 4 characterizes the set of feasible exchange direct mechanisms. Section 5 formulates the optimization problem of maximizing a linear combination of allocative efficiency and revenue over this set. Section 6 solves for optimal mechanisms, discusses some of their properties, and then presents an illustrative example. Section 7 states and then proves the implementation result. Section 8 concludes.

2. Related Literature

In the literature on mechanism design, analysis of optimal trading mechanisms dates back at least as far as Myerson and Satterthwaite (1983). However, most existing work studies trading mechanisms under assumptions which are not as well-suited to financial markets for several reasons. For instance, most models assume that traders' utilities are linear in the quantity of the asset allocated to them. Most also assume that traders are exogenously designated as buyers or sellers ex-ante. Many models focus on the large market setting or derive asymptotic results as the number of traders diverges. These assumptions may be made in part due to tractability: when there are a finite number of traders with nonlinear utility, the set of implementable allocation rules is often difficult to optimize over. However, they are at odds with some of the core features of financial markets.

First, an essential role of financial markets is to facilitate risk sharing which can only be modeled in a setting where traders have nonlinear utility over asset allocations. Second, a potentially important source of allocative inefficiency in financial markets concerns the strategic avoidance of price impact. An appropriate analysis of this requires modeling a finite number of traders. Third, traders may choose to be either buyers or sellers depending on the terms of trade: as mentioned earlier, this has important implications for allocative efficiency, leading to results that contrast with Myerson and Satterthwaite (1983). The contribution of this paper is to study mechanism design in a framework which is consistent with each of these features.

A closely related paper is Lu and Robert (2001). Lu and Robert (2001) study optimal trading mechanisms in a setting with a finite number of traders who have inelastic demand for a finite quantity of the asset which is common knowledge. Whether traders are buyers or sellers depends endogenously on the mechanism. Traders have linear utility in allocation up to a satiation point determined by a private value. They show that the optimal allocation rule ranks traders in terms of their virtual values. It then allocates the asset to traders with higher virtual values, moving down the list once a trader is satiated. As in the current paper, they also find that there is an intermediate range of types where traders' participation constraints bind who expect to trade zero.

This property also appears in Biais, Martimort, and Rochet (2000) who compute the revenue-optimal mechanism when there is a single trader who has linear-quadratic preferences which are similar to those assumed in this paper. Because there is a single trader, they allow the designer to absorb some quantity of the asset. They do not study exchange mechanisms which requires the presence of multiple traders.

Another related paper is Tatur (2005). Tatur (2005) derives the minimal level of allocative inefficiency, for any given target revenue, achievable by a mechanism in the large market limit. The paper shows that a double auction with a fixed transaction fee can achieve this minimal level in the large market limit. However, he does also show that the constrained efficient outcome can be implemented exactl by a double auction with more complicated though still fixed transaction fees. This is similar to the implementation result in the present paper except that we require more general transaction fees. Moreover, our papers differ in that Tatur (2005) studies the case with single unit demand in which traders are designated as buyers or sellers ex-ante. In contrast, we study a case with multi-unite demand and linear-quadratic utility where a trader may be either a buyer or seller. We also present some analytical examples which show that there are cases when the number of traders is small (5 in the example) such that the double auction with a fixed transaction fee is nowhere near the frontier in our setting.

A closely related paper which studies double auctions with side payments (or transaction costs) is Jantschgi, Nax, Pradelski, and Pycia (2022). In a model with ex-ante buyers or sellers with unit demand Jantschgi, Nax, Pradelski, and Pycia (2022) classify transaction costs based on whether they preserve desirable properties of the double auction mechanism without transaction costs (such as strategy-proofness, asymptotic allocative efficiency as the number of traders grows large). They show that if a trader has a vanishing impact on transaction costs as the number of traders diverges then the desirable properties of the double auction mechanism persist with transaction costs. As in our paper, they allow for essentially any transaction costs which can condition both on the price and quantities traded in the double auction.

There is also a literature in finance that studies the impact of side payments known as make-take/take-make fees on trading in central limit-order books (Malinova and Park, 2015; Colliard and Foucault, 2012; Foucault, Kadan, and Kandel, 2013). Relative to these papers, we allow for essentially any side payment rules and then derive the optimal ones.

3. Model

There are $N > 1$ traders in the market for a single asset. Traders are indexed by $i \in \{1, 2, ..., N\}$. Trader i is endowed with a random quantity Z_i of the asset prior to trade which is his private information. It is common knowledge that Z_i is drawn from the CDF F_i supported on the interval $[\underline{Z}, \overline{Z}] \subset \mathbb{R}$ with continuous density $f_i > 0$. We assume that traders' endowments $(Z_1, ..., Z_N)$ are jointly independent. This assumption allows us to avoid full surplus extracting mechanisms: see, for instance, Crémer and McLean (1988).

A *mechanism* is a tuple $(M_i, X_i, T_i)_{i=1}^N$ where, for each $i \in \{1, 2, .., N\}$,

1. M_i is the *set of messages* that trader i can send,
2. $X_i : \Pi_{i=1}^N M_i \to \mathbb{R}$ is the (trader i) *allocation rule* which maps a profile of received messages to the quantity purchased by trader i,
3. $T_i : \Pi_{i=1}^N M_i \to \mathbb{R}$ is the (trader i) *transfer rule* which maps a profile of received messages to the amount paid by trader i.

An *exchange mechanism* is a mechanism $(M_i, X_i, T_i)_{i=1}^N$ with the property that the total quantity traded is zero for each profile of submitted messages:

$$\sum_{i=1}^{N} X_i = 0. \tag{1}$$

The utility of trader i from participating in the mechanism $(M_i, X_i, T_i)_{i=1}^N$ is

$$u_i(Z_i, m_i, m_{-i}) = -\frac{1}{2\kappa_i}\big(Z_i + X_i(m_i, m_{-i})\big)^2 - T_i(m_i, m_{-i}) \tag{2}$$

if he reports the message $m_i \in M_i$ and the profile of the other traders' reports is $m_{-i} \in \Pi_{j \neq i} M_j$.[2] Trader i incurs a quadratic holding cost from retaining a net position in the

[2] If we multiply the right-hand side of equation (2) by the risk capacity κ_i, then all traders have the same holding costs but different marginal utilities for transfers. It may be interesting for future work to analyze the model under this interpretation. This has been studied under the large market assumption with linear utility and ex-ante buyers or sellers by Dworczak, Kominers, and Akbarpour (2021).

asset post-trade. Above, $\kappa_i > 0$ denotes trader i's *risk capacity*. When it is larger, trader i's holding cost is lower. For tractability, we assume that risk capacities are common knowledge. If trader i does not participate in the mechanism, then his utility is simply

$$-\frac{1}{2\kappa_i} Z_i^2.$$

Preferences of the form in (2) have been assumed by a wide range of papers on financial market microstructure (Du and Zhu, 2017a; Sannikov and Skrzypacz, 2016; Chen and Duffie, 2021; Rostek and Yoon, 2021). These papers also assume an information structure where endowments are private information but risk capacities are common knowledge. The majority of these papers assume that trading takes place by a double auction which is tractable and resembles trading on lit-exchanges relatively well. This paper seeks to derive optimal trading mechanisms while retaining the other aspects of the model. We then ask whether these derived mechanisms resemble anything like what we see in practice.

Since traders' utilities are quasilinear in payments, it is natural to measure *allocative efficiency* by the sum of traders' utilities from their post-trade positions gross of any payments

$$\sum_{i=1}^{N} -\frac{1}{2\kappa_i}\big(Z_i + X_i(m_i, m_{-i})\big)^2.$$

In what follows, we consider maximizing convex combinations of expected revenue and expected allocative efficiency over all allocations and transfers that are implementable by *exchange mechanisms*.[3] By the revelation principle, it is without loss to restrict attention to *exchange direct mechanisms*. These are *exchange mechanisms* with the additional property that $M_i = [\underline{Z}, \overline{Z}]$ for each $i \in \{1, 2, .., N\}$. Later, it will be clear that these optimal exchange direct mechanisms would also be implementable if traders instead had CARA utility assuming that the distribution of the underlying asset's payoff is standard Gaussian.[4] It will also turn out that the indirect implementation result of Section 7 extends to this case.

4. Feasible Exchange Direct Mechanisms

Following Myerson (1981), in this section we characterize feasible exchange direct mechanisms. These are exchange direct mechanisms where it is individually rational for trader

[3] We mean specifically Bayes-Nash implementability.

[4] However, the mechanisms may or may not be optimal in this setup. The main difference between this alternative setup and the current model is that traders are risk averse over transfers in the former. While we assume that the asset has zero payoffs in expectation, the results can be extended straightforwardly to allow for payoffs which are nonzero in expectation provided the expectation exists.

i to participate in the mechanism for any realization of his endowment Z_i and Bayesian incentive compatible for him to truthfully report his endowment Z_i.

