Bilateral Trading with Maxmin Agents*

Alexander Wolitzky
Stanford University
August 1, 2013

Abstract

This paper revisits the classical bilateral trading problem of Myerson and Satterthwaite (1983) under the assumption that agents disagree about the distribution of values, in that each is maximally pessimistic about the other’s distribution subject to knowing its mean and upper and lower bounds on its support, as well as possibly additional information (which is the information structure that emerges when agents start with a common prior but are pessimistic about what information the other might acquire before participating in the mechanism). An exact characterization of the parameters for which efficient trade is possible is obtained, when the mechanism is allowed to run a surplus. Simple reference rules (which do not run a surplus) are efficient trading mechanisms when the average types of each agent do not have strict gains from trade with each other. In other cases, more sophisticated mechanisms are required, while in others still the Myerson-Satterthwaite impossibility result persists.

*For helpful comments, I thank Pierpaolo Battigalli, Aaron Bodoh-Creed, Subir Bose, Gabe Carroll, Aytek Erdil, Ben Golub, Fuhito Kojima, Massimo Marinacci, Paul Milgrom, Suresh Mutuswami, Ilya Segal, Andy Skrzypacz, and participants in several seminars. I thank Jason Huang for excellent research assistance.
1 Introduction

It has been long been recognized that uncertainty on the part of economic agents about each other’s valuation of a good can preclude efficiency in trade or in the provision of public goods. Standard models of trade under uncertainty assume that agents are Bayesian, and typically also assume that agents’ beliefs are derived from a common prior, often with independent types. The canonical result in this setting is the theorem of Myerson and Satterthwaite (1983), which states that with a common prior, independent types, and uncertain gains from trade, no ex post efficient (E), interim incentive compatible (IC), interim individually rational (IR), and ex post budget balanced (BB) trading mechanism exists. Their elegant proof of this result uses the envelope theorem to show that the sum of the lowest-valuation agents’ expected utilities must be negative in any mechanism satisfying E, IC, and BB.

This paper revisits the classical Myerson-Satterthwaite setting under an alternative assumption on agents’ behavior. Specifically, I assume for most of the paper that each agent knows at least the mean of the distribution of her opponent’s value, as well as upper and lower bounds on its support, and is maximally pessimistic (“maxmin,” “ambiguity-averse”) about this distribution subject to knowing the mean and bounds (as well as possibly knowing some additional information). There are two related but distinct motivations for this approach. First, as I explain below, this information structure is the one that emerges when agents have a common prior on values at an ex ante stage and are maxmin about the information that the other might acquire before entering the mechanism. This may sometimes be an appealing alternative to the Bayesian approach of specifying a prior over the set of experiments that one’s opponent may have access to, especially when this set is large (e.g., consists of all possible experiments) or the agents’ interaction is one-shot.¹ Second, precisely because Myerson and Satterthwaite’s proof is so elegant in how the various terms in the agents’ expected utilities cancel out, their setting may be one where relaxing the common prior

¹More broadly, maxmin agents have a long pedigree in economics, dating back to the axiomatizations of Milnor (1954) and Arrow and Hurwicz (1972), with the modern ambiguity-aversion literature dating to Gilboa and Schmeidler (1989). While my exposition emphasizes maxmin (“ambiguity-averse”) agents, similar results hold with minmax (“Bayesian pessimistic”) agents, so it is not really necessary for my approach that agents are non-Bayesian (see Section 5.4).
assumption is particularly appealing. An interesting conclusion of this paper will be that, despite this apparent dependence on a common prior, the Myerson-Satterthwaite theorem sometimes continues to hold when agents are maxmin about their opponents’ information acquisition technology as described above—but sometimes not.

In particular, the first main result of the paper completely characterizes when efficient trade is possible, when the mechanism is allowed to run a surplus (weak budget balance (WBB)). In a standard bilateral trading setting where the range of possible seller and buyer values is $[0, 1]$, the average seller value is $c^*$, and the average buyer value is $v^*$, Figure 1 indicates the combination of parameters $(c^*, v^*)$ for which a mechanism satisfying E, IC, IR, and WBB exists. Above the curve—the formula for which is

$$\frac{c^*}{1 - c^*} \log \left( 1 + \frac{1 - c^*}{c^*} \right) + \frac{1 - v^*}{v^*} \log \left( 1 + \frac{v^*}{1 - v^*} \right) = 1$$

—the Myerson-Satterthwaite theorem persists, despite the lack of a common prior or independent types, even when the mechanism is allowed to run a surplus. Below the curve, the Myerson-Satterthwaite theorem fails.

I call the mechanism that allows for efficient trading for all parameters below the curve in Figure 1 the screening mechanism. It is defined by specifying that the agents trade when the buyer values the good more than the seller, and that the price is increasing in both the buyer’s and seller’s reported values in a particular way. Specifically, the mechanism has the property that the “worst-case” belief of a buyer who reports value $\hat{v}$ about the distribution of the seller’s value is that the seller’s value is always either $\hat{v}$ (the most favorable value for which there are no gains from trade) or 0 (the most favorable value possible). If the buyer shades $\hat{v}$ down to try to get a better price, the requirement that the expected seller value is fixed at $c^*$ forces the buyer’s worst-case belief to put more weight on $\hat{v}$ and less weight on 0, reducing his expected probability of trade. Since reducing the probability of trade is more costly for higher-value buyers, the mechanism successfully screens buyers (when

---

2 This is in line with Gilboa’s exhortation in his monograph on decision making under uncertainty to “[consider] the MMEU [Maxmin Expected Utility] model when a Bayesian result seems to crucially depend on the existence of a unique, additive prior, which is common to all agents. When you see that, in the course of some proof, things cancel out too neatly, this is the time to wonder whether introducing a little bit of uncertainty may provide more realistic results,” (Gilboa, 2009, p.169).
Figure 1: With weak budget balance, efficient trade is possible in the region below the curve and impossible in the region above it.

prices are set so that truth-telling satisfies buyers’ first-order conditions. Sellers are treated symmetrically, and it turns out that the resulting mechanism runs a surplus precisely when \((c^*, v^*)\) lies below the curve in Figure 1.

Conversely, when \((c^*, v^*)\) lies above the curve in Figure 1, no mechanism satisfies E, IC, IR, and WBB. The proof adapts the proof of the Myerson-Satterthwaite theorem, with some significant differences. The first conceptual difference is that the appropriate envelope condition for payoffs is no longer “derivative of utility with respect to type equals probability of trade,” but rather “derivative of utility with respect to type is at least the smallest possible probability of trade.”\(^3\) This smallest possible probability of trade equals the probability of trade under the worst-case belief in the screening mechanism, which explains why the screening mechanism gives an “if and only if” characterization of when trade is possible. The second conceptual difference is that there is no longer a common prior under which it can be shown that the sum of the lowest-valuation agents’ expected utilities must be negative under any mechanism satisfying E, IC, and WBB. However, since agents are

\(^3\)This is related to results of Bodoh-Creed (2012) and Milgrom and Segal (2002), but differs in that the maxmin problem agents face in the current paper may not be a saddle point problem.
maximin, their subjective expected utilities are lower than their expected utilities would be under any common prior, and the result is proved by choosing a benchmark common prior appropriately.

The second main result of the paper takes a complementary approach: rather than asking when efficient trade can be implemented by any mechanism, it asks when efficient trade can be implemented by a particularly simple kind of mechanism, called a reference rule. A reference rule works by setting a “reference price” $p^*$ and specifying that if the reported values lie on opposite sides of $p^*$, then the price is $p^*$; while if the reported values lie on the same side of $p^*$, then the price is whichever report is closer to $p^*$ (i.e., price is $v$ if $c \leq v \leq p^*$; price is $c$ if $p^* \leq c \leq v$). There are four reasons why this alternative approach is of interest. First, the condition characterizing when efficient trade is possible with reference rules is simpler and easier to interpret than the corresponding condition for general mechanisms: efficient trade can be implemented with a reference rule if and only if $c^* \geq v^*$ (which defines in the region below the 45° line in Figure 1); that is, if and only if the average types of each agent do not have strict gains from trade with each other. Second, reference rules are exactly budget balanced (i.e., they do not run a surplus). Third, the characterization result for general mechanism requires the assumption that a buyer who reports value $\hat{v}$ (say) finds it possible that the seller’s value is always either $\hat{v}$ or 0, while the characterization result for references rules does not require this. Fourth, reference rules have some established market design relevance, having been proposed by Erdil and Klemperer (2010) as practical payment rules for core-selecting package auctions. Erdil and Klemperer show that, although reference rules are not Bayesian incentive compatible in their setting, they perform well in terms of the sum of types’ “local incentives to deviate.” From this perspective, this paper complements Erdil and Klemperer’s emphasis on reference rules’ practical appeal and robustness to local deviations by demonstrating a sense in which reference rules can be exactly incentive compatible.

The intuition for why reference rules are incentive compatible when $c^* \geq v^*$ and $p^* \in [v^*, c^*]$ is captured in Figure 2. Observe that every buyer with value $v \leq c^*$ may be certain that no gains from trade exist, as he may believe that the distribution of seller values is the degenerate distribution on $c^*$ (because this distribution has mean $c^*$). Hence, certainty
Figure 2: Reference rules with \( p^* \in [v^*, c^*] \) are incentive compatible.

of no-trade is a worst-case belief for these buyers, and they are therefore willing to reveal their values truthfully.\(^4\) In contrast, buyers with value \( v > c^* \) do believe that gains from trade exist with positive probability. But it is optimal for these buyers to reveal their values truthfully as well: misreporting some \( \hat{v} > c^* \) does not affect the price regardless of the seller’s value (as price equals \( c \) if \( c > p^* \) and equals \( p^* \) if \( c \leq p^* \)), and misreporting some \( \hat{v} \leq c^* \) again gives payoff 0 in the worst-case (as certainty that the seller’s value equals \( c^* \) would again be a worst-case belief). Therefore, truth-telling is optimal for every buyer type, and the argument for sellers is symmetric. However, this argument breaks down if \( p^* \notin [v^*, c^*] \), which explains why efficient trade with reference rules is impossible when \( c^* < v^* \).\(^5\)

This paper joins a small but growing literature on games and mechanisms with maxmin (“ambiguity-averse”) agents. Most closely related are three recent papers that discuss bilateral trade with ambiguity-averse agents: Bodoh-Creed (2012), de Castro and Yannelis (2010), and Bose and Mutuswami (2012). Bodoh-Creed and de Castro and Yannelis both

\(^4\)A potential concern here is that agents who do not expect to trade are indifferent over a wide range of reports. A partial response is given in Section 5.3, which shows how to ensure that truth-telling is not weakly dominated for any type. A second response is that there need not be “many” types who do not expect to trade (unlike in Figure 2).

\(^5\)A more general intuition for why efficient trade is possible when \( c^* \geq v^* \) is the following: Buyers with value \( v \leq c^* \) and sellers with cost \( c \geq v^* \) may be certain of no-trade and and therefore are willing to reveal their types truthfully. If \( c^* \geq v^* \) then the remaining intervals of buyers (i.e., those with \( v > c^* \)) and sellers (i.e., those with \( c < v^* \)) do not overlap, so they may trade at a fixed price in between \( v^* \) and \( c^* \). In other words, when \( c^* \geq v^* \), “certainty of no-trade types” and “trade at fixed price types” cover the type space, but when \( c^* < v^* \) they do not.
point out that with no restrictions on agents’ beliefs about the distribution of their opponents’
values, efficient trade is always possible. The proof is simple: when all beliefs are possible,
certainty of no-trade is a worst-case belief for every agent, so all agents are willing to reveal
their values truthfully. In contrast, Bose and Mutuswami introduce a “small” amount of
ambiguity into the Myerson-Satterthwaite setting and show that efficient trade can then be
achieved using dynamic mechanisms that leverage the time-inconsistency of ambiguity-averse
agents. Their main idea there is clearly quite different from mine, as I retain a standard
Nash implementation (i.e., direct mechanism) perspective. Relative to all three of these
papers, it is noteworthy that in my model the Myerson-Satterthwaite theorem sometimes
continues to hold, despite the lack of a common prior or independent types. The mechanisms
proposed in the current paper are also very different from those in this literature.

A second related literature concerns the possibility of full surplus extraction in mecha-
nism design when agents have a common prior ex ante but have different beliefs at the
interim stage due to correlated types or information acquisition. The seminal papers here
are Crémer and McLean (1985, 1988) and McAfee and Reny (1992) (who show that the
Myerson-Satterthwaite theorem fails with correlated types). Bikhchandani (2010) considers
information acquisition. Lopomo, Rigotti, and Shannon (2009) show that full surplus extrac-
tion may be impossible when agents have incomplete preferences à la Bewley (1986). The
form of belief differences as well as the analysis involved in this literature is quite different
from that in the current paper.

The paper proceeds as follows. Section 2 presents the model. Sections 3 and 4 give the
characterization results for general mechanisms and reference rules, respectively. Section 5

---

6Indeed, this observation is not the main result of either of their papers. Bodoh-Creed establishes an
envelope condition for payoffs with maxmin agents, related to my Lemma 1 below. Among other results, he
also provides some comparative statics on optimal bilateral trading mechanisms with respect to the amount
of ambiguity, for versions of these concepts that differ from what I consider here. De Castro and Yannelis
establish a general equivalence between ex post efficiency and incentive compatibility when agents’ beliefs
are completely unrestricted.

7Other less closely related papers on mechanism design with ambiguity-averse agents include Salo and
Bose and Daripa (2009), Bose and Renou (2012), and Di Tillio, Kos, and Messner (2012).
contains several extensions of the model and results, including the details of the information acquisition interpretation of the model described above, a discussion of how the results change with minmax ("Bayesian pessimistic") rather than maxmin ("ambiguity-averse") agents, and a description of how both the screening and reference mechanisms can be modified to avoid a reliance on weakly dominated strategies. Section 6 concludes. An appendix contains omitted proofs.

2 Model

Agents and Preferences: Two agents must make a binary social choice, \( y \in \{0, 1\} \) (e.g., \{no trade, trade\}, \{no provision, provision\}). Each agent \( i = 1, 2 \) has a real-valued type \( \theta_i \), which is her value for outcome \( y = 1 \) over \( y = 0 \). Agents have quasilinear utility. Thus, if a type \( \theta_i \) agent receives transfer \( t_i \), her payoff is

\[- \theta_i y + t_i.\]

Agent \( i \)'s value \( \theta_i \) is her private information. Agent \( i \)'s opponent, agent \( j \), knows at least the following features of the distribution of agent \( i \)'s value:\n
1. \( \theta_i \) lies in the interval \( \Theta_i = [\bar{\theta}_i, \tilde{\theta}_i] \).

2. The expectation of \( \theta_i \) is \( \theta_i^* \in (\bar{\theta}_i, \tilde{\theta}_i) \).

Formally, let \( \Delta_i \) be the set of Borel measures \( \phi_i \) on \( \Theta_i \) such that \( E^{\phi_i} [\theta_i] = \theta_i^* \), endowed with the weak* topology. I assume that the set of distributions of agent \( i \)'s value that agent \( j \) finds possible is some nonempty, compact, convex subset \( \Delta'_i \subseteq \Delta_i \). Each agent \( i \) evaluates her expected utility with respect to the worst possible distribution of her opponent’s value among those distributions in \( \Delta'_j \); that is, the agents are maxmin optimizers.

---

\(^8\)Section 5.1 shows that this assumption is appropriate if agents have a common prior on values with mean \((\bar{\theta}_1^*, \bar{\theta}_2^*)\) and support \([\bar{\theta}_1, \bar{\theta}_1] \times [\bar{\theta}_2, \bar{\theta}_2]\) at an ex ante stage and may acquire additional information prior to entering the mechanism. Section 5.2 considers weaker assumptions. The assumption that \( \theta_i^* \) lies in the interior of \( \Theta_i \) is without loss of generality: if \( \theta_i^* \in [\bar{\theta}_i, \bar{\theta}_i] \), then there would be no uncertainty about agent \( i \)'s value, and efficient trade could always be implemented with a Vickrey-Glarke-Groves mechanism.
Two special kinds of distributions $\phi_i$ will play an important role in the analysis. Let $\delta_{\theta_i^*}$ be the Dirac measure at $\theta_i^*$, so that $\delta_{\theta_i^*} \in \Delta'_i$ corresponds to the possibility that agent $i$’s value is $\theta_i^*$ for sure.\footnote{In the information acquisition interpretation of Section 5.1, $\delta_{\theta_i^*} \in \Delta'_i$ corresponds to the possibility that agent $i$ may acquire no new information about her value before entering the mechanism.} Let $\delta_{\theta_i^*,\theta_i^h}$ be the measure consisting of atoms at $\theta_i^*$ and $\theta_i^h$, satisfying $E_{\delta_{\theta_i^*,\theta_i^h}}[\theta_i] = \theta_i^*$; that is, $\delta_{\theta_i^*,\theta_i^h}$ is given by $\theta_i = \theta_i^*$ with probability $\frac{\theta_i^h - \theta_i^*}{\theta_i^h - \theta_i^*}$ and $\theta_i = \theta_i^h$ with probability $\frac{\theta_i^* - \theta_i^h}{\theta_i^h - \theta_i^*}$. Thus, $\delta_{\theta_i^*,\theta_i^h} \in \Delta'_i$ corresponds to the possibility that agent $i$’s value may take on only value $\theta_i^j$ or $\theta_i^h$.$^{10}$ The main results to follow require $\delta_{\theta_i^*} \in \Delta'_i$, and sometimes also require $\delta_{\theta_i^j,\theta_i^h} \in \Delta'_i$ for certain values of $\theta_i^j, \theta_i^h$ (but Section 5.2 relaxes these assumptions).

Mechanisms: A (direct) mechanism $x = (y,t)$ consists of a measurable function $y : \theta_1 \times \theta_2 \rightarrow [0,1]$ and a measurable and bounded function $t : \theta_1 \times \theta_2 \rightarrow \mathbb{R} \times \mathbb{R}$ such that $y(\theta_1,\theta_2)$ is the probability of outcome $y = 1$ and $t(\theta_1,\theta_2)$ is the transfer to agent $i$.

Given a mechanism $x$, let

$$U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right) = \theta_i y\left(\hat{\theta}_i, \theta_j\right) + t_i\left(\hat{\theta}_i, \theta_j\right),$$
$$U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) = \mathbb{E}_{\phi_j} \left[U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)\right],$$
$$U_i\left(\theta_i\right) = \inf_{\phi_j \in \Delta'_j} U_i\left(\theta_i, \phi_j; \theta_i\right).$$

Thus, $U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$ is agent $i$’s utility from reporting type $\hat{\theta}_i$ against type $\theta_j$ given true type $\theta_i$, $U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$ is agent $i$’s expected utility from reporting type $\hat{\theta}_i$ against belief $\phi_j$ given true type $\theta_i$, and $U_i\left(\theta_i\right)$ is agent $i$’s “worst-case” expected utility from reporting her true type $\theta_i$.$^{11}$ A mechanism is maxmin incentive compatible (IC) if

$$\theta_i \in \arg \max_{\hat{\theta}_i \in \Theta_i, \phi_j \in \Delta'_j} \inf_{\theta_j \in \Theta_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) \text{ for all } \theta_i \in \Theta_i.$$ 

I restrict attention to incentive compatible direct mechanisms throughout. The motivation for doing so is the following: Suppose the agents play an arbitrary game. Then every

\footnote{The term “worst-case” is only used heuristically in this paper, but the meaning is generally that if $\min_{\phi_j \in \Delta'_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$ exists, then a minimizer is a worst-case belief; while if the minimum does not exist (which is possible, as $U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$ may not be continuous in $\phi_j$), then a limit point of a sequence $\{\phi_j\}_n$ that attains the infimum is a worst-case belief.}
Nash equilibrium of the game (i.e., every strategy profile where each type maximizes her expected payoff given her opponent’s strategy, where the expectation is taken with respect to the worst possible distribution of her opponent’s types) corresponds to the outcome of an incentive compatible direct mechanism. This is just the standard revelation principle, used here in the same way as in the prior literature on mechanism design with ambiguity-averse agents.\(^{12}\)

Besides incentive compatibility, I also consider the following desiderata.

- **Ex Post Efficiency (E)**: \( y(\theta_1, \theta_2) = 1 \) if \( \theta_1 + \theta_2 \geq 0 \), \( y(\theta_1, \theta_2) = 0 \) if \( \theta_1 + \theta_2 < 0 \).
- **Interim Individual Rationality (IR)**: \( U_i(\theta_i) \geq 0 \) for all \( \theta_i \in \Theta_i \).
- **Ex Post Weak Budget Balance (WBB)**: \( t_1(\theta_1, \theta_2) + t_2(\theta_1, \theta_2) \leq 0 \) for all \( \theta_1 \in \Theta_1, \theta_2 \in \Theta_2 \).
- **Ex Post Strong Budget Balance (SBB)**: \( t_1(\theta_1, \theta_2) + t_2(\theta_1, \theta_2) = 0 \) for all \( \theta_1 \in \Theta_1, \theta_2 \in \Theta_2 \).

I say that efficient trade with (weak, strong) budget balance is possible if there exists a mechanism satisfying E, IC, IR, and (W,S)BB.

In interpreting these conditions, note that the results to follow have both positive and negative parts, so strengthening or weakening any condition will make one direction easier to prove and the other harder. Efficiency is self-explanatory. Interim individual rationality is imposed with respect to the agents’ own worst-case beliefs. All results also hold with ex post individual rationality (i.e., \( U_i(\theta_i, \theta_j; \theta_i) \geq 0 \) for all \( \theta_i \in \Theta_i, \theta_j \in \Theta_j \)). The ex post version of budget balance seems appropriate, since there is no common prior. The difference between weak and strong budget balance is that with weak budget balance the mechanism is allowed to run a surplus. For comparison, recall that weak budget balance is all that is needed in the standard proof of the Myerson-Satterthwaite theorem.\(^{13}\)

\(^{12}\)Dekel, Fudenberg, and Levine (2004) illustrate difficulties in providing learning foundations for Nash equilibrium when agents have heterogeneous beliefs. This objection may be less salient in the context of mechanism design, where perhaps the designer can tell the agents how to play.

\(^{13}\)Efficient trade is always possible if the mechanism is allowed to run a deficit: the Vickrey-Clarke-Groves
The model covers the classical bilateral trade and bilateral public good provision settings, as follows. An advantage of the current notation is that it treats the two agents symmetrically.

**Bilateral Trade:** Agent 1 is the seller and agent 2 is the buyer. They can trade \((y = 1)\) or not \((y = 0)\). The seller’s type \(\theta_1\) is \(-1\) times her value of retaining the object (or equivalently her cost of providing it). The buyer’s type \(\theta_2\) is his value of acquiring the object. \(t_1(\theta_1, \theta_2)\) is the price received by the seller, and \(t_2(\theta_1, \theta_2)\) is \(-1\) times the price paid by the buyer.