Formally, an exchange direct mechanism $(X_i, T_i)_{i=1}^N$ is *feasible* if the following hold

1. for each Z_i and $\tilde{Z}_i$ in $[\underline{Z}, \overline{Z}]$

$$\text{(IC)} \qquad \mathbb{E}[u_i(Z_i, Z_i, Z_{-i})|Z_i] \geq \mathbb{E}[u_i(Z_i, \tilde{Z}_i, Z_{-i})|Z_i].$$

2. for each Z_i in $[\underline{Z}, \overline{Z}]$

$$\text{(IR)} \qquad \mathbb{E}[u_i(Z_i, Z_i, Z_{-i})|Z_i] \geq -\frac{1}{2\kappa_i} Z_i^2.$$

Recall that by definition, (1) must hold as well. In what follows, we derive simple necessary and sufficient conditions on $(X_i, T_i)_{i=1}^N$ for it to be feasible. To ease notation, let

$$U_i(Z_i) := \mathbb{E}[u_i(Z_i, Z_i, Z_{-i})|Z_i].$$

Local incentive compatibility of any feasible exchange direct mechanism implies that

$$(3) \qquad U_i(Z_i) + \frac{1}{2\kappa_i} Z_i^2 = \int_{\underline{Z}}^{Z_i} -\frac{1}{\kappa_i} \mathbb{E}\left[X_i(s, Z_{-i})\right] ds + U_i(\underline{Z}) + \frac{1}{2\kappa_i} \underline{Z}^2.$$

Equation (3) is the utility gain from participation of a trader of type Z_i. Notice that the utility gain is linear in the allocation rule which allows for a tractable analysis. This is because holding costs are quadratic. Under a feasible exchange direct mechanism, individual rationality (IR) is satisfied if the right-hand side of (3) is nonnegative.

Using equation (3) together with (2) we see that the interim transfer rule must satisfy

$$(4) \quad \mathbb{E}[T_i(Z_i, Z_{-i})|Z_i] = \mathbb{E}\left[-\frac{1}{2\kappa_i} X_i(Z_i, Z_{-i})^2 - \frac{1}{\kappa_i} Z_i X_i(Z_i, Z_{-i}) + \frac{1}{\kappa_i} \int_{\underline{Z}}^{Z_i} X_i(s, Z_{-i}) ds \middle| Z_i \right] - U_i(\underline{Z}) - \frac{1}{2\kappa_i} \underline{Z}^2$$

in any feasible exchange direct mechanism. Thus, as in Myerson (1981) we are able to express the interim transfer rule in terms of the allocation rule and the utility of the lowest type, though the allocation rule enters nonlinearly in our case.

Using equation (3) we can also derive a necessary and sufficient condition for gobal incentive compatibility (IC). To see this, by inspecting equation (2) for the utility of trader i, we see that when a type Z_i trader i deviates to report $\tilde{Z}_i$ he obtains utility

$$U_i(\tilde{Z}_i) - \frac{1}{2\kappa_i}\mathbb{E}\left[\left(Z_i + X_i(\tilde{Z}_i, Z_{-i})\right)^2 | Z_i\right] + \frac{1}{2\kappa_i}\mathbb{E}\left[\left(\tilde{Z}_i + X_i(\tilde{Z}_i, Z_{-i})\right)^2 | Z_i\right].$$

Thus (IC) is equivalent to the condition

$$U_i(Z_i) - U_i(\tilde{Z}_i) \geq -\frac{1}{2\kappa_i}(Z_i^2 - \tilde{Z}_i^2) - \frac{1}{\kappa_i}(Z_i - \tilde{Z}_i)\mathbb{E}[X_i(\tilde{Z}_i, Z_{-i})|Z_i]$$

for each Z_i, $\tilde{Z}_i \in [\underline{Z}, \overline{Z}]$. Then, by equation (3), (IC) is equivalent to

$$\int_{Z_i}^{\tilde{Z}_i} \mathbb{E}[X_i(s, Z_{-i})]ds \geq (\tilde{Z}_i - Z_i)\mathbb{E}[X_i(\tilde{Z}_i, Z_{-i})|\tilde{Z}_i]$$

for each Z_i, $\tilde{Z}_i \in [\underline{Z}, \overline{Z}]$ such that $Z_i \geq \tilde{Z}_i$. This is in turn equivalent to the interim allocation rule $\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i]$ being weakly decreasing in Z_i.

We sum up the results of this section in the following Lemma 1.

Lemma 1. *An exchange direct mechanism $(X_i, T_i)_{i=1}^N$ is feasible if and only if the following hold for each trader i:*

1. *the utility gain from participation given by (3) is nonnegative for each realization of Z_i in $[\underline{Z}, \overline{Z}]$.*
2. *the interim transfer rule $\mathbb{E}[T_i(Z_i, Z_{-i})|Z_i]$ satisfies (4) for each realization of Z_i in $[\underline{Z}, \overline{Z}]$.*
3. *the interim allocation rule $\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i]$ is weakly decreasing in the realization of Z_i.*

Lemma 1 is an adaptation of Myerson's lemma to our model setting. We use Lemma 1 to tractably formulate the optimization problem which we now describe.

5. The Optimization Problem

We seek to compute the exchange direct mechanisms that maximize convex combinations of revenue and allocative efficiency, subject to budget balance. In this section, we set up the optimization problem.

Using equation (4) for the interim transfer rule, taking an unconditional expectation, integrating by parts, and summing across traders we can compute the expected revenue of a given exchange direct mechanism with allocation rule $(X_i)_{i=1}^N$ and $\underline{Z}$-type utilities $(U_i(\underline{Z}))_{i=1}^N$:

$$R\Big((X_i, U_i(\underline{Z}))_{i=1}^N\Big) := \sum_{i=1}^N \mathbb{E}\left[-\frac{1}{\kappa_i}\left(Z_i - \frac{1-F_i(Z_i)}{f_i(Z_i)}\right) X_i(Z_i, Z_{-i})\right] \\ - \sum_{i=1}^N \frac{1}{2\kappa_i}\mathbb{E}[X_i(Z_i, Z_{-i})^2] - U_i(\underline{Z}) - \frac{1}{2\kappa_i}\underline{Z}^2. \tag{5}$$

We denote the expected allocated efficiency of the mechanism by

$$A\Big((X_i)_{i=1}^N\Big) := \mathbb{E}\left[\sum_{i=1}^N -\frac{1}{2\kappa_i}\left(Z_i + X_i(Z_i, Z_{-i})\right)^2\right].$$

Then, using Lemma 1, the objective is, for a given $\alpha \in [0, 1]$,

$$\max_{(X_i, U_i(\underline{Z}))_{i=1}^N} \alpha R\Big((X_i, U_i(\underline{Z}))_{i=1}^N\Big) + (1-\alpha) A\Big((X_i)_{i=1}^N\Big) \tag{6}$$

such that,

$$R\Big((X_i, U_i(\underline{Z}))_{i=1}^N\Big) \geq 0, \tag{7}$$

$$\sum_{i=1}^N X_i(Z_i, Z_{-i}) = 0, \tag{8}$$

for each $i \in \{1, 2, ..., N\}$,

$$\int_{\underline{Z}}^{Z_i} -\frac{1}{\kappa_i}\mathbb{E}\left[X_i(s, Z_{-i})\right] ds + U_i(\underline{Z}) + \frac{1}{2\kappa_i}\underline{Z}^2 \geq 0 \tag{9}$$

for each $Z_i \in [\underline{Z}, \overline{Z}]$, and in addition

$$\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i] \text{ is weakly decreasing in } Z_i. \tag{10}$$

Equation (7) is the (ex-ante) budget balance condition. Equation (8) is the condition that the designer can not absorb any net quantity of the asset. Equations (9) and (10) are necessary and sufficient for feasibility by Lemma 1. In Appendix A we prove, using a projection argument, that there exists a unique solution to (6).

6. Optimal Exchange Direct Mechanisms

In this section, we compute, under some conditions, the direct mechanism that solves the problem formulated in the last section. We then discuss, in an analytical example, the basic properties of the mechanism, how it changes with α, and how it differs from the double auction without side payments. The derivation of the optimal mechanism follows a similar roadmap to that of Biais, Martimort, and Rochet (2000). Biais, Martimort, and Rochet (2000) compute the revenue-maximizing mechanism in a different setting but where a trader also has linear-quadratic utility. However, they restrict attention to the case of a single trader and thus do not analyze exchange mechansisms.

To compute the optimal mechanism we first ignore the monotonicity constraint on the interim allocation rule. We then formulate the Lagrangian and then optimize. We show that the solution of the relaxed problem satisfies the monotonicity constraint under a condition on the distributions $(F_i)_{i=1}^N$.