**Public Good Provision:** Agents 1 and 2 can either share the cost \(C \in \mathbb{R}\) of providing a public good \((y = 1)\) or not \((y = 0)\). An agent’s type \(\theta_i\) is her benefit from the good, net of a benchmark payment of \(\frac{C}{2}\). \(t_i(\theta_i, \theta_j)\) is \(-1\) times what agent \(i\) pays for the good in addition to \(\frac{C}{2}\).

### 3 Efficient Trade with General Mechanisms

This section states and proves my most general result, which gives a complete characterization of when efficient trade is possible under weak budget balance, a richness condition on the sets of possible beliefs, and a mild restriction on the range of possible values.

A preliminary observation is that if no type of some agent \(i\) has gains from trade with the average type of agent \(j\), then efficient trade is always possible. In this case, efficient trade can be implemented by simply giving the entire surplus to agent \(j\): certainty of no-trade is then a worst-case belief for every type of agent \(i\), so truthtelling is trivially optimal for agent \(i\), and truthtelling is optimal for agent \(j\) by the usual Vickrey-Clarke-Groves logic.

**Proposition 1** Assume that \(\bar{\theta}_i + \theta_j^* \leq 0\) and \(\delta_{\theta_j^*} \in \Delta_j^*\) for some \(i \in \{1, 2\}\). Then efficient trade with strong budget balance is possible.

In light of Proposition 1 (the proof of which is in the appendix), the main result of this section considers the non-trivial case where some type of each agent has gains from trade with the average type of the other agent (i.e., \(\bar{\theta}_i + \theta_j^* > 0\) for \(i = 1, 2\)). I also assume the mechanism with \(t_i(\theta_i, \theta_j) = \theta_j\) when \(\theta_i + \theta_j \geq 0\) satisfies E, IR, and dominant strategy incentive compatibility, which is stronger than maxmin incentive compatibility.
that some type of each agent does not have strict gains from trade with the average type of the other agent (i.e., $\theta_i + \theta_j^* \leq 0$ for $i = 1, 2$); this is the type space restriction mentioned above.\footnote{When this restriction is not satisfied, the proof of the negative part of Theorem 1 still gives an impossibility result, but there is a gap between this impossibility result and the positive part of Theorem 1.} Put together, the assumptions that $\bar{\theta}_i + \theta^*_j > 0$ and $\theta_i + \theta^*_j \leq 0$ for $i = 1, 2$ say that the intervals $\Theta_i$ and $\Theta_j$ are “wide.” As an example, these assumptions hold in the bilateral trade setting when the set of possible values for both the seller and the buyer is $[0, 1]$ (regardless of $c^*$ and $v^*$).

The result is the following.

**Theorem 1** Assume that $\bar{\theta}_i + \theta^*_j > 0$ and $\theta_i + \theta^*_j \leq 0$, and that $\delta_{\theta_i, \bar{\theta}_i} \in \Delta'_i$ for all $\theta_i \in [\theta_i, \theta_i^*]$, for $i = 1, 2$.\footnote{Note that $\delta_{\theta_i^*, \bar{\theta}_i} = \delta_{\theta_i}$, so we have $\delta_{\theta_i^*} \in \Delta'_i$.} Then efficient trade with weak budget balance is possible if and only if

$$
\left( \frac{\bar{\theta}_1 + \min \{ \bar{\theta}_2, -\theta_1^* \}}{\theta_1 + \theta_2} \right) \left( \frac{\bar{\theta}_1 - \theta_1^*}{\theta_1^* + \min \{ \theta_2, -\theta_1^* \}} \right) \log \left( 1 + \frac{\theta_1^* + \min \{ \bar{\theta}_2, -\theta_1 \}}{\theta_1 - \theta_1^*} \right) \\
+ \left( \frac{\bar{\theta}_2 + \min \{ \bar{\theta}_1, -\theta_2 \}}{\theta_1 + \theta_2} \right) \left( \frac{\bar{\theta}_2 - \theta_2^*}{\theta_2^* + \min \{ \theta_1, -\theta_2 \}} \right) \log \left( 1 + \frac{\theta_2^* + \min \{ \bar{\theta}_1, -\theta_2 \}}{\theta_2 - \theta_2^*} \right) \geq 1. \quad (*)
$$

Theorem 1 shows that, under mild restrictions, efficient trading between maxmin agents is possible if and only if Condition (*) holds. In other words, the Myerson-Satterthwaite impossibility result persists with maxmin agents if and only if Condition (*) fails. To interpret Condition (*), first set aside the first term in each of the products on the left-hand side (i.e., the $\frac{\bar{\theta}_1 + \min \{ \bar{\theta}_2, -\theta_1 \}}{\theta_1 + \theta_2}$ and $\frac{\bar{\theta}_2 + \min \{ \bar{\theta}_1, -\theta_2 \}}{\theta_1 + \theta_2}$ terms). These terms take care of the possibility that the sets $\Theta_1$ and $\Theta_2$ may not “line up.” For example, they vanish in the bilateral trade setting when the set of possible seller and buyer values is $[0, 1]$. Next, note that each of the remaining products is of the form $\frac{1}{x} \log(1 + x)$, which is decreasing in $x$. In particular, increasing $\theta_i^*$ makes Condition (*) harder to satisfy; a very rough intuition is that increasing $\theta_i^*$ makes agent $j$ more confident that he will trade, which makes shading his report more tempting. Another observation is that Condition (*) always holds when $\theta_1^* + \theta_2^* \leq 0$; that is, when the average types of each agent do not have strict gains from trade with each other (e.g., this is why the curve in Figure 1 lies above the $45^\circ$ line). This follows
because, using the inequality $\log 1 + x \geq \frac{x}{1+x}$, the left-hand side of Condition (*) is at least

$$\frac{\theta_1 - \theta_1^*}{\theta_1 + \theta_2} + \frac{\theta_2 - \theta_2^*}{\theta_1 + \theta_2} = 1 - \frac{\theta_1^* + \theta_2^*}{\theta_1 + \theta_2}.$$ 

This is consistent with Theorem 2 below, which shows that efficient trade is implementable with reference rules when $\theta_1^* + \theta_2^* \leq 0$. In particular, the parameters for which efficient trade is possible with weak budget balance but not with reference rules are precisely those that satisfy Condition (*) but would violate Condition (*) if the $\log 1 + x$ terms were approximated by $\frac{x}{1+x}$. This gives one measure of how restrictive reference rules are.

Before presenting the proof of Theorem 1, I give a heuristic derivation of Condition (*) in the bilateral trade setting when the set of possible seller and buyer values is $[0, 1]$. In this case, Condition (*) becomes

$$\frac{c^*}{1-c^*} \log \left( 1 + \frac{1 - c^*}{c^*} \right) + \frac{1 - v^*}{v^*} \log \left( 1 + \frac{v^*}{1 - v^*} \right) \geq 1$$

(which gives Figure 1 in the introduction). I show that the “screening mechanism” (defined formally in the proof of Theorem 1) runs a surplus if and only if this condition holds. As indicated in the introduction, this mechanism has the property that the worst-case belief of a buyer who reports value $\hat{v}$ puts positive mass only on $c = \hat{v}$ and $c = 0$; that is, his worst-case belief is $\delta_{0,\hat{v}}$.\(^{16}\) Thus, letting $p_B(c, v)$ be the price paid by the buyer given reported values $(c, v)$, the maxmin payoff of a type $v$ buyer who reports type $\hat{v}$ is

$$\frac{\hat{v} - c^*}{\hat{v}} (v - p_B(0, \hat{v})).$$

The first-order condition for truth-telling to be optimal is

$$\frac{\partial}{\partial v} p_B(0, v) = \frac{c^*}{v(v - c^*)} (v - p_B(0, v)).$$

Solving this differential equation for $p_B(0, v)$ yields

$$p_B(0, v) = \frac{v}{v - c^*} (k + c^* \log v),$$

\(^{16}\)This is imprecise. If a type $v$ buyer reports value $\hat{v} < v$, he makes a profit from trading with a type $\hat{v}$ seller, so the belief $\delta_{0,\hat{v}+\varepsilon}$ is worse for him than is $\delta_{0,\hat{v}}$. In this case, the infimum in the maxmin problem is attained by the sequence $\{\delta_{0,\hat{v}+\varepsilon}\}_{\varepsilon \to 0}$.
where \( k \) is a constant of integration. The constant that keeps price bounded as \( v \to c^* \) is
\[
k = -c^* \log c^* ,
\]
which gives
\[
p_B (0, v) = \frac{ve^*}{v - c^*} \log \left( 1 + \frac{v - c^*}{c^*} \right).
\]
A similar argument for sellers gives
\[
p_S (c, 1) = 1 - \frac{(1 - c)(1 - v^*)}{v^* - c} \log \left( 1 + \frac{v^* - c}{1 - v^*} \right).
\]
Thus, the mechanism runs a surplus when \( c = 0 \) and \( v = 1 \) if and only if \( p_B (0, 1) \geq p_S (0, 1) \), which yields (1). While this argument only shows that (1) is a necessary condition for the screening mechanism to run a surplus, it turns out that, when the remaining prices \( p_B (c, v) \) and \( p_S (c, v) \) are filled in so that \( \delta_{0,v} \) and \( \delta_{c,1} \) are indeed worst-case beliefs, it is also sufficient.

The main ideas behind the proof of Theorem 1 are described in the introduction. Rather than trying to give more intuition here, I discuss the conceptual issues in the proof in footnotes along the way.

### 3.1 Proof of Theorem 1: Positive Direction

The screening mechanism is defined by

\[
y (\theta_i, \theta_j) = \begin{cases} 1 & \text{if } \theta_i + \theta_j \geq 0 \\ 0 & \text{if } \theta_i + \theta_j < 0 \end{cases}, \text{ and}
\]
\[
t_i (\theta_i, \theta_j) = \begin{cases} (1 - \alpha_i (\theta_i)) \theta_j - \alpha_i (\theta_i) \min \{ \theta_i, -\theta_j \} & \text{if } \theta_i + \theta_j \geq 0 \\ 0 & \text{if } \theta_i + \theta_j < 0 \end{cases},
\]

for \( i = 1, 2 \), where
\[
\alpha_i (\theta_i) = \begin{cases} \frac{\theta_j - \theta_i^*}{\theta_j + \min \{ \theta_i, -\theta_j \}} \log \left( 1 + \frac{\theta_i + \min \{ \theta_i, -\theta_j \}}{\theta_j - \theta_i^*} \right) & \text{if } \theta_i > -\theta_j^* \\ 1 & \text{if } \theta_i \leq -\theta_j^* \end{cases}.
\]

This mechanism is clearly efficient. I show that it satisfies IR and IC, and that it satisfies WBB if and only if Condition (*) holds.

**Claim 1:** The screening mechanism satisfies (ex post) IR.
Proof: If $\theta_i + \theta_j < 0$, then $U_i(\theta_i, \theta_j; \theta_i) = 0$. If $\theta_i + \theta_j \geq 0$, then

$$U_i(\theta_i, \theta_j; \theta_i) = \theta_i + (1 - \alpha_i(\theta_i)) \theta_j - \alpha_i(\theta_i) \min \{\theta_i, -\theta_j\}$$

$$\geq (1 - \alpha_i(\theta_i)) (\theta_i + \theta_j).$$

Now $\alpha_i(\theta_i)$ is of the form $\frac{1}{x} \log (1 + x)$ for $x > 0$, and $\frac{1}{x} \log (1 + x) \in (0, 1)$ for $x > 0$, so $\alpha_i(\theta_i) \in (0, 1)$ for all $\theta_i$, which yields ex post IR (and hence IR).

Claim 2: The screening mechanism satisfies IC.

Proof: Suppose $\theta_i \leq -\theta_j^*$. By IR, $U_i(\theta_i) \geq 0$. By ex post IR and WBB, $t_i(\hat{\theta}_i, \theta_j) \leq \theta_j$ for all $\hat{\theta}_i, \theta_j$, and therefore $U_i(\hat{\theta}_i, \theta_j^*; \theta_i) \leq \max \{\theta_i + \theta_j^*, 0\} \leq 0$ for all $\hat{\theta}_i$. Hence, $\delta_{\theta_j^*} \in \Delta_j$ implies that $U_i(\theta_i) \geq U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) \geq \inf_{\phi_j \in \Delta_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$, which yields IC.

For the remainder of the proof, suppose $\theta_i > -\theta_j^*$. The next four paragraphs show that no misreport $\hat{\theta}_i$ can be profitable.

If $\hat{\theta}_i > \theta_i$, I claim that $U_i(\hat{\theta}_i, \theta_j; \theta_i) \geq U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$ for all $\theta_j$. The key step is the following observation (proof in appendix).

Lemma 1 In the screening mechanism, $t_i(\theta_i, \theta_j)$ is non-increasing in $\theta_i$ in the region where $\theta_i + \theta_j \geq 0$.

Now, if $\theta_i + \theta_j \leq 0$, then $U_i(\theta_i, \theta_j; \theta_i) = 0$, while ex post IR and WBB imply that $U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right) \leq \max \{\theta_i + \theta_j, 0\} \leq 0$. If instead $\theta_i + \theta_j > 0$, then efficiency and Lemma 1 imply that $U_i(\theta_i, \theta_j; \theta_i) \geq U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$. The claim follows, and therefore $U_i(\theta_i) \geq \inf_{\phi_j \in \Delta_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$.

If $\hat{\theta}_i \leq -\theta_j^*$, then $U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) = 0$. Hence, $\delta_{\theta_j^*} \in \Delta_j$ and ex post IR imply that $U_i(\theta_i) \geq \inf_{\phi_j \in \Delta_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$.

The hardest case is $\hat{\theta}_i \in (-\theta_j^*, \theta_i]$. In this case, I claim that

$$\inf_{\phi_j \in \Delta_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) = \frac{\theta_j^* + \min \{\hat{\theta}_i, -\theta_j\}}{\hat{\theta}_j + \min \{\hat{\theta}_i, -\theta_j\}} \left(\theta_i + t_i\left(\hat{\theta}_i, \theta_j\right)\right).$$

\[^{17}\text{Intuitively, the claim is that } \delta_{\max\{-\theta_i, \theta_j\}} \text{ is a worst-case belief for an agent of type } \theta_i \text{ who misreports as type } \hat{\theta}_i \in (-\theta_j^*, \theta_i].\]
To see that \( \inf_{\phi_j \in \Delta'_i} U_i (\hat{\theta}_i, \phi_j; \theta_i) \) is at least this much, note that \( \delta_{\text{max}\{\hat{\theta}_i - \varepsilon, \theta_j\}} \in \Delta'_i \) for sufficiently small \( \varepsilon \), and that

\[
\lim_{\varepsilon \to 0} U_i (\hat{\theta}_i, \delta_{\text{max}\{\hat{\theta}_i - \varepsilon, \theta_j\}}; \theta_i) = \lim_{\varepsilon \to 0} \frac{\theta_j^* + \min \{ \hat{\theta}_i + \varepsilon, -\theta_j \} }{\hat{\theta}_j + \min \{ \hat{\theta}_i + \varepsilon, -\theta_j \}} \left( \theta_i + t_i \left( \hat{\theta}_i, \theta_j \right) \right).
\]

To see that \( \inf_{\phi_j \in \Delta'_i} U_i (\hat{\theta}_i, \phi_j; \theta_i) \) is at most this much, note that

\[
U_i (\hat{\theta}_i, \phi_j; \theta_i) = \min \left( \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} \right) \left( \theta_j \mid \theta_j \geq -\hat{\theta}_i \right) \geq \min \left( \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} \right) \left( \theta_j \mid \theta_j \geq -\hat{\theta}_i \right)
\]

Thus, if there exists a convergent sequence of opposing beliefs that gives a lower limit payoff than does \( \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} \), it must be that either the limit \( \Pr \left( \theta_j \geq -\hat{\theta}_i \right) \) is lower than \( \frac{\theta_j^* + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} }{\hat{\theta}_j + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \}} \) or the limit \( \Pr \left( \theta_j \geq -\hat{\theta}_i \right) \) is lower than \( \frac{\theta_j^* + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} }{\hat{\theta}_j + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \}} \). The limit \( \Pr \left( \theta_j \geq -\hat{\theta}_i \right) \) cannot be lower than \( \frac{\theta_j^* + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} }{\hat{\theta}_j + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \}} \) by Chebyshev’s inequality.\(^{18}\)

To see that the limit \( \Pr \left( \theta_j \geq -\hat{\theta}_i \right) \) cannot be lower than \( \frac{\theta_j^* + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} }{\hat{\theta}_j + \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \}} \), note that

\[
\Pr \left( \theta_j < -\hat{\theta}_i \right) \left( \theta_j \mid \theta_j < -\hat{\theta}_i \right) + \Pr \left( \theta_j \geq -\hat{\theta}_i \right) \left( \theta_j \mid \theta_j \geq -\hat{\theta}_i \right) = \theta_j^* \text{ for all } \phi_j \in \Delta'_i,
\]

so the fact that \( \left\{ \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} \right\} \) maximizes both \( \Pr \left( \theta_j < -\hat{\theta}_i \right) \) and \( \Pr \left( \theta_j \geq -\hat{\theta}_i \right) \) implies that \( \left\{ \min \{ \hat{\theta}_i - \alpha_i \hat{\theta}_j \} \right\} \) minimizes \( \Pr \left( \theta_j \geq -\hat{\theta}_i \right) \left( \theta_j \mid \theta_j \geq -\hat{\theta}_i \right) \). This proves the claim.

Therefore,

\[
\sup_{\hat{\theta}_i \in \left( -\theta_j^*, \theta_i \right]} \inf_{\phi_j \in \Delta'_i} U_i (\hat{\theta}_i, \phi_j; \theta_i) = \sup_{\hat{\theta}_i \in \left( -\theta_j^*, \theta_i \right]} \frac{\theta_j^* + \min \{ \hat{\theta}_i - \theta_j \} }{\hat{\theta}_j + \min \{ \hat{\theta}_i - \theta_j \}} \left( \theta_i + t_i \left( \hat{\theta}_i, \theta_j \right) \right).
\]

\(^{18}\) The form of Chebyshev’s inequality I use throughout the paper is, for random variable \( X \) with mean \( x^* \) and upper bound \( \bar{x} \), \( \Pr (X \geq x) \geq \frac{x^* - x}{\bar{x} - x} \). This follows because \( x^* \leq \Pr (X \geq x) \bar{x} + \Pr (X < x) x \). See, for example, p. 319 of Grimmett and Stirzaker (2001).
I now show that the objective is non-decreasing in \( \hat{\theta}_i \) over \((-\theta_j^*, \theta_i]\). In particular, if \( \hat{\theta}_i > -\theta_j \) then the derivative of the objective is 0. If instead \( \hat{\theta}_i \in (-\theta_j^*, -\theta_j) \), then

\[
\alpha_i' \left( \hat{\theta}_i \right) = -\frac{1}{\theta_j^* + \hat{\theta}_i} \left( \alpha_i \left( \hat{\theta}_i \right) - \frac{\bar{\theta}_j - \theta_j^*}{\bar{\theta}_j + \hat{\theta}_i} \right),
\]

and

\[
\frac{\partial}{\partial \hat{\theta}_i} t_i \left( \hat{\theta}_i, \theta_j \right) = -\frac{\bar{\theta}_j - \theta_j^*}{\theta_j^* + \hat{\theta}_i} \left( 1 - \alpha_i \left( \hat{\theta}_i \right) \right),
\]

where the last inequality follows because \( \theta_i \geq \hat{\theta}_i \geq -\theta_j^* \geq -\bar{\theta}_j \).

Combining the last four paragraphs yields \( U \left( \theta_i \right) \geq \inf_{\phi_j \in \Delta_j} U \left( \hat{\theta}_i, \phi_j; \theta_i \right) \) for all \( \hat{\theta}_i > \theta_i \), for all \( \hat{\theta}_i \leq -\theta_j^* \), and for all \( \hat{\theta}_i \in (-\theta_j^*, \theta_i]\). Thus, the screening mechanism satisfies IC.

**Claim 3:** The screening mechanism satisfies WBB if and only if Condition (*) holds.

**Proof:** WBB is trivially satisfied when \( \theta_1 + \theta_2 < 0 \), so suppose that \( \theta_1 + \theta_2 \geq 0 \).

If \( \theta_1 < -\theta_2 \) and \( \theta_2 < -\theta_1 \),

\[
t_1 \left( \theta_1, \theta_2 \right) + t_2 \left( \theta_1, \theta_2 \right) = \left( \theta_1 + \theta_2 \right) \left( 1 - \alpha_1 \left( \theta_1 \right) - \alpha_2 \left( \theta_2 \right) \right).
\]

Since \( \alpha_i \left( \theta_i \right) \) is non-increasing in \( \theta_i \) (as \( \frac{1}{x} \log (1 + x) \) is decreasing in \( x \)), this expression is non-positive for all \( \theta_1, \theta_2 \) with \( \theta_1 + \theta_2 \geq 0 \) if and only if \( \alpha_1 \left( \theta_1 \right) + \alpha_2 \left( \theta_2 \right) \geq 1 \). Condition (*) implies \( \alpha_1 \left( \theta_1 \right) + \alpha_2 \left( \theta_2 \right) \geq 1 \) and is equivalent to this inequality when \( \hat{\theta}_i \leq -\theta_j \) for \( i = 1, 2 \).

If \( \theta_1 \geq -\theta_2 \) and \( \theta_2 \geq -\theta_1 \),

\[
t_1 \left( \theta_1, \theta_2 \right) + t_2 \left( \theta_1, \theta_2 \right) = \left( 1 - \alpha_1 \left( \theta_1 \right) \right) \theta_2 + \left( 1 - \alpha_2 \left( \theta_2 \right) \right) \theta_1 + \alpha_1 \left( \theta_1 \right) \theta_2 + \alpha_2 \left( \theta_2 \right) \theta_1
\]

\[= \theta_1 + \theta_2 - \alpha_1 \left( \theta_1 \right) \left( \theta_2 - \theta_1 \right) - \alpha_2 \left( \theta_2 \right) \left( \theta_1 - \theta_1 \right).
\]

This expression is non-decreasing in \( \theta_1 \) and \( \theta_2 \) (for example, it is non-decreasing in \( \theta_1 \) holding \( \alpha_1 \left( \theta_1 \right) \) fixed, and non-increasing in \( \alpha_1 \left( \theta_1 \right) \) holding \( \theta_1 \) fixed), so it is non-positive for all \( \theta_1, \theta_2 \) with \( \theta_1 + \theta_2 \geq 0 \) if and only if

\[
\theta_1 + \theta_2 - \alpha_1 \left( \theta_1 \right) \left( \theta_2 - \theta_1 \right) - \alpha_2 \left( \theta_2 \right) \left( \theta_1 - \theta_1 \right) \leq 0.
\]
Moving the product terms to the right-hand side and dividing by $\bar{\theta}_1 + \bar{\theta}_2$ (which is positive) shows that this inequality is equivalent to Condition (*) when $\bar{\theta}_i \geq -\bar{\theta}_j$ for $i = 1, 2$ (which is the case under consideration).