Let $\varOmega_i$ denote the Lagrange multiplier associated with trader i's participation constraint (9). $\varOmega_i$ is a nonnegative nondecreasing function of bounded variation on $[\underline{Z}, \overline{Z}]$. By complementary slackness, it is increasing only when the pariticipation constraint is binding. Let λ, a nonnegative constant, denote the Lagrange multiplier on the budget balance constraint (7). Then, after integrating by parts and ignoring some constant terms, the Lagrangian associated with the relaxed version of (6) is:

$$(11) \quad \max_{(X_i, U_i(\underline{Z}))_{i=1}^N} \Bigg\{ \sum_{i=1}^N \mathbb{E}\left[-\frac{1}{\kappa_i} \left(Z_i(1+\lambda) - \frac{(\alpha+\lambda)(1-F_i(Z_i)) + \varOmega_i(Z_i) - \varOmega_i(\overline{Z})}{f_i(Z_i)} \right) X_i(Z_i, Z_{-i}) \right] \\ - \sum_{i=1}^N \mathbb{E}\left[\frac{1}{2\kappa_i} (1+\lambda)\, X_i(Z_i, Z_{-i})^2 \right] - \sum_{i=1}^N \left(\alpha + \lambda - \varOmega_i(\overline{Z})\right) \left(U_i(\underline{Z}) + \frac{1}{2\kappa_i} \underline{Z}^2 \right) \Bigg\}.$$

such that (8) holds. (11) can be solved pointwise. Each pointwise problem is a finite linear-quadratic program with a single linear equality constraint. These problems are standard and it is well known that they can be solved using the method of Lagrange multipliers to incorporate the constraint (8). Note that we do not incorporate (8) directly into the Lagrangian (11) but instead retain it in the definition of the domain: to prove that it is necessary and sufficient for a mechanism to solve (11) (for some multipliers satisfying complementary slackness) to also solve the relaxed version of (6), we rely on a theorem in Luenberger (1997) which requires the existence of a point in the domain such that the constraints are slack. See Appendix A for details.

We now construct the solution to (11) and the multipliers. By inspection, for there to be an interior solution to (11), it must be that $\Omega_i(\overline{Z}) = \alpha + \lambda$ for each trader i. Thus, as one might expect, the participation constraint must be binding for some types of each trader i. Solving (11), we derive

$$X_i(Z_i, Z_{-i}) = -V_i(Z_i) + \frac{\kappa_i}{\sum_{j=1}^N \kappa_j} \sum_{j=1}^N V_j(Z_j) \tag{12}$$

where

$$V_j(Z_j) := \frac{1}{1+\lambda}\left(Z_j - \frac{\Omega_j(Z_j) - (\alpha+\lambda)F_j(Z_j)}{f_j(Z_j)}\right).$$

Following Myerson (1981), we refer to V_j as trader j's virtual value (though related, it has a somewhat different interpretation in the current setting: see equation (5)). Before deriving the equations that pin down the multipliers, we first note some interesting properties of the optimal allocation rule that we can deduce from equation (12). First, notice that trader i unloads his virtual value which is absorbed by the other traders. In return, trader i absorbs a fraction proportional to his risk capacity of the aggregate virtual value.

Next, by setting the multipliers and α to zero, we see that the first best allocation (the allocation that minimizes the sum of traders' holding costs without imposing budget balance) is the following.

Lemma 2. *The first best allocation sets*

$$X_i(Z_i, Z_{-i}) = -Z_i + \frac{\kappa_i}{\sum_{j=1}^N \kappa_j} \sum_{j=1}^N Z_j.$$

Thus the optimal allocation is of the same form as the first best allocation except with virtual values in place of traders' types. Similar results have been found in other settings that analyze trading mechanisms such as Wilson (1985) and Lu and Robert (2001) which we reviewed in the introduction. Wilson (1985) terms this the *generalized double auction in virtual values* property. In Section 7, we show that it is literally possible to implement the optimal allocation by a double auction when allowing for side payments in the current model setting.

We now turn our attention to computing the multipliers. For this, it will be useful to assume the following two conditions.

Condition 1. *The interim allocation rule corresponding to the first best allocation is negative at $\overline{Z}$ and positive at $\underline{Z}$ for each trader i.*

Condition 2. $\frac{1-F_i(Z_i)}{f_i(Z_i)}$ *is decreasing and* $\frac{F_i(Z_i)}{f_i(Z_i)}$ *is increasing in Z_i for each $i \in \{1, 2, ..., N\}$.*

Condition 1 states that under the first best allocation, a trader with the highest endowment realization will on average sell and a trader with the lowest endowment realization will on average buy. Condition 2 is a regularity condition on the distributions F_i. It is not very restrictive. A sufficient condition is that the density f_i is continuously differentiable and log-concave which is satisfied by many common distributions. Using these conditions, we can prove the following result.

Lemma 3. *Under Conditions 1 and 2 the types Z_i of trader i with binding participation constraints are such that*

$$\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i] = 0.$$

The set is an interval, $[Z_{ia}, Z_{ib}] \subseteq [\underline{Z}, \overline{Z}]$.[5] *The candidate optimal allocation rule (12) is weakly decreasing in Z_i and thus the associated interim allocation rule is weakly decreasing.*

Proof. Inspecting equation (9), by Condition 1, the only types Z_i for which the participation constraint (9) can be binding are such that

$$\mathbb{E}\left[X_i(Z_i, Z_{-i})|Z_i\right] = 0.$$

Condition 1 ensures that this set is nonempty and lies in the interior of $[\underline{Z}, \overline{Z}]$. We shall argue that this set is an interval of the type $[Z_{ia}, Z_{ib}] \subseteq [\underline{Z}, \overline{Z}]$. Suppose for contradiction that there are points $\underline{Z} \le a < b < c \le \overline{Z}$ such that the participation constraint is binding at $Z_i = a$ and $Z_i = c$ but not at $Z_i = b$. Let b^* denote the infimum over types $Z_i > a$ such that the participation constraint is not binding. By continuity of the utility gain in (9), there is a neighborhood $(b^*, b^* + \epsilon)$ for $\epsilon > 0$ small such that the participation constraint is not binding. Thus, Ω_i is nonincreasing in this region by complementary slackness. But Condition 2 implies that

$$V_i(Z_i) = \frac{1}{1+\lambda}\left(Z_i - \frac{\Omega(Z_i) - (\alpha+\lambda)F(Z_i)}{f(Z_i)}\right)$$

is thereby increasing in this neighborhood. To see this, observe that since F/f is increasing, the only way for V_i to be decreasing is if f is decreasing. But, if f is decreasing, since $\Omega_i \le \alpha + \lambda$ and since $(1 - F)/f$ is decreasing by Condition 2, V_i still must be increasing

[5] We allow for the case when $Z_{ib} = Z_{ia}$ so the interval is just a point.

in the neighborhood. This implies that $\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i] < 0$ and is decreasing in the neighborhood. The only way for $\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i]$ to begin increasing again and eventually reach 0 at c is if Ω_i starts increasing. But this can never happen by complementary slackness since the participation constraint can only bind when $\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i] = 0$. That is, the only way to reach $\mathbb{E}[X_i(Z_i, Z_{-i})|Z_i] = 0$ again after b^* is if it has already happened which is a contradiction. Thus, we have that in the interval $[\underline{Z}, Z_{ia}]$,

$$V_i(Z_i) = \frac{1}{1+\lambda}\left(Z_i + (\alpha+\lambda)\frac{F_i(Z_i)}{f_i(Z_i)}\right).$$

In the interval $[Z_{ia}, Z_{ib}]$ where the constraint binds,

$$V_i(Z_i) = \frac{\kappa_i}{\sum_{j\neq i}\kappa_j}\mathbb{E}\left[\sum_{j\neq i} V_j(Z_j)\right]. \tag{13}$$

In the interval $[Z_{ib}, \overline{Z}]$,

$$V_i(Z_i) = \frac{1}{1+\lambda}\left(Z_i - (\alpha+\lambda)\frac{1-F_i(Z_i)}{f_i(Z_i)}\right).$$

V_i is therefore weakly increasing, implying that X_i is weakly decreasing in Z_i. □

Lemma 3 states that (12) is weakly decreasing and therefore a candidate solution to (6). Moreover the types for whom the participation constraint binds lie in an interval and are such that the expected trade quantity is zero. Using these facts, we will derive a system of equations that characterizes the multipliers. To ease notation, let us denote

$$\eta_i := \frac{\kappa_i}{\sum_{j\neq i}\kappa_j}\mathbb{E}\left[\sum_{j\neq i} V_j(Z_j)\right].$$

Then, writing out the expectation

$$\eta_i = \frac{\kappa_i}{\sum_{j\neq i}\kappa_j}\frac{1}{1+\lambda}\sum_{j\neq i}\left(-\int_{Z_{ja}}^{Z_{jb}}\Omega_j(s)ds + \big(1-(\alpha+\lambda)\big)\mathbb{E}[Z_j] + (\alpha+\lambda)(Z_{jb} - \underline{Z})\right). \tag{14}$$