Finally, if $\theta_1 < -\theta_2$ and $\theta_2 \geq -\theta_1$ (which is the hardest case),\(^{19}\)

$$t_1 (\theta_1, \theta_2) + t_2 (\theta_1, \theta_2) = (1 - \alpha_1 (\theta_1) - \alpha_2 (\theta_2)) \theta_1 + (1 - \alpha_1 (\theta_1)) \theta_2 + \alpha_2 (\theta_2) \bar{\theta}_1 + \bar{\theta}_2 \alpha_2 (\theta_2).$$

This expression is non-positive for all $\theta_1, \theta_2$ with $\theta_1 + \theta_2 > 0$ if and only if

$$\alpha_1 (\theta_1) + \frac{\theta_1 - \theta_1}{\theta_1 + \theta_2} \alpha_2 (\theta_2) \geq 1$$

for all such $\theta_1, \theta_2$. The left-hand side of this expression is decreasing in $\theta_2$, so it holds for all $\theta_1, \theta_2$ with $\theta_1 + \theta_2 \geq 0$ if and only if

$$\alpha_1 (\theta_1) + \frac{\theta_1 - \theta_1}{\theta_1 + \theta_2} \alpha_2 (\theta_2) \geq 1 \quad (2)$$

for all $\theta_1$. If $\theta_1 \leq -\theta_2^*$, then $\alpha_1 (\theta_1) = 1$ so (2) holds. Suppose toward a contradiction that (2) fails for some $\theta_1 \in [-\theta_2^*, \min \{\bar{\theta}_1, -\bar{\theta}_2\}]$. Observe first that (2) holds at $\theta_1 = \min \{\bar{\theta}_1, -\bar{\theta}_2\}$: this has already been shown if $\bar{\theta}_1 \geq -\bar{\theta}_2$, and if $\bar{\theta}_1 < -\bar{\theta}_2$ it follows by noting that at $\bar{\theta}_1$ (2) is equivalent to Condition (*) when $\bar{\theta}_1 < -\bar{\theta}_2$ and $\bar{\theta}_2 \geq -\bar{\theta}_1$. Since the left-hand side of (2) is continuous in $\theta_1$ and (2) holds for $\theta_1 = \bar{\theta}_1$ and for all $\theta_1 \geq -\bar{\theta}_2$, (2) fails somewhere on the interval $[-\theta_2^*, \min \{\bar{\theta}_1, -\bar{\theta}_2\}]$ if and only if it fails at a local minimum in $(-\theta_2^*, -\bar{\theta}_2)$. Hence, the argument may be completed by showing that no local minimum in $(-\theta_2^*, -\bar{\theta}_2)$ exists. To see this, note that for $\theta_1 \in (-\theta_2^*, -\bar{\theta}_2)$,

$$\alpha'_1 (\theta_1) = -\frac{1}{\theta_2^* + \theta_1} \left( \alpha_1 (\theta_1) - \frac{\bar{\theta}_2 - \theta_2^*}{\theta_2^* + \theta_1} \right),$$

and therefore the first-order condition for an extremum is

$$\frac{\bar{\theta}_2 + \theta_1}{(\bar{\theta}_2 + \theta_1)^2} \alpha_2 (\bar{\theta}_2) = \frac{1}{\theta_2^* + \theta_1} \left( \alpha_1 (\theta_1) - \frac{\bar{\theta}_2 - \theta_2^*}{\theta_2^* + \theta_1} \right).$$

In addition, the second derivative of the left-hand side of (2) equals

$$-\frac{(\bar{\theta}_2 - \theta_2^*) (2 (\bar{\theta}_2 + \theta_1) + \theta_2^* + \theta_1)}{(\theta_2^* + \theta_1)^2 (\bar{\theta}_2 + \theta_1)^2} + 2 \frac{\alpha_1 (\theta_1)}{(\theta_2^* + \theta_1)^2} - 2 \frac{\bar{\theta}_2 + \theta_1}{(\theta_2^* + \theta_1)^3} \alpha_2 (\bar{\theta}_2).$$

---

\(^{19}\)This assumes $\theta_1 + \theta_2 > 0$. If instead $\theta_1 + \theta_2 = 0$, then $t_1 (\theta_1, \theta_2) + t_2 (\theta_1, \theta_2) = -\alpha_2 (\theta_2) (\theta_1 - \theta_1) \leq 0$. 

18
At an extremum, using the first-order condition implies that this equals

\[
-\frac{(\bar{\theta}_2 - \theta_2^*) (2 (\bar{\theta}_2 + \theta_1) + \theta_2^* + \theta_1)}{(\theta_2^* + \theta_1)^2 (\theta_2 + \theta_1)^2} + 2 \frac{\alpha_1(\theta_1)}{(\theta_2^* + \theta_1)^2} - 2 \frac{1}{(\theta_2 + \theta_1) (\theta_2^* + \theta_1)} \left( \alpha_1(\theta_1) - \frac{\bar{\theta}_2 - \theta_2^*}{\theta_2 + \theta_1} \right)
\]

\[
= \frac{\bar{\theta}_2 - \theta_2^*}{(\theta_2^* + \theta_1)^2 (\theta_2 + \theta_1)} \left[ \alpha_1(\theta_1) + \left( \alpha_1(\theta_1) - \frac{\bar{\theta}_2 - \theta_2^*}{\theta_2 + \theta_1} \right) - 1 \right]
\]

\[
= \frac{\bar{\theta}_2 - \theta_2^*}{(\theta_2^* + \theta_1)^2 (\theta_2 + \theta_1)} \left[ \alpha_1(\theta_1) + \frac{(\theta_2^* + \theta_1) (\bar{\theta}_2 + \theta_1)}{(\theta_2 + \theta_1)^2} \alpha_2(\bar{\theta}_2) - 1 \right].
\]

Next, observe that

\[
\frac{\theta_1 - \theta_1}{\theta_1 + \theta_2} \geq \frac{(\theta_2^* + \theta_1) (\bar{\theta}_2 + \theta_1)}{(\theta_2 + \theta_1)^2},
\]

which may be seen by cross-multiplying by \((\bar{\theta}_2 + \theta_1)^2\) and noting that \(\theta_1 - \theta_1 \geq \theta_2^* + \theta_1\) (as \(\theta_1 + \theta_2^* \leq 0\)) and \(\bar{\theta}_2 + \theta_1 \geq \bar{\theta}_1 + \bar{\theta}_2\). Therefore,

\[
\alpha_1(\theta_1) + \frac{\theta_1 - \theta_1}{\theta_1 + \theta_2} \alpha_2(\bar{\theta}_2) \geq \alpha_1(\theta_1) + \frac{(\bar{\theta}_2 + \theta_1) (\theta_2^* + \theta_1)}{(\theta_1 + \theta_2)^2} \alpha_2(\bar{\theta}_2),
\]

so if \((2)\) fails then the second derivative is negative at any local extremum. That is, any local extremum in \((-\theta_2^*, -\bar{\theta}_2)\) must be a local maximum, so no local minimum in \((-\theta_2^*, -\theta_2)\) exists, completing the proof. The argument for \(\theta_1 \geq -\bar{\theta}_2\) and \(\theta_2 < -\theta_1\) is symmetric.

### 3.2 Proof of Theorem 1: Negative Direction

Assume Condition (*) fails. I suppose there exists a mechanism \(x\) satisfying E, IC, IR, and WBB, and derive a contradiction. By E,

\[
y(\theta_i, \theta_j) = \begin{cases} 
1 & \text{if } \theta_i + \theta_j \geq 0 \\
0 & \text{if } \theta_i + \theta_j < 0 
\end{cases}.
\]

Let

\[
y(\theta_i, \varphi_j) = E^{\varphi_j} [y(\theta_i, \theta_j)],
\]

let

\[
\tilde{y}_i(\theta_i) = \begin{cases} 
0 & \text{if } \theta_i \leq -\theta_j^* \\
\frac{\theta_i^* + \theta_i}{\theta_j^* + \theta_i} & \text{if } \theta_i \in (-\theta_2^*, -\theta_j) \\
1 & \text{if } \theta_i \geq -\theta_j
\end{cases},
\]

19
and note that

$$\tilde{y}_i(\theta_i) \leq y(\theta_i, \phi_j) \quad \text{for all } \phi_j \in \Delta'_j$$

(this is immediate for the $\theta_i \leq -\theta^*_j$ and $\theta_i \geq -\theta^*_j$ cases, and follows by Chebyshev’s inequality in the $\theta_i \in (-\theta^*_j, -\theta^*_j)$ case).

The following lemma gives a weaker version of the usual envelope characterization of payoffs.\(^{20}\)

**Lemma 2** For any mechanism $x$ satisfying $E$ and $IC$,

$$U_i(\theta_i) \geq \int_{\Theta_i}^{\theta_i} \tilde{y}_i(\theta) \, d\theta + U_i(\overline{\theta}_i) \quad \text{for all } \theta_i.$$

**Proof.** A straightforward consequence of $IC$ is that $U_i$ is non-decreasing, and hence differentiable almost everywhere. Given a point of differentiability $\theta^0_i \in (\overline{\theta}_i, \overline{\theta}_i)$, fix a sequence $(\theta^k_i)_{k=1}^\infty$ converging to $\theta^0_i$ from above. We have

$$U_i(\theta^k_i) - U_i(\theta^0_i) \geq \inf_{\phi_j \in \Delta'_j} U_i(\theta^0_i, \phi_j; \theta^k_i) - \inf_{\phi_j \in \Delta'_j} U_i(\theta^0_i, \phi_j; \theta^0_i)$$

$$\geq \inf_{\phi_j \in \Delta'_j} \left( U_i(\theta^0_i, \phi_j; \theta^k_i) + (\theta^k_i - \theta^0_i) y(\theta^0_i, \phi_j) \right) - \inf_{\phi_j \in \Delta'_j} U_i(\theta^0_i, \phi_j; \theta^0_i)$$

$$\geq \inf_{\phi_j \in \Delta'_j} (\theta^k_i - \theta^0_i) y(\theta^0_i, \phi_j) \geq (\theta^k_i - \theta^0_i) \tilde{y}(\theta^0_i).$$

Hence,

$$\frac{U_i(\theta^k_i) - U_i(\theta^0_i)}{\theta^k_i - \theta^0_i} \geq \tilde{y}(\theta^0_i),$$

and therefore $U'_i(\theta^0_i) \geq \tilde{y}_i(\theta^0_i)$. The result follows by non-decreasingness and differentiability almost everywhere. \(\blacksquare\)