Next, equation (13) which implies that

$$\eta_i = \frac{1}{1+\lambda}\left(Z_i - \frac{\Omega_i(Z_i) - (\alpha+\lambda)F_i(Z_i)}{f_i(Z_i)}\right) \tag{15}$$

for $Z_i \in [Z_{ia}, Z_{ib}]$. Taking an expectation of both sides conditional on $Z_i \in [Z_{ia}, Z_{ib}]$, we derive

$$\eta_i\big(F_i(Z_{ib}) - F_i(Z_{ia})\big)(1+\lambda) = Z_{ib} - Z_{ia} - \big(1-(\alpha+\lambda)\big)\int_{Z_{ia}}^{Z_{ib}} F_i(s)ds - \int_{Z_{ia}}^{Z_{ib}} \Omega_i(s)ds. \tag{16}$$

Next the interval endpoints must satisfy

$$Z_{ia} + (\alpha+\lambda)\frac{F_i(Z_{ia})}{f_i(Z_{ia})} = \eta_i \tag{17}$$

and

$$Z_{ia} + (\alpha+\lambda)\frac{F_i(Z_{ia})}{f_i(Z_{ia})} = Z_{ib} - (\alpha+\lambda)\frac{1-F_i(Z_{ib})}{f_i(Z_{ib})} \tag{18}$$

which follow from (13) and the fact that $\Omega_i(Z_{ib}) = \alpha + \lambda$. The last condition is the complementary slackness condition on the budget balance constraint:

$$\lambda R\Big((X_i, U_i(\underline{Z}))_{i=1}^N\Big) = 0 \tag{19}$$

Together, equations (14), (16), (17), (18), (19) form a system of equations which can be used to derive the unknowns $\Big(Z_{ia}, Z_{ib}, \int_{Z_{ia}}^{Z_{ib}} \Omega(s)ds, \eta_i\Big)_{i=1}^N$. Once this system has been solved $(\Omega_i)_{i=1}^N$ can be derived from (15).

Theorem 1. *Under Conditions 1 and 2, an optimal allocation rule $(X_i)_{i=1}^N$ exists and is uniquely optimal up to measure zero differences. It satisfies (12) for Lagrange multipliers and participation intervals that solve equations (14), (16), (17), (18), and (19).*

Proof. See Appendix A. □

At a first glance, the equation system may seem unwieldy. However it is useful to note that given λ and the participation intervals $(Z_{ia}, Z_{ib})_{i=1}^N$, the remaining unknowns $\Big(\eta_i, \int_{Z_{ia}}^{Z_{ib}} \Omega(s)ds\Big)_{i=1}^N$ solve a system of linear equations and are therefore characterized explicitly. It is also worthwhile to note that either $\lambda = 0$ or the optimal allocation rule corresponds to the constrained efficient allocation where $\alpha = 0$. This simplifies the possibilities down further. It is numerically tractable to solve the system for two groups of traders such that traders may differ across groups but are symmetric within them. It also turns out that there is a special case when everything is computable explicitly, and which may therefore serve as a useful laboratory.

Corollary 1.1. *Suppose that it is known that the budget balance constraint does not bind in that* $\lambda = 0$ *as, for example, under the revenue maximizing mechanism. Then the optimal mechanism can be computed explicitly when each* F_i *is the uniform distribution as the solution to a system of linear equations.*[6]

Proof. In the special case when each F_i is uniform, the equation system (15), (16), (17), and (18) is linear in the unknowns $\left(Z_{ia}, Z_{ib}, \int_{Z_{ia}}^{Z_{ib}} \Omega(s)ds, \eta_i\right)_{i=1}^{N}$. The linear system is reported in Appendix A. □

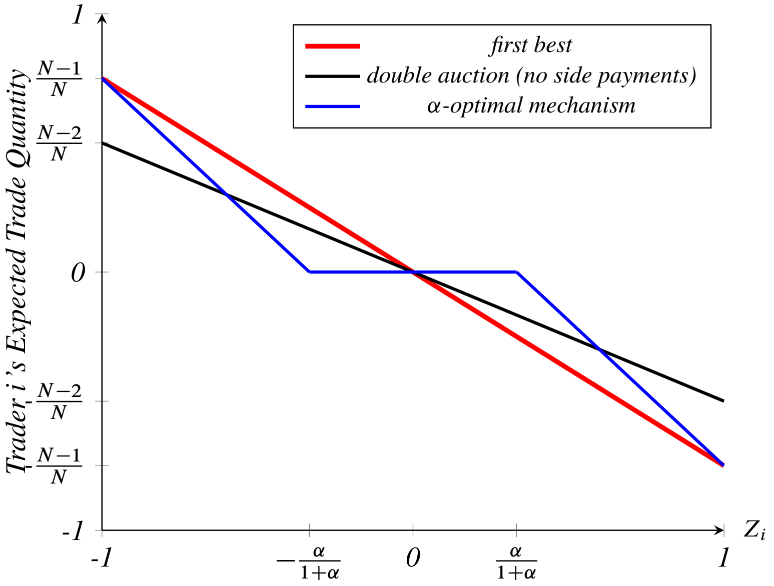
Corollary 1.2. *It is possible to implement the optimal allocation rule* (12) *with transfers such that trader i 's transfer is deterministic conditional on his endowment* Z_i*. This can be done by setting* $T_i(Z_i)$ *equal to the right-hand side of* (4)*. Thus, if each trader i instead had CARA utility and the asset had standard Gaussian payoffs then the allocation rule* (12) *would still be implementable.*

Theorem 1 characterizes optimal mechanisms. Below, we present an analytical example to illustrate some of their properties.

Example 1. *Suppose that each trader i has risk capacity* $\kappa_i = 1/2$ *and distribution* F_i *over asset endowment* Z_i *which is uniform on* $[-1, 1]$*. Then it can be shown that* $Z_{ib} = -\frac{\alpha}{1+\alpha}$, $Z_{ia} = \frac{\alpha}{1+\alpha}$, *and* $\Omega_i(Z_i) = \frac{1+\alpha}{2}Z_i + \frac{\alpha}{2}$ *for* $Z_i \in [Z_{ia}, Z_{ib}]$*. We display the interim allocation rule of any given trader i in the graph below in blue. We also plot the interim allocation rule associated with the symmetric linear equilibrium of the double auction without side payments in black and that of the first best allocation in red.*[7]

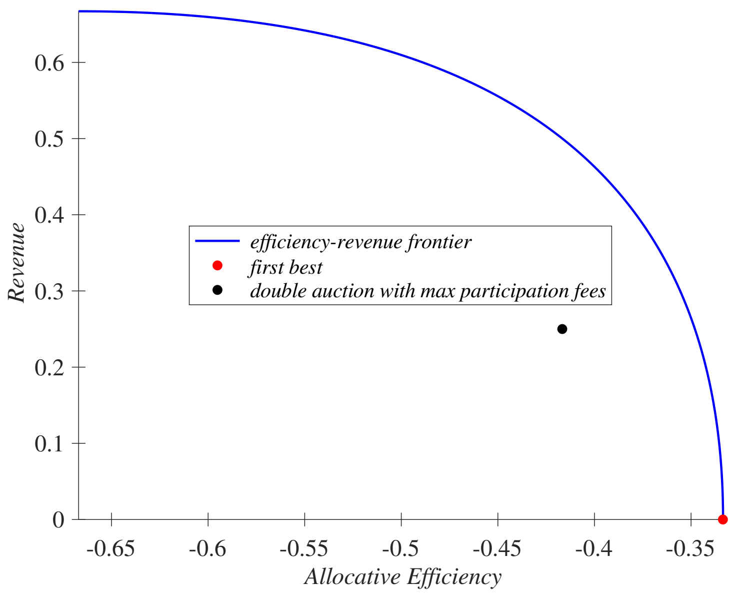
[6] The result also goes through if each F_i is supported on any interval $[\underline{Z}_i, \overline{Z}_i] \subseteq [\underline{Z}, \overline{Z}]$ For expositional convenience we assumed $f_i > 0$ on all of $[\underline{Z}, \overline{Z}]$ but it is straightforward to extend the model to accomodate differing supports for traders' endowments.

[7] The black line is valid only for $N > 2$. If $N = 2$ it is well known that there is no symmetric linear equilibrium of the double auction without side payments.



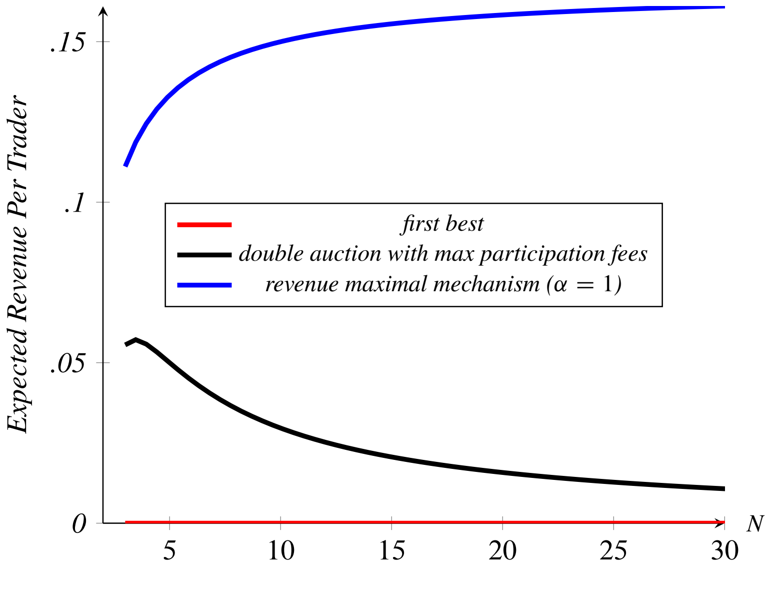
When compared with first best, the double auction distorts the allocation of traders with more extreme endowments relatively more with the distortion decreasing towards the center. Given the convexity of traders' holding costs this is suboptimal for any weight α in the objective. In contrast, the α-optimal mechanism distorts the allocation of intermediate types in order to extract more revenue from types with more extreme endowments who benefit greatly from unloading their positions. The degree of the distortion in this example is determined by the length of the participation interval which shrinks as the weight on revenue in the objective decreases. In the extreme case when $\alpha = 1$, participation constraints bind for half of the interval. Thus trade does not occur at all with probability $1/2^N$. On the other hand, when $\alpha = 0$, the α-optimal mechanism coincides with first best and trade occurs almost surely. Thus, it turns out that in this example, the first best allocation of Lemma 2, is implementable with ex-ante budget balance though this is not generally true as we later discuss.

Below, we plot the frontier of attainable combinations of revenue and allocative efficiency for the case when the number of traders is $N = 5$.

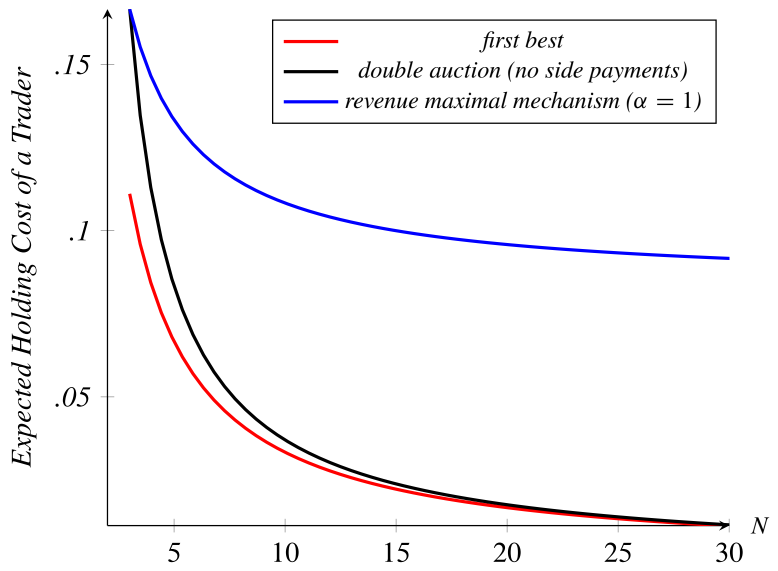


As seen above, the double auction with participation fees does not lie on the frontier. For the same level of allocative efficiency, there is another mechanism which can obtain roughly twice the revenue.

The following two graphs illustrate how expected revenue and allocative efficiency vary across the mechanisms as the number of traders increases starting from $N = 3$.



Under the revenue maximal mechanism, the expected revenue per trader is increasing in the number of traders and asymptotes at 1/6. Total revenue therefore diverges. In contrast, for the double auction with maximum participation fees, revenue per trader converges to 0. It can be shown that total revenue asymptotes at 1/3.



While the double auction and first best are such that holding costs eventually disappear, under the revenue maximal mechanism it remains strictly positive for each trader and asymptotes to $1/12$*.*

In Example 1, it turned out that first best is attainable with budget balance. This is not generally true.

Proposition 1. *The first best allocation which minimizes the sum of traders' holding costs is achievable with ex-ante budget balance if and only if the vector of the means of traders' endowments* $\left[\mathbb{E}[Z_1], ..., \mathbb{E}[Z_N]\right]$ *and the vector of their risk capacities* $[\kappa_1, ..., \kappa_N]$ *are linearly dependent.*

Proof. Compute the minimal utility for $\underline{Z}$-types under the first best allocation (2). Substitute these into the equation for revenue (5) to see that revenue is nonnegative if and only if the condition in Proposition 1 holds. □

A special case of Proposition 1 is when traders are ex-ante symmetric. This case was proven by McAfee (1991) who studies a similar model with hidden endowments. Here we show that it holds in some cases when traders are asymmetric. The intuition is the similar to that of Cramton, Gibbons, and Klemperer (1987) who show that when agents own shares of a partnership, it is possible to achieve the efficient outcome provided shares are not too asymetrically distributed. This is because each agent of each type receives positive surplus. Provided that this is large enough, the designer can charge participation fees to cover the costs of subsidies needed to achieve efficiency. Similarly, in this paper, traders can be either buyers or sellers. Thus each type of each trader earns a surplus from participating in the mechanism. Provided the surplus is high enough, efficiency is attainable. This happens to be the case under the condition in Proposition 1 where it turns out that ex-ante, no trader is a net buyer or seller of the asset in expectation.

7. Implementation Result

We have thus far characterized optimal exchange direct mechanisms and discussed their main properties. Do these direct mechanisms have indirect implementations which resemble trading mechanisms used in practice for exchange of financial assets? In this section, we show that an optimal exchange direct mechanism has an indirect implementation in the form of a double auction with side payment rules which are trader-specific (in the special case when traders are ex-ante symmetric, then the side payment rules are also the same for each trader). These side payment rules are conditioned only on the quantity that a given trader i purchases in the auction and the clearing price. The proof is constructive: we show how to calibrate the side payment rule to implement a given optimal exchange direct mechanism.

The intuition behind this result is the following. By inspecting the optimal allocation rule (12), we see that all traders must absorb the aggregate virtual value. This is the common component or aggregate risk (in the sense of Malamud and Rostek (2017)) which all traders must be exposted to. We may think of this common component as corresponding to the clearing price in a double auction. In a double auction mechanism, traders' demands condition on the price which links their purchases together through market clearing. Thus, to obtain the implementation result, we will later show that we can find demand schedules (one for each trader) which lead to the clearing price being the aggregate value, and side payment rules such that these demand schedules constitute an equilibrium.

Before stating the result, we first specify the protocol for a *double auction with side payment rules* $\left(S_i : \mathbb{R}^2 \to \mathbb{R}\right)_{i=1}^N$.

1. After observing his endowment Z_i, each trader i submits a measurable demand schedule $q_i : \mathbb{R} \to \mathbb{R}$ which specifies the quantity that trader will purchase in the auction for each realization of the clearing price p.

2. The clearing price p is computed such that

$$\sum_{i=1}^{N} q_i(p) = 0.$$

   If there does not exist a clearing price or the clearing price is not unique then no trades are executed.

3. Each trader i receives the quantity $q_i(p)$ but must pay $pq_i(p) + S_i(p, q_i(p))$.

Theorem 2 presents the side payment rules which implement a given optimal exchange direct mechanism. It also characterizes the associated equilibrium in demand schedules of the double auction with these side payment rules.

Theorem 2. *The allocation rule* $(X_i)_{i=1}^N$ *given by* (12) *is implementable by a double auction with side payments where trader i's side payment rule is*

$$S_i(p, q_i(p)) = \frac{c_i}{2} q_i(p)^2 - T_i\left(-q_i(p) - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j} p\right) + \tau.$$

Above, c_i *is the constant defined by*

$$c_i = -\frac{1}{\kappa_i} + \frac{\sum_{j=1}^N \kappa_j}{\kappa_i} - \frac{\sum_{j=1}^N \kappa_j}{\sum_{j \neq i}^N \kappa_j},$$

$T_i : \mathbb{R} \to \mathbb{R}$ *is the function defined by*[8]

$$T_i(l) = \left(\sum_{j=1}^N \kappa_j \frac{l^2}{2} - \int_{\underline{Z}}^{l} V_i^{-1}(s) ds\right) \frac{\sum_{j \neq i} \kappa_j}{\kappa_i \sum_{j=1}^N \kappa_j}$$

for $l \in [\underline{Z}, \overline{Z}]$*, and* τ *is a constant representing the participation fee which may be computed explicitly to maximize revenue.*

In the implementing equilibrium of the double auction with side payment rule so defined, trader i submits the demand schedule

$$q_i(p) = -V_i(Z_i) - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j} p \tag{20}$$

given his type $Z_i \in [\underline{Z}, \overline{Z}]$.