\(^{20}\)Lemma 2 is related to Theorem 1 of Bodoh-Creed (2012). Bodoh-Creed’s result gives a full characterization of payoffs using Milgrom and Segal’s (2002) envelope theorem for saddle point problems. The difference comes because the maxmin problem defining $IC$ is a saddle point problem in Bodoh-Creed but not in the present paper; the reason is that Bodoh-Creed assumes that $y(\theta_i, \phi_j)$ is continuous in $\theta_i$ and $\phi_j$ (his assumption A8), which is not the case here. My approach of bounding $U_i(\theta_i)$ bears some resemblance to Segal and Whinston (2002), Carbajal and Ely (2013), or Kos and Messner (2013), with the difference that the difficulty here is not relating $U_i(\theta_i)$ to $\int_{\Theta_i}^{\theta_i} U'_i(\theta) \, d\theta$, but rather relating $U'_i(\theta)$ to (bounds on) $y(\theta, \phi_j)$.
Integrating by parts now gives that for every measure $\phi_i$,

$$U_i(\theta_i) \leq \int_{\mathcal{Q}_i}^{\partial i} U_i(\theta_i) \, d\phi_i - \int_{\mathcal{Q}_i}^{\partial i} \tilde{y}(\theta_i) (1 - F_i(\theta_i)) \, d\theta_i,$$

where $F_i(\theta_i) = \phi_i([\theta_i, \theta_i])$. Hence, for every measure $\phi_i$ and every measure $\phi_j \in \Delta'_{j}$,

$$U_i(\theta_i) \leq \int_{\mathcal{Q}_i}^{\partial i} U_i(\theta_i) \, d\phi_i - \int_{\mathcal{Q}_i}^{\partial i} \tilde{y}(\theta_i) (1 - F_i(\theta_i)) \, d\theta_i$$

$$\leq \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} (y(\theta_i, \theta_j) \theta_i + t_i(\theta_i, \theta_j)) \, d\phi_j \, d\phi_i - \int_{\mathcal{Q}_i}^{\partial i} \tilde{y}(\theta_i) (1 - F_i(\theta_i)) \, d\theta_i$$

$$= \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} (y(\theta_i, \theta_j) \theta_i + t_i(\theta_i, \theta_j)) \, d\phi_j \, d\phi_i - \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} y(\theta_i, \theta_j) (1 - F_i(\theta_i)) \, d\phi_j \, d\theta_i$$

$$+ \int_{\mathcal{Q}_i}^{\partial i} (1 - F_i(\theta_i)) ((1 - F_j(-\theta_i)) - \tilde{y}_i(\theta_i)) \, d\theta_i.$$

where the second line follows by the definition of $U_i$ and the fact that $\phi_j \in \Delta'_{j}$, and the third line follows because $\int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} (1 - F_i(\theta_i)) \, d\phi_j \, d\theta_i = \int_{\mathcal{Q}_i}^{\partial i} (1 - F_i(\theta_i)) (1 - F_j(-\theta_i)) \, d\theta_i$.

For measures $\phi_i \in \Delta'_{i}$ and $\phi_j \in \Delta'_{j}$, summing over agents yields

$$U_i(\theta_i) + U_j(\theta_j) \leq \sum_{i=1}^{2} \left[ \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} (y(\theta_i, \theta_j) \theta_i + t_i(\theta_i, \theta_j)) \, d\phi_j \, d\phi_i - \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} y(\theta_i, \theta_j) (1 - F_i(\theta_i)) \, d\phi_j \, d\theta_i \right]$$

$$+ \sum_{i=1}^{2} \int_{\mathcal{Q}_i}^{\partial i} (1 - F_i(\theta_i)) ((1 - F_j(-\theta_i)) - \tilde{y}_i(\theta_i)) \, d\theta_i.$$

Now

$$\sum_{i=1}^{2} \left[ \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} (y(\theta_i, \theta_j) \theta_i + t_i(\theta_i, \theta_j)) \, d\phi_j \, d\phi_i - \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} y(\theta_i, \theta_j) (1 - F_i(\theta_i)) \, d\phi_j \, d\theta_i \right]$$

$$\leq \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} y(\theta_i, \theta_j) (\theta_i + \theta_j) \, d\phi_j \, d\phi_i - \sum_{i=1}^{2} \int_{\mathcal{Q}_i}^{\partial i} \int_{\mathcal{Q}_j}^{\partial j} y(\theta_i, \theta_j) (1 - F_i(\theta_i)) \, d\phi_j \, d\theta_i,$$

where the order of the first double integral can be reversed because the integrand is bounded, and the inequality then comes from WBB. Exactly as in the standard proof of the Myerson-Satterthwaite setting.

---

21 This is where I use the fact that an agent’s subjective expected utility is less than her expected utility under any common prior.

22 Intuitively, $\phi_i$ and $\phi_j$ are playing the role of the independent common priors on $\Theta_i$ and $\Theta_j$ in the standard Myerson-Satterthwaite setting.
Satterthwaite theorem, this expression equals

\[- \int_{\theta_i}^{-\theta_i} (1 - F_i(\theta)) (1 - F_j(-\theta)) \, d\theta.\]

Hence, by IR,

\[0 \leq - \int_{\theta_i}^{-\theta_i} (1 - F_i(\theta)) (1 - F_j(-\theta)) \, d\theta + \sum_{i=1}^{2} \left[ \int_{\theta_i}^{\bar{\theta}_i} (1 - F_i(\theta)) ((1 - F_j(-\theta)) - \tilde{y}_i(\theta_i)) \, d\theta \right].\]

(3)

Let \(\phi_i = \delta_{\max\{\bar{\theta}_i, -\bar{\theta}_j\}} \bar{\theta}_i\) (which is assumed to be an element of \(\Delta_i^{'})\) for \(i = 1, 2\). I show that the failure of Condition (*) implies that the right-hand side of (3) is negative, contradicting (3). Let \(\beta_i = \frac{\sigma_i^* + \min\{\bar{\theta}_i, -\bar{\theta}_j\}}{\bar{\theta}_i + \min\{\bar{\theta}_i, -\bar{\theta}_j\}}\), which is the mass on \(\bar{\theta}_i\) under \(\delta_{\max\{\bar{\theta}_i, -\bar{\theta}_j\}} \bar{\theta}_i\). Observe the following (where the last equality uses the assumption that \(\bar{\theta}_i + \theta_j^* \leq 0\)).

\[
\int_{\bar{\theta}_i}^{-\bar{\theta}_i} (1 - F_i(\theta)) (1 - F_j(-\theta)) \, d\theta = \int_{\min\{\bar{\theta}_i, -\bar{\theta}_j\}}^{\max\{\bar{\theta}_i, -\bar{\theta}_j\}} \beta_i \beta_j \, d\theta
\]

\[
= \left( \min\{\bar{\theta}_i, -\bar{\theta}_j\} + \min\{\bar{\theta}_j, -\bar{\theta}_i\} \right) \beta_i \beta_j,
\]

\[
\int_{\bar{\theta}_i}^{\bar{\theta}_i} (1 - F_i(\theta)) (1 - F_j(-\theta)) \, d\theta = \int_{\min\{\bar{\theta}_i, -\bar{\theta}_j\}}^{\max\{\bar{\theta}_i, -\bar{\theta}_j\}} \beta_i \beta_j \, d\theta
\]

\[
= \left( \min\{\bar{\theta}_i, -\bar{\theta}_j\} + \min\{\bar{\theta}_j, -\bar{\theta}_i\} \right) \beta_i \beta_j
\]

\[
+ \max\{0, \bar{\theta}_i + \theta_j^*\} \beta_i
\]

\[
\int_{\bar{\theta}_i}^{\bar{\theta}_i} (1 - F_i(\theta)) \tilde{y}_i(\theta_i) \, d\theta = \int_{\min\{\bar{\theta}_i, -\bar{\theta}_j\}}^{\max\{\bar{\theta}_i, -\bar{\theta}_j\}} \beta_i \left( \frac{\theta_j^* + \theta}{\theta_j + \theta} \right) \, d\theta + \int_{\min\{\bar{\theta}_i, -\bar{\theta}_j\}}^{\bar{\theta}_i} \beta_i \, d\theta
\]

\[
= (\theta_j^* + \min\{\bar{\theta}_i, -\bar{\theta}_j\}) \beta_i
\]

\[
- (\bar{\theta}_j - \theta_j^*) \beta_i \log \left( 1 + \frac{\theta_j^* + \min\{\bar{\theta}_i, -\bar{\theta}_j\}}{\theta_j + \theta_j^*} \right)
\]

\[+ \max\{0, \bar{\theta}_i + \theta_j^*\} \beta_i.\]
Combining these observations, the right-hand side of (3) equals

\[
(\min \{ \theta_1, -\theta_2 \} + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}) \beta_1 \beta_2 - (\theta_1^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \}) \beta_1 - (\theta_2^* + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}) \beta_2 \\
+ (\bar{\theta}_2 - \theta_2^*) \beta_1 \log \left( 1 + \frac{\theta_2^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \}}{\theta_2 - \theta_2^*} \right) + (\bar{\theta}_1 - \theta_1^*) \beta_2 \log \left( 1 + \frac{\theta_1^* + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}}{\theta_1 - \theta_1^*} \right)
\]

\[
= -\frac{(\bar{\theta}_1 + \bar{\theta}_2) (\theta_2^* + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}) (\theta_1^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \})}{(\theta_1 + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}) (\theta_2 + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \})} \log \left( 1 + \frac{\theta_1^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \}}{\theta_2 - \theta_2^*} \right) \\
+ \frac{(\bar{\theta}_2 - \theta_2^*) (\theta_2^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \})}{\theta_2 + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \}} \log \left( 1 + \frac{\theta_1^* + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}}{\theta_1 - \theta_1^*} \right).
\]

Dividing by \( \frac{(\bar{\theta}_1 + \bar{\theta}_2) (\theta_2^* + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}) (\theta_1^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \})}{(\theta_1 + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}) (\theta_2 + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \})} \) (which is positive) yields

\[
-1 + \left( \frac{\bar{\theta}_1 + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}}{\theta_1 + \theta_2} \right) \left( \frac{\theta_1 - \theta_1^*}{\theta_1^* + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}} \right) \log \left( 1 + \frac{\theta_1^* + \min \{ \bar{\theta}_2, -\bar{\theta}_1 \}}{\theta_1 - \theta_1^*} \right) \\
+ \left( \frac{\bar{\theta}_2 + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \}}{\theta_1 + \theta_2} \right) \left( \frac{\theta_2 - \theta_2^*}{\theta_2^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \}} \right) \log \left( 1 + \frac{\theta_2^* + \min \{ \bar{\theta}_1, -\bar{\theta}_2 \}}{\theta_2 - \theta_2^*} \right).
\]

The failure of Condition (*) implies that this expression is negative. Hence, the right-hand side of (3) is negative, contradicting (3).

### 4 Efficient Trade with Reference Rules

This section presents a characterization of when efficient trade is possible with reference rules. The advantage of this result over Theorem 1 is that it involves simpler mechanisms, makes weaker assumptions on \( \Delta \), uses strong rather than weak budget balance, and provides some support for reference rules as a market design solution (cf Erdil and Klemperer, 2010).

Of course, the disadvantage is that it is less general.

I define a reference rule as follows.

**Definition 1** An efficient mechanism \( x \) is a reference rule if there exist transfers \( t_i^* \in \mathbb{R} \) for
\(i = 1, 2\) such that \(t^*_i = -t^*_j\) and

\[
t_i(\theta_i, \theta_j) = \begin{cases} 
  t^*_i & \text{if } \theta_i \geq -t^*_i, \theta_j \geq -t^*_j, \theta_i + \theta_j \geq 0 \\
  -\theta_i & \text{if } \theta_i < -t^*_i, \theta_j \geq -t^*_j, \theta_i + \theta_j \geq 0 \\
  \theta_j & \text{if } \theta_i \geq -t^*_i, \theta_j < -t^*_j, \theta_i + \theta_j \geq 0 \\
  0 & \text{if } \theta_i + \theta_j < 0 
\end{cases}
\]

With a reference rule, when the agents trade they receive the “reference transfers” \((t^*_1, -t^*_1)\) if this gives each of them a non-negative payoff. Otherwise, the transfer of the agent who would get a negative payoff with the reference transfers is set to give her payoff 0, and the other agent’s transfer is set to balance the budget. For example, in the bilateral trade setting, a reference rule corresponds to setting a reference price \(p^*\), trading at price \(p^*\) if \(c \leq p^* \leq v\), and otherwise trading at whichever value is closer to \(p^*\) (i.e., price is \(v\) if \(c \leq v < p^*\); price is \(c\) if \(p^* < c \leq v\)).