Proof. Given $\epsilon > 0$, let $\hat{V}_i(Z_i)$ denote

$$\hat{V}_i(Z_i) = V_i(Z_i) + \epsilon \max\{\min\{Z_{ib}, Z_i\} - Z_{ia}, 0\}.$$

This is the virtual value of trader i with type Z_i, but perturbed over the interval $[Z_{ia}, Z_{ib}]$ of binding participation constraints so that the function is strictly increasing. Let $\hat{X}_i(Z_i, Z_{-i})$

[8] V_i is not invertible at η_i but this does not affect the value of the integral: we can assign any finite value to the inverse at that point.

denote

$$\hat{X}_i(Z_i, Z_{-i}) = -\hat{V}_i(Z_i) + \frac{\kappa_i}{\sum_{j=1}^N \kappa_j} \sum_{j=1}^N \hat{V}_j(Z_j). \tag{21}$$

Then as $\epsilon \to 0$, $\hat{X}_i \to X_i$ pointwise. Also, note that by construction $\sum_{i=1}^N \hat{X}_i = 0$ so that the auctioneer does not retain any net position in the asset.

Suppose that each trader i submits the demand schedule given by

$$\hat{q}_i(p) = -\hat{V}_i(Z_i) - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j} p \tag{22}$$

to the double auction. Then the resulting market clearing price is

$$p = \sum_{j=1}^N \hat{V}_j(Z_j).$$

One can verify that the resulting trade quantities coincide with the allocation (21). Moreover, the inverse residual demand curve facing trader i can be computed as follows. By market clearing,

$$\hat{q}_i + \sum_{j \neq i} -\hat{V}_j(Z_j) - \frac{\sum_{j \neq i} \kappa_j}{\sum_{j=1}^N \kappa_j} p = 0.$$

This implies that

$$p = \left(\hat{q}_i + \sum_{j \neq i} -\hat{V}_j(Z_j) \right) \frac{\sum_{j=1}^N \kappa_j}{\sum_{j \neq i} \kappa_j}.$$

Thus, the total impact of trader i on the price is $\frac{\sum_{j=1}^N \kappa_j}{\sum_{j \neq i} \kappa_j} \hat{q}_i$.

Next, we construct the side payment rule such that it is optimal for trader i to submit the demand schedule (22). We consider a side payment rule of the form

$$S_i(p, q_i(p)) = \frac{\hat{c}_i}{2} q_i(p)^2 - \hat{T}_i \left(-q_i - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j} p \right)$$

where the constant $\hat{c}_i$ and function $\hat{T}_i$ are to be derived.

To derive them, we consider trader i's demand submission problem. For each price p, it must be optimal to purchase $\hat{q}_i(p)$ units: a deviation to purchasing $\hat{q}_i(p) + \Delta$ units must be suboptimal for all $\Delta \neq 0$. Thus $\Delta = 0$ must solve

$$\max_{\Delta} -\frac{1}{2\kappa_i}(Z_i + \hat{q}_i + \Delta)^2 - \left(p + \Delta\frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j}\right)(\hat{q}_i + \Delta) \\ + \hat{T}_i\left(-\hat{q}_i - \Delta - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j}\left(p + \Delta\frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j}\right)\right) - \frac{\hat{c}_i}{2}(\hat{q}_i + \Delta)^2$$

where to ease notation we have omitted the argument in $\hat{q}_i$. Taking a first derivative,

$$-\frac{1}{\kappa_i}(Z_i + \hat{q}_i + \Delta) - 2\hat{c}_i(\hat{q}_i + \Delta) - p - \Delta\frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j} - \frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j}(\hat{q}_i + \Delta) \\ - \hat{T}'\left(-\hat{q}_i - \Delta - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j}\left(p + \Delta\frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j}\right)\right)\left(1 + \frac{\kappa_i}{\sum_{j\neq i} \kappa_j}\right) = 0.$$

This must hold at $\Delta = 0$ when $\hat{q}_i$ is as in (22). That is,

$$-\frac{1}{\kappa_i}\left(Z_i - \hat{V}_i(Z_i) - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j}p\right) - \hat{c}_i\left(-\hat{V}_i(Z_i) - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j}p\right) = \\ p + \frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j}\left(-\hat{V}_i(Z_i) - \frac{\kappa_i}{\sum_{j=1}^N \kappa_j}p\right) + \hat{T}_i'(\hat{V}_i(Z_i))\left(1 + \frac{\kappa_i}{\sum_{j\neq i} \kappa_j}\right).$$

Gathering the terms involving only p gives

$$\frac{1}{\sum_{j=1}^N \kappa_j}p + \hat{c}_i\frac{\kappa_i}{\sum_{j=1}^N \kappa_j}p = p - \frac{\kappa_i}{\sum_{j\neq i} \kappa_j}p.$$

Thus we may ensure that these terms are consistent by setting

$$\hat{c}_i = -\frac{1}{\kappa_i} + \frac{\sum_{j=1}^N \kappa_j}{\kappa_i} - \frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j}.$$

We next gather only terms involving Z_i:

$$\hat{V}_i(Z_i)\left(\frac{1}{\kappa_i} + \hat{c}_i + \frac{\sum_{j=1}^N \kappa_j}{\sum_{j\neq i} \kappa_j}\right) - \frac{1}{\kappa_i}Z_i = \hat{T}_i'(\hat{V}_i(Z_i))\left(1 + \frac{\kappa_i}{\sum_{j\neq i} \kappa_j}\right).$$

Writing l in place of $\hat{V}_i(Z_i)$ and integrating both sides we have

$$\hat{T}_i(l) - \hat{T}_i(\hat{V}_i(\underline{Z})) = \left(\sum_{j=1}^{N} \kappa_j \left(\frac{l^2}{2} - \frac{\underline{Z}^2}{2} \right) - \int_{\underline{Z}}^{l} \hat{V}_i^{-1}(s) ds \right) \frac{\sum_{j \neq i} \kappa_j}{\kappa_i \sum_{j=1}^{N} \kappa_j}.$$

Above, $\hat{V}_i^{-1}$ is well-defined as we have perturbed the allocation rule to ensure that $\hat{V}_i$ is strictly monotone.

To ensure global optimality of $\Delta = 0$ under $\hat{T}_i$ and $\hat{c}_i$ it suffices to show that the objective is globally concave. Notice that the integral term is convex since $\hat{V}_i$ is increasing. Thus, ignoring this term, (which is actually not twice-differentiable), by a simple computation (which we omit) we can show that the second derivative of the objective ignoring this term is nonpositive. The objective is therefore globally concave in Δ.

Thus we have achieved an indirect implementation of the perturbed allocation rule in (21). To derive an exact implementation of (12), we take limits as $\epsilon \to 0$. One can show that in the limit $\hat{T}_i$ converges to T_i pointwise and so by relabeling $\hat{c}_i$ by c_i we obtain the side payment rule in the statement of the theorem. Incentive compatibility must hold at the limit. If this were not true, for ϵ sufficiently small, one can show that incentive compatibility is also violated for the perturbed allocation rule under the perturbed side payment rule which is a contradiction.

□

Conditional on the clearing price, trader i's payment in the double auction is known to him. Thus, the double auction with side payments will also implement the same allocation if instead the trader had CARA utility with risk aversion coefficient $1/(2\kappa_i)$ if the distribution of the asset's payoff is standard Gaussian. An interesting feature of the implementing side payment rule is that it is not differentiable: T_i has a kink at η_i. This discontinuity occurs at a point corresponding to when traders transition from being expected sellers to expected buyers. When the market is large and the law of large numbers kicks in, the kink resembles, to some extent, a bid-ask spread.

In practice, there are trading mechanisms which resemble double auction mechanisms with side payment rules. For instance, make-take and take-make fees are widely used in lit exchanges. When viewing the double auction as an idealization of the limit-order book, Theorem 2 offers a microfoundation for these. That is, the optimal trading mechanisms for the preferences and information structure often assumed in the market microstructure literature do resemble, to some degree, double auction mechanisms which are widely used in practice. An implication of Theorem 2 is also that the revenue and efficiency of the double auction mechanism may be sensitive to the particular side payments used. Thus there may be cause for regulation of these side payments or on the flipside, opportunities to use them as regulatory tools as well.

An unattractive feature of Theorem 2 is that the implementing side payment rules are sensitive to the fine details of the environment such as traders' risk capacities or the distributions over their endowments. This is to be expected. Even in the classic single item auction setting of Myerson (1981) the reservation prices for optimal auctions depend on the distributions over buyers' private types. A takeaway from the classic result of Myerson (1981) is that it is optimal for a seller to fix a standard auction format and focus only on calibrating the reservation price. Similarly, for the present model, an exchange operator may without loss fix a double auction mechanism, and only focus on calibrating the side payment rules. Of course, side payment rules are functions and thus much higher-dimensional objects than reservation prices. But in my view this seems to be a meaningful simplification of the design problem.