Reference rules clearly satisfy E, (ex post) IR, and SBB, so an IC reference rule satisfies all of our criteria for implementing efficient trade. The following result characterizes when IC reference rules exist; that is, when efficient trade is implementable with reference rules.

**Theorem 2** Assume that \(\delta_{\theta^*_i} \in \Delta'_i\) for \(i = 1, 2\). Then efficient trade is implementable with reference rules if and only if either

1. \(\theta^*_1 + \theta^*_2 \leq 0\) (the average types do not have strict gains from trade), or

2. \(\theta^*_1 + \theta^*_2 \geq 0\) (every pair of types has gains from trade).

The proof of the positive direction shows that reference rules with \(t^*_1 \in [\theta^*_2, -\theta^*_1]\) satisfy IC if \(\theta^*_1 + \theta^*_2 \leq 0\). The reason why was explained in the introduction for the bilateral trade setting; the proof just fills in the details. The intuition for the negative direction may be seen in Figure 3 (again for the bilateral trade setting). Suppose the reference price \(p^*\) is greater than \(c^*\). Consider a buyer with value \(v \in (c^*, p^*)\). If he reports his value truthfully, then whenever he trades under the reference rule he does so at price \(v\), which gives him payoff 0. Suppose he instead shades his report down to some \(\hat{v} \in (c^*, v)\). Then whenever

\[\text{Note that this defines } t_i(\theta_i, \theta_j) \text{ for all } \theta_i, \theta_j, \text{ because if } \theta_i < -t^*_i \text{ and } \theta_j < -t^*_j \text{ then } \theta_i + \theta_j < -t^*_i - t^*_j = 0.\]
he trades the price is \( \hat{v} \), which gives him a positive payoff, and in addition he expects to trade with positive probability (since \( \hat{v} > c^* \)). So he will shade down. The same argument shows that in any reference rule a seller with \( c \in (p^*, v^*) \) shades up. Figure 3 shows that a consequence of this argument is that a reference rule cannot be IC for both agents when \( c^* < v^* \), regardless of where the reference price \( p^* \) is set.

**Proof of Theorem 2. Positive Direction:**

When \( \theta_1 + \theta_2 \geq 0 \), any reference rule with \( t_i^* \in [-\theta_i, \theta_j] \) satisfies IC.

When \( \theta_1^* + \theta_2^* \leq 0 \), I show that any reference rule with \( t_i^* \in [\theta_1^*, -\theta_2^*] \) satisfies IC.

First, suppose that \( \theta_i < -\theta_j^* \). Observe that \( U_i(\hat{\theta}_i, \theta_j; \theta_i) \leq 0 \) for all \( \hat{\theta}_i \). This follows because if \( \hat{\theta}_i + \theta_j^* < 0 \) then \( U_i(\hat{\theta}_i, \theta_j^*; \theta_i) = 0 \), while if \( \hat{\theta}_i + \theta_j^* \geq 0 \) then \( U_i(\hat{\theta}_i, \theta_j^*; \theta_i) = \theta_i + \theta_j^* < 0 \). Hence, for all \( \hat{\theta}_i \), IR and \( \delta \theta_j^* \in \Delta_j \) imply that \( U_i(\theta_i) \geq 0 \geq \inf_{\phi_j \in \Delta_j} U_i(\hat{\theta}_i, \phi_j; \theta_i) \), which yields IC.

Next, suppose that \( \theta_i \geq -\theta_j^* \). First, note that misreports of \( \hat{\theta}_i \leq -\theta_j^* \) cannot be profitable because \( U_i(\theta_i) = 0 = U_i(\hat{\theta}_i, \theta_j^*; \theta_i) \) and \( \delta \theta_j^* \in \Delta_j \). Next, consider misreports of \( \hat{\theta}_i > -\theta_j^* \). If \( y(\theta_i, \theta_j) = y(\hat{\theta}_i, \theta_j) \), then \( t_i(\theta_i, \theta_j) = t_i(\hat{\theta}_i, \theta_j) \) (as \( \theta_i, \hat{\theta}_i \geq -\theta_j^* \geq -t_i^* \)), and hence \( U_i(\theta_i, \theta_j; \theta_i) = U_i(\hat{\theta}_i, \theta_j; \theta_i) \). In addition, if \( y(\theta_i, \theta_j) = 1 \) and \( y(\hat{\theta}_i, \theta_j) = 0 \) then \( U_i(\theta_i, \theta_j; \theta_i) \geq 0 = U_i(\hat{\theta}_i, \theta_j; \theta_i) \) by ex post IR. Finally, if \( y(\theta_i, \theta_j) = 0 \) and \( y(\hat{\theta}_i, \theta_j) = 1 \) then \( U_i(\hat{\theta}_i, \theta_j; \theta_i) \leq \theta_i + \theta_j < 0 = U_i(\theta_i, \theta_j; \theta_i) \). Hence, \( U_i(\theta_i, \theta_j; \theta_i) \geq U_i(\hat{\theta}_i, \theta_j; \theta_i) \) for all \( \theta_j \), so misreports of \( \hat{\theta}_i > -\theta_j^* \) cannot be profitable, either. This yields IC.

**Negative Direction:**
Since \( \theta_1^* + \theta_2^* > 0 \), every reference rule satisfies either \( t_1^* < \theta_2^* \) or \( t_2^* > -\theta_1^* \), and hence \( t_i^* < \theta_j^* \) for some \( i \in \{1, 2\} \). Fix this choice of \( i \).

First, suppose that \( \theta_i = -t_i^* \). Then there exists a type \( \theta_j \in (-\theta_j^*, -t_i^*] \cap \Theta_i \). Note that \( U_i(\theta_i, \theta_j; \theta_i) = 0 \) for all \( \theta_i \). However, \( U_i(\hat{\theta}_i, \theta_j; \hat{\theta}_i) = \theta_i - \hat{\theta}_i > 0 \) whenever \( \hat{\theta}_i < \theta_i \) and \( \hat{\theta}_i + \theta_j \geq 0 \). In addition, for all \( \hat{\theta}_i \in (-\theta_j^*, \theta_i) \) and all \( \phi_j \in \Delta_j^* \), we have \( \Pr^\phi_j(\hat{\theta}_i + \theta_j \geq 0) \geq \frac{\theta_j^* + \theta_i}{\theta_j + \theta_i} \) (by Chebyshev’s inequality), and therefore \( U_i(\hat{\theta}_i, \phi_j; \theta_i) \geq \frac{\theta_j^* + \theta_i}{\theta_j + \theta_i}(\theta_i - \theta_j) > 0 \). Hence, \( \inf_{\phi_j \in \Delta_j^*} U_i(\hat{\theta}_i, \phi_j; \theta_i) > 0 = \inf_{\phi_j \in \Delta_j^*} U_i(\theta_i, \phi_j; \theta_i) \), so IC fails.

Next, suppose that \( \theta_i > -t_i^* \). Then there exists a type \( \theta_j \in (-\theta_i, -t_i^*) \cap \Theta_j \) (as \( \theta_j < -\theta_2^* < t_i^* = -t_2^* < \theta_1^* < \theta_j^* \)). Note that \( U_j(\theta_j, \theta_i; \phi_j) = 0 \) for all \( \theta_i \). However, \( U_j(\hat{\theta}_j, \theta_i; \phi_j) = \theta_j - \hat{\theta}_j > 0 \) whenever \( \hat{\theta}_j < \theta_j \) and \( \theta_i + \hat{\theta}_j \geq 0 \). Now for all \( \hat{\theta}_j \in (-\theta_i, \theta_j) \) and all \( \phi_j \in \Delta_j^* \), we have \( \Pr^\phi_j(\theta_i + \hat{\theta}_j \geq 0) = 1 \) and therefore \( U_j(\hat{\theta}_j, \phi_j; \theta_j) = \theta_j - \hat{\theta}_j \). Hence, \( \inf_{\phi_j \in \Delta_j^*} U_j(\hat{\theta}_j, \phi_j; \theta_j) = \theta_j - \hat{\theta}_j > 0 = \inf_{\phi_j \in \Delta_j^*} U_j(\theta_j, \phi_j; \theta_j) \), so IC fails.

5 Discussion and Extensions

This section discusses several aspects of the results and provides some extensions.

5.1 Information Acquisition Interpretation

The assumption that agents know the mean and bounds on the support of the distribution of each other’s value emerges naturally when agents share a common prior at an ex ante stage but are uncertain about the information acquisition technology that the opponent can access prior to entering the mechanism. This section provides the details of this argument.

Consider the following extension of the model. Each agent \( i \)'s ex post utility is

\[
\tilde{\theta}_i y + t_i,
\]

where \( \tilde{\theta}_i \in \mathbb{R} \) is her realized ex post value. There is an ex ante stage at which the agents’ beliefs about the ex post values \( (\tilde{\theta}_1, \tilde{\theta}_2) \) are given by a common product measure on \([\tilde{\theta}_1, \hat{\theta}_1] \times [\tilde{\theta}_2, \hat{\theta}_2]\) with mean \((\theta_1^*, \theta_2^*)\) (the common prior). Before entering the mechanism, each agent observes the outcome of some experiment that is informative of her own ex post value but independent of her opponent’s. Agent \( i \)'s interim value, \( \theta_i \) (which corresponds to her type in
the main model), is then her posterior expectation of $\tilde{\theta}_i$ after observing the outcome of this experiment.

The following observation allows for the desired interpretation of the main model.

**Remark 1** If a measure $\phi_i$ on $\mathbb{R}$ is the distribution of posterior expectations of $\tilde{\theta}_i$ corresponding to some experiment, then $\phi_i \in \Delta_i$ (i.e., $E^{\phi_i} [\theta_i] = \theta_i^*$ and the support of $\phi_i$ is contained in $[\tilde{\theta}_i, \bar{\theta}_i]$).

The fact that $E^{\phi_i} [\theta_i] = \theta_i^*$ is the law of iterated expectation. The fact that the support of $\phi_i$ is contained in $[\tilde{\theta}_i, \bar{\theta}_i]$ follows because $\tilde{\theta}_i \in [\theta_i, \bar{\theta}_i]$ with probability 1 under the prior. Thus, assuming that agent $j$ finds any distribution in some subset $\Delta'_i \subseteq \Delta_i$ possible amounts to assuming that he finds it possible that player $i$ may have access to some subset of all possible experiments. With this interpretation, the assumption that $\delta_{\theta_i^*, \theta_i^b} \in \Delta'_i$ means that agent $j$ finds it possible that agent $i$ acquires no information about her value before entering the mechanism (beyond the common prior), while the assumption that $\delta_{\theta_i^l, \theta_i^b} \in \Delta'_i$ means that agent $j$ finds it possible that agent $i$ observes a binary signal of her value, where the “bad” realization lowers her expectation of $\tilde{\theta}_i$ to $\theta_i^l$ and the “good” realization raises her expectation of $\tilde{\theta}_i$ to $\theta_i^h$.

### 5.2 Relaxing Known Means

The assumption that each agent knows the mean of her opponent’s distribution seems natural in view of the above interpretation of the model in terms of information acquisition. However, this assumption can be relaxed for both Theorems 1 and 2.