In the past few decades, the trading landscape for financial assets has become more complex and fragmented. Trade is now split across over a dozen lit exchanges, thirty dark pools, and across many dealers in over-the-counter markets. Fragmentation is in part a consequence of regulatory policies such as Reg NMS in the US and MiFid II in Europe. A growing body of literatures studies the implications of fragmentation (Chen and Duffie, 2021; Rostek and Yoon, 2021; Wittwer, 2021; Antill and Duffie, 2021). When allowing for side payment rules, and for the special (though widely assumed) modeling environment we consider, Theorem 2 implies that a centralized double auction mechanism is always at least as good as a fragmented market for any reasonable objective that balances revenue and allocative efficiency for some side payment rule. However, it is important to note that exchange operators will have different objectives than a social planner and will likely place higher weight on revenue. Fragmentation may be beneficial since competition among exchange operators may lead them to place more weight on allocative efficiency and set side payment rules which are more favorable to traders.

8. Conclusion

We have derived optimal exchange direct mechanisms which maximize a convex combination of revenue and allocative efficiency for an economic environment widely assumed in the financial market microstructure literature. The majority of this literature fixes the trading mechanism and studies various properties of the resulting equilibrium. This paper brings the perspective of mechanism design. We have given a sharp characterization of optimal mechanisms and presented a simple analytical example describing how their properties vary with the number of traders and the weight placed on revenue versus allocative efficiency in the objective. In the example, we the regions where participation constraints bind decreases in size as more weight is placed on allocative efficiency in the objective. We have also shown that optimal mechanisms can be implemented indirectly via double auctions with side payments. Thus side payments can lead to a wide range of outcomes and can be calibrated to attain any point on the efficiency-revenue frontier. There may therefore be cause to regulate side payments or to use them as effective policy tools. An important direction for future work is to study the effects of competition on the side payment rules set by exchanges.

A. **Omitted Proofs**

We first prove the following Lemma which shows that the relaxed version of problem (6) has a unique solution.

Lemma 4. *There exists a unique solution (within the set of equivalence classes on $\mathcal{L}^2(\mathbb{R}^N) \times \mathbb{R}_+^N$) to the relaxed version of problem* (6) *which ignores the constraint* (10).

Proof. Without loss of generality, we may consider only allocation rules $(X_i)_{i=1}^N$ which lie in $\mathcal{L}^2(\mathbb{R}^N)$. This is because the assumption that $f_i > 0$ ensures that the underlying probability measure and Leb($\mathbb{R}^N$) are mutually absolutely continuous and because the terms which are quadratic enter negatively into the objective. Thus any $(X_i)_{i=1}^N \notin \mathcal{L}^2(\mathbb{R}^N)$ leads to $-\infty$ for the objective. One might be tempted to change variables and in place of $U_i(\underline{Z})$ define a new control by $u_i = \sqrt{U_i(\underline{Z})}$. Then, by completing the square, the objective corresponds to the distance between two points on $\mathcal{L}^2(\mathbb{R}^N) \times \mathbb{R}_+^N$ where $\mathbb{R}_+$ denotes the nonnegative reals. One might then seek to apply the Hilbert projection theorem for $\mathcal{L}^2(\mathbb{R}^N) \times \mathbb{R}_+^N$ to prove Lemma 4. Unfortunately, the domain is no longer convex and so this is not possible. However, the same basic proof strategy does work with some minor alterations.

Rewrite the relaxed version of problem (6) in the form

$$\min_{(x,m)\in V} ||x-y||^2_{\mathcal{L}^2(\mathbb{R}^N)} + m \cdot e$$

where $V \subset \mathcal{L}^2(\mathbb{R}^N) \times \mathbb{R}_+^N$ is closed and convex and e denotes the vector of ones in $\mathbb{R}_+^N$. Let

$$I = \inf_{(x,m)\in V} ||x-y||^2_{\mathcal{L}^2(\mathbb{R}^N)} + m \cdot e.$$

Then there exists a sequence $(x_n, m_n)_{n=1}^\infty$ contained in V such that

$$||x_n - y_n||^2_{\mathcal{L}^2(\mathbb{R}^N)} + m_n \cdot e \to I. \tag{23}$$

Moreover, one can label things such that for n, l sufficiently large,

$$||x_n - y_n||^2_{\mathcal{L}^2(\mathbb{R}^N)} + m_n \cdot e + ||x_l - y_l||^2_{\mathcal{L}^2(\mathbb{R}^N)} + m_l \cdot e - 2I \leq \frac{1}{n} + \frac{1}{l}.$$

Since $\mathcal{L}^2(\mathbb{R}^N)$ is a Hilbert space, the polarization identity holds:

$$||x+y||^2_{\mathcal{L}^2(\mathbb{R}^N)} + ||x-y||^2_{\mathcal{L}^2(\mathbb{R}^N)} = 2\left(||x||^2_{\mathcal{L}^2(\mathbb{R}^N)} + ||y||^2_{\mathcal{L}^2(\mathbb{R}^N)}\right)$$

for any $x, y \in \mathcal{L}^2(\mathbb{R}^N)$. Using this fact, it is easy to see that

$$||x_n - x_l||^2 + 4||\frac{x_n + x_l}{2} - y||^2_{\mathcal{L}^2(\mathbb{R}^N)} = 2\left(||x_n - y||^2_{\mathcal{L}^2(\mathbb{R}^N)} + ||x_l - y||^2_{\mathcal{L}^2(\mathbb{R}^N)}\right).$$

We therefore have

$$\frac{1}{2}||x_n - x_l||^2_{\mathcal{L}^2(\mathbb{R}^N)} + m_n \cdot e + 2||\frac{x_n + x_l}{2} - y||^2_{\mathcal{L}^2(\mathbb{R}^N)} + m_l \cdot e - 2I \leq \frac{1}{n} + \frac{1}{l}.$$

Since V is convex it contains $\left(\frac{x_n+x_l}{2}, \frac{m_n+m_l}{2}\right)$ and so by the definition of I, it follows that

$$\frac{1}{2}||x_n - x_l||^2_{\mathcal{L}^2(\mathbb{R}^N)} \leq \frac{1}{n} + \frac{1}{l}.$$

Thus $(x_n)_{n=1}^{\infty}$ is Cauchy and must converge. This then implies that $(m_n)_{n=1}^{\infty}$ must also converge as otherwise (23) would be violated. (Here m_n represents the vector of $U_i(\underline{Z})$ possibly scaled by κ_i to fit into the framework. But, $U_i(\underline{Z})$ must be calibrated so that the participation constraint (at least nearly) binds which is determined by the allocation rule by (9). Thus m_n must converge if x_n converges.) Since V is closed and convex, it follows that the limit point (x, m) is contained in V. Thus a solution exists and we are in fact justified in writing a minimum in front of the objective. Uniqueness follows by inspecting the objective. If two distinct points attain the minimum value they must do so with different allocation rules. One could then take the convex combination of the two and obtain a strict improvement which is a contradiction. □

It is clear that we can prove existence and uniqueness for problem (6) using the same method by redefining V to additionally incorporate the monotonicity constraint. However, Lemma 4 is more directly useful for the proof of Theorem 1.

Completing the proof of Theorem 1. Consider the problem

$$\min f(x) \text{ such that } x \in \Omega, G(x) \leq \theta \tag{24}$$

where Ω is a convex subset of a vector space X, f is a real-valued convex functional on Ω, and G is a convex mapping from Ω into a normed space Z. Let P be a convex cone in Z. For $z, y \in Z$ write $z \geq y$ if $z - y \in P$. The cone, P, which defines this relation, is called the *positive cone* in Z.

The relaxed version of problem (6) which omits the monotonicity constraint fits into this framework with the following translation.

1. Let
$$f := -\left[\alpha R\left((X_i, U_i(\underline{Z}))_{i=1}^N\right) + (1-\alpha)A\left((X_i)_{i=1}^N\right)\right].$$

2. Ω is the set $(X_i, U_i(\underline{Z}))_{i=1}^N$ such that $\sum_{i=1}^N X_i = 0$ where each X_i is a measurable function on $[\underline{Z}, \overline{Z}]$.

3. G is the convex function which maps
$$(X_i, U_i(\underline{Z}))_{i=1}^N \in \Omega \mapsto \left(\int_{\underline{Z}}^{Z_i} \frac{1}{\kappa_i}\mathbb{E}\left[X_i(s, Z_{-i})\right] ds - U_i(\underline{Z}) - \frac{1}{2\kappa_i}\underline{Z}^2, -R\left((X_i, U_i(\underline{Z}))_{i=1}^N\right)\right)$$

4. Z is the set $\mathcal{C}[\underline{Z}, \overline{Z}] \times \mathbb{R}$.

5. $\theta \in Z$ is $(0, 0)$ where, with abuse of notation, $0 \in \mathcal{C}[\underline{Z}, \overline{Z}]$ denotes the constant function which is zero everywhere on $[\underline{Z}, \overline{Z}]$.

6. The dual Z^* is the set $\text{BV}[\underline{Z}, \overline{Z}] \times \mathbb{R}$ where $\text{BV}[\underline{Z}, \overline{Z}]$ is the set of functions of bounded variation on $[\underline{Z}, \overline{Z}]$.