For Theorem 1, it is easily checked that the proof only uses the fact that $E [\theta_i] \geq \theta_i^*$ for all $\phi_i \in \Delta'_i$, rather than $E [\theta_i] = \theta_i^*$ for all $\phi_i \in \Delta'_i$. The intuition is that, with maxmin agents, bounding how bad agents’ beliefs can be is important, while bounding how good they can be is not. This gives the following.

---

24 More precisely, $\Delta'_i$ is jointly determined by the set of experiments player $i$ may have access to and the prior. For example, it is possible that $\Delta'_i$ equals $\Delta_i$ if and only if the prior puts probability 1 on agent $i$’s ex post value being either $\tilde{\theta}_i$ or $\bar{\theta}_i$ (see, for example, Theorem 1 of Shmaya and Yariv (2009) or Proposition 1 of Kamenica and Gentzkow (2011)).
Proposition 2 Let $\tilde{\Delta}_i$ be the set of Borel measures $\phi_i$ on $\Theta_i$ such that $E^\phi_i [\theta_i] \geq \theta_i^*$. Theorem 1 continues to hold when the assumption that $\Delta_i' \subseteq \Delta_i$ is relaxed to $\Delta_i' \subseteq \tilde{\Delta}_i$.

For Theorem 2, what is essential for the positive result is that there is a pair of values $(\theta_1^*, \theta_2^*)$ such that agent 1 may believe that agent 2’s value is almost surely less than $\theta_2^*$, agent 2 may believe that agent 1’s value is almost surely less than $\theta_1^*$, and types $\theta_1^*$ and $\theta_2^*$ do not have strict gains from trade with each other. The values $\theta_1^*$ and $\theta_2^*$ need not have anything to do with the agents’ average values. Similarly, what is essential for the negative result is that there is a pair of values $(\theta_1^*, \theta_2^*)$ such that $\theta_1^* + \theta_2^* > 0$ and $\Pr^\phi_i (\theta_i \geq \theta_i^*)$ is bounded away from 0 for all $\phi_i \in \Delta_i'$. In particular, straightforward modifications to the proof of Theorem 2 yield the following.

Proposition 3 Let $\Delta_1', \Delta_2'$ be nonempty, compact, convex sets of Borel measures on $\Theta_1, \Theta_2$ (not necessarily all with the same mean). Efficient trade is implementable with reference rules if either

1. There exist $\theta_1^*, \theta_2^* \in \mathbb{R}$ and $\phi_1^* \in \Delta_1', \phi_2^* \in \Delta_2'$ such that $\Pr^{\phi_1^*} (\theta_1 \leq \theta_1^*) = \Pr^{\phi_2^*} (\theta_2 \leq \theta_2^*) = 1$ and $\theta_1^* + \theta_2^* \leq 0$, or
2. $\theta_1 + \theta_2 \geq 0$.

Conversely, efficient trade is not implementable with reference rules if

1. There exist $\theta_1^*, \theta_2^* \in \mathbb{R}$ and $\varepsilon > 0$ such that $\Pr^{\phi_i} (\theta_i \geq \theta_i^*) \geq \varepsilon$ for all $\phi_i \in \Delta_i'$ (for $i = 1, 2$), and $\theta_1^* + \theta_2^* > 0$, and
2. $\theta_1 + \theta_2 < 0$.

5.3 Disallowing Weakly Dominated Strategies

A potential criticism of the efficient trading mechanisms employed in the proofs of both Theorems 1 and 2 is that truth-telling is weakly dominated for some types. In particular, in both mechanisms an agent of type $\theta_i < -\theta_j^*$ gets payoff 0 against every opposing type if she reports truthfully, while if she shades her report down she gets payoff 0 against some
opposing types but may get a positive payoff against others. This section shows that the mechanisms can be modified to address this concern.

The modification required for Theorem 1 is simple: when \( \theta_i < -\theta_j^* \), modify \( t_i(\theta_i, \theta_j) \) by replacing \( \alpha_i(\theta_i) = 1 \) with \( \alpha_i(\theta_i) = 1 - \varepsilon \) for some \( \varepsilon > 0 \). In other words, modify the mechanism so that an agent with \( \theta_i < -\theta_j^* \) gets fraction \( \varepsilon \) of the gains from any trade she is involved in, rather than getting none of the gains from trade. Inspecting the proof of Theorem 1 shows that this modification does not lead to a violation of WBB if \( \varepsilon \) is sufficiently small.

A similar modification of reference rules ensures the truth-telling is not weakly dominated: when \( \theta_i < -\theta_j^* \), change \( t_i(\theta_i, \theta_j) \) from \( -\theta_i \) to \( -(1 - \varepsilon) \theta_i + \varepsilon \theta_j \). However, since reference rules are strongly budget balanced, this modification violates (both weak and strong) budget balance unless \( t_j(\theta_i, \theta_j) \) is also changed from \( \theta_i \) to \( (1 - \varepsilon) \theta_i - \varepsilon \theta_j \). This change can in turn lead to a violation of IC for agent \( j \). Nonetheless, it turns out that IC is preserved if \( \varepsilon \) is not too large and \( \Delta_j^s \) satisfies the same richness condition as in Theorem 1, as the following result shows (proof in appendix).\(^{25}\)

**Proposition 4** Assume that \( \theta_1^* + \theta_2^* \leq 0 \) and that \( \delta_{\theta_i, \theta_i} \in \Delta^s_i \) for all \( \theta_i \in [\bar{\theta}_i, \bar{\theta}_i^*] \). Then for every \( t_i^* \in (\max \{ \theta_j^*, -\bar{\theta}_i \}, \min \{ -\theta_i^*, \bar{\theta}_j \}) \), the \( \varepsilon \)-modified reference rule given by

\[
\begin{align*}
t_i(\theta_i, \theta_j) = \begin{cases} 
t_i^* & \text{if } \theta_i \geq -t_i^*, \theta_j \geq -t_j^*, \theta_i + \theta_j \geq 0 \\
-(1 - \varepsilon) \theta_i + \varepsilon \min \{ \theta_j, -\bar{\theta}_j \} & \text{if } \theta_i < -t_i^*, \theta_j \geq -t_j^*, \theta_i + \theta_j \geq 0 \\
(1 - \varepsilon) \theta_j - \varepsilon \min \{ \theta_i, -\bar{\theta}_i \} & \text{if } \theta_i \geq -t_i^*, \theta_j < -t_j^*, \theta_i + \theta_j \geq 0 \\
0 & \text{if } \theta_i + \theta_j < 0
\end{cases}
\end{align*}
\]

satisfies E, IR, and SBB, and satisfies IC for all \( \varepsilon \in \left(0, \min_{i \in \{1, 2\}} \frac{\bar{\theta}_i + t_i^*}{\min \{ \bar{\theta}_i + \theta_j, \bar{\theta}_j - \bar{\theta}_i \}} \right) \). In addition, under such a mechanism truth-telling is not weakly dominated for any type.

The statement of Proposition 4 does not cover the cases where \( \theta_1^* + \theta_2^* = 0 \) or \( \bar{\theta}_1 + \bar{\theta}_2 \leq 0 \). However, the same proof shows that when \( \theta_1^* + \theta_2^* = 0 \) the \( \varepsilon \)-modified reference rule with \( t_i^* = \theta_j^* \) is IC for all \( \varepsilon \in \left(0, \min_{i \in \{1, 2\}} \frac{\bar{\theta}_i + t_i^*}{\min \{ \bar{\theta}_i + \theta_j, \bar{\theta}_j - \bar{\theta}_i \}} \right) \), while when \( \bar{\theta}_1 + \bar{\theta}_2 \leq 0 \) it may be

\(^{25}\)Note that the modification in Proposition 4 that preserves IC is actually from \( -\theta_i \) to \( -(1 - \varepsilon) \theta_i + \varepsilon \min \{ \theta_j, -\bar{\theta}_j \} \) rather than \( -(1 - \varepsilon) \theta_i + \varepsilon \theta_j \).
checked that the $\varepsilon$-modified reference rule with $t^*_i = \tilde{\theta}_j$ is IC for all $\varepsilon \in (0, 1)$. Therefore, whenever $\theta^*_1 + \theta^*_2 \leq 0$ there exists an IC $\varepsilon$-modified reference rule for some $\varepsilon > 0$, so weakly dominated strategies can always be disallowed.

A related remark concerns “smoother” models of ambiguity aversion (e.g., Klibanoff, Marinacci, and Mukherjee, 2006) and “smoothness” requirements on agents’ worst-case beliefs (e.g., requiring that every belief $\phi_i \in \Delta^*_i$ has positive density on $\Theta_i$). Requiring robustness in these directions is stronger than disallowing weakly dominated strategies: for example, with an $\varepsilon$-modified reference rule, a type $\theta_i < -\theta^*_j$ agent whose belief has positive density faces a first-order gain from shading her report down against a second-order loss, and will therefore misreport. Thus, the results in this paper—like most results in the literature on mechanism design with ambiguity-averse agents—directly concern only maxmin agents and not smoothly ambiguity-averse ones.

5.4 Minmax Agents

As remarked above, the maxmin problem defining IC is not a saddle point problem, and the max and min do not in general commute. Therefore, the results established in this paper for maxmin (“ambiguity-averse”) agents do not necessarily hold for minmax (“Bayesian pessimistic”) ones. However, the maxmin problem defining IC becomes a saddle point problem when attention is restricted to continuous mechanism (i.e., mechanisms where $y$ and $t$ are continuous functions). Continuous mechanisms cannot be exactly efficient, which is why the main mechanisms considered in this paper are discontinuous. Nonetheless, the equivalence of the maxmin and minmax problems under continuous mechanisms suggests that approximate versions of the results for maxmin agents do hold for minmax agents. In addition, the results on efficient trade with reference rules turn out to hold exactly for minmax agents. This section formalizes these arguments.

Say that a mechanism is \textit{minmax incentive compatible} if for all $\theta_i \in \Theta_i$ there exists

$$\phi^\theta_{j,i} \in \arg\min_{\phi_j \in \Delta^*_j} \sup_{\hat{\theta}_i} U_i\left(\hat{\theta}_i, \phi^\theta_{j,i}; \theta_i\right)$$

such that

$$\theta_i \in \arg\max_{\hat{\theta}_i} U_i\left(\hat{\theta}_i, \phi^\theta_{j,i}; \theta_i\right).$$
Both the positive and negative result on efficient trade with reference rules (Theorem 2) hold without further modification when maxmin IC is replaced by minmax IC, by essentially the same proof. The proofs of both the positive and negative results on efficient trade with general mechanisms (Theorem 1) do not go through when maxmin IC is replaced by minmax IC, and whether or not these results hold exactly with minmax IC is an open question. However, approximate version of these results may be expected to hold with minmax IC. Formally, say that a mechanism is \( \varepsilon \)-efficient if

\[
y(\theta_1, \theta_2) \geq 1 - \varepsilon \quad \text{if } \theta_1 + \theta_2 \geq \varepsilon, \text{ and}
\]

\[
y(\theta_1, \theta_2) \leq \varepsilon \quad \text{if } \theta_1 + \theta_2 \leq -\varepsilon.
\]

A straightforward adaptation of the proof of the negative direction Theorem 1 shows that when Condition (*) fails, \( \varepsilon \)-efficient trade is impossible with continuous mechanisms and minmax IC, for \( \varepsilon \) sufficiently small (proof in appendix).

**Proposition 5** Suppose the assumptions of Theorem 1 hold and Condition (*) fails. Then for some \( \varepsilon > 0 \) there does not exist an \( \varepsilon \)-efficient and continuous mechanism that satisfies minmax IC, IR, and WBB.

Conversely, it is to be expected that when Condition (*) holds