7. P is the set $\mathcal{C}_+[\underline{Z}, \overline{Z}] \times \mathbb{R}_+$ where $\mathcal{C}_+[\underline{Z}, \overline{Z}]$ is the set of nonnegative functions in $\mathcal{C}[\underline{Z}, \overline{Z}]$ and $\mathbb{R}_+$ is the set of nonnegative real numbers.

By inspection, P is nonempty, closed, and contains an interior point. We can also see that there exists a point $x \in \Omega$ such that $G(x) < 0$.[9] Namely, let $x = \left((X_i, U_i(\underline{Z}))_{i=1}^N\right)$ correspond to the candidate revenue optimal mechanism (or any mechanism that generates positive revenue) with transfers reduced by ϵ for some $\epsilon > 0$ small for each trader i. Then $G(x) < 0$.

To complete the proof of Theorem 1, we appeal to the following two theorems, which originally appear in Luenberger (1997). Theorem 3 states that a necessary condition for x_0 to solve (24) is that x_0 minimizes the Lagrangian for some Lagrange multipliers z_0^* in Z^* such that the complementary slackness conditions hold.

Theorem 3 (restatement of Theorem 1 in Luenberger (1997)). *Let X be a linear vector space, Z a normed space, Ω a convex subset of X, and P the positive cone in Z. Assume that P contains an interior point.*

[9] The existence of such a point is why we incorporate the constraint that total trade quantity nets to zero into the definition of Ω rather then include it in the Lagrangian (11).

Let f be a real-valued convex functional on Ω, and G a convex mapping from Ω into Z. Assume the existence of a point $x_1 \in \Omega$ for which $G(x_1) < \theta$ (i.e., $G(x_1)$ is an interior point of $N = -P$).

Let

$$\mu_0 = \inf f(x) \text{ subject to } x \in \Omega, G(x) \leq \theta$$

and assume μ_0 is finite. Then there is an element $z_0^* \geq \theta$ in Z^* such that

$$\mu_0 = \inf_{x \in \Omega} \{f(x) + \langle G(x), z_0^* \rangle\}. \tag{25}$$

Furthermore, if the infimum is achieved in (24) by an $x_0 \in \Omega$, $G(x_0) \leq \theta$ it is achieved by x_0 in (25) and

$$\langle G(x_0), z_0 \rangle = 0. \tag{26}$$

Theorem 4 states that it is sufficient that (x_0, z_0^*) is a saddle point of the Lagrangian for x_0 to solve (24).

Theorem 4 (restatement of Theorem 2 in Luenberger (1997)). *Let P denote the positive cone of Z and suppose P is closed. Suppose there exists $z_0^* \in Z^*$, $z_0^* \geq \theta$, and an $x_0 \in \Omega$ such that the Lagrangian $L(x, z^*) = f(x) + \langle G(x), z^* \rangle$ possess a saddle point at x_0, z_0^*. That is,*

$$L(x_0, z^*) \leq L(x_0, z_0^*) \leq L(x, z_0^*)$$

for all $x \in \Omega$, $z_0^* \geq \theta$. Then x_0 solves (24).

In the main text, we constructed the allocation rule and Lagrange multipliers such that the allocation rule maximizes the Lagrangian (11) with those multipliers, the complementary slackness conditions hold, and is such that the allocation rule is weakly decreasing. The allocation rule and multipliers we constructed form a saddle point of the Lagrangian.

Thus by Theorem 3, Theorem 4, and Lemma 4 there is a unique solution to the relaxed problem which happens to be monotone. (Note that we are using Theorem 3 and Lemma 4 to ensure the existence of a solution to the system of equations in Theorem 1) This must therefore be the unique solution of the original problem (6).

□

Linear system in Corollary 1.1.

$$Z_{ia} + (\alpha + \lambda)(Z_{ia} - \underline{Z}) = \eta_i. \tag{27}$$

$$Z_{ia} + (\alpha + \lambda)\left(Z_{ia} - \underline{Z}\right) = Z_{ib} - (\alpha + \lambda)\left(\overline{Z} - Z_{ib}\right). \tag{28}$$

$$\eta_i \frac{(1+\lambda)\sum_{j\neq i}\kappa_j}{\kappa_i} = \sum_{j\neq i}\left(-\int_{Z_{ja}}^{Z_{jb}} \Omega_j(s)ds + \big(1-(\alpha+\lambda)\big)\frac{\underline{Z}+\overline{Z}}{2} + (\alpha+\lambda)(Z_{jb} - \underline{Z})\right). \tag{29}$$

$$\eta_i \frac{\left(Z_{ib} - Z_{ia}\right)}{\overline{Z} - \underline{Z}}(1+\lambda) = Z_{ib} - Z_{ia} - \frac{1-(\alpha+\lambda)}{\overline{Z}-\underline{Z}}\left(\frac{(Z_{ib}-Z_{ia})(Z_{ib}+Z_{ia})}{2} + Z_{ib} - Z_{ia}\right) - \int_{Z_{ia}}^{Z_{ib}} \Omega_i(s)ds. \tag{30}$$

Note that (28) implies that

$$Z_{ib} - Z_{ia} = \frac{(\overline{Z} - \underline{Z})(\alpha+\lambda)}{1+\alpha+\lambda}$$

so the system is linear. □

References

Antill, S., and D. Duffie (2021): “Augmenting markets with mechanisms,” *The Review of Economic Studies*, 88(4), 1665–1719.

Biais, B., D. Martimort, and J.-C. Rochet (2000): “Competing mechanisms in a common value environment,” *Econometrica*, 68(4), 799–837.

Chen, D., and D. Duffie (2021): “Market fragmentation,” *American Economic Review*, 111(7), 2247–74.

Colliard, J.-E., and T. Foucault (2012): “Trading fees and efficiency in limit order markets,” *The Review of Financial Studies*, 25(11), 3389–3421.

Cramton, P., R. Gibbons, and P. Klemperer (1987): “Dissolving a partnership efficiently,” *Econometrica: Journal of the Econometric Society*, pp. 615–632.

Crémer, J., and R. P. McLean (1988): “Full extraction of the surplus in Bayesian and dominant strategy auctions,” *Econometrica: Journal of the Econometric Society*, pp. 1247–1257.

Du, S., and H. Zhu (2017a): “Are CDS auctions biased and inefficient?,” *The Journal of Finance*, 72, 2589–2628.

——— (2017b): “What is the Optimal Trading Frequency in Financial Markets?,” *The Review of Economic Studies*, 84(4), 1606–1651.

Dworczak, P., S. D. Kominers, and M. Akbarpour (2021): “Redistribution through markets,” *Econometrica*, 89(4), 1665–1698.

Foucault, T., O. Kadan, and E. Kandel (2013): “Liquidity cycles and make/take fees in electronic markets,” *The Journal of Finance*, 68(1), 299–341.

Jantschgi, S., H. H. Nax, B. Pradelski, and M. Pycia (2022): “Double Auctions and Transaction Costs,” .

Lu, H., and J. Robert (2001): “Optimal trading mechanisms with ex ante unidentified traders,” *Journal of Economic Theory*, 97(1), 50–80.

Luenberger, D. G. (1997): *Optimization by vector space methods*. John Wiley & Sons.

Malamud, S., and M. Rostek (2017): “Decentralized Exchange,” *American Economic Review*, 107(11), 3320–3362.

MALINOVA, K., AND A. PARK (2015): “Subsidizing liquidity: The impact of make/take fees on market quality,” *The Journal of Finance*, 70(2), 509–536.

MCAFEE, R. P. (1991): “Efficient allocation with continuous quantities,” *Journal of Economic Theory*, 53(1), 51–74.

MYERSON, R. B. (1981): “Optimal auction design,” *Mathematics of operations research*, 6(1), 58–73.

MYERSON, R. B., AND M. A. SATTERTHWAITE (1983): “Efficient mechanisms for bilateral trading,” *Journal of economic theory*, 29(2), 265–281.

ROSTEK, M., AND J. H. YOON (2021): “Exchange design and efficiency,” *Econometrica*, 89(6), 2887–2928.

SANNIKOV, Y., AND A. SKRZYPACZ (2016): “Dynamic Trading: Price Inertia and Front-Running,” Working paper, Graduate School of Business, Stanford University.

TATUR, T. (2005): “On the trade off between deficit and inefficiency and the double auction with a fixed transaction fee,” *Econometrica*, 73(2), 517–570.

WILSON, R. (1985): “Efficient trading,” in *Issues in contemporary microeconomics and welfare*, pp. 169–208. Springer.

WITTWER, M. (2021): “Connecting disconnected financial markets?,” *American Economic Journal: Microeconomics*, 13(1), 252–82.

ZHANG, A. L. (2020): “Competition and Manipulation in Derivative Contract Markets,” SSRN Scholarly Paper ID 3413265, Social Science Research Network, Rochester, NY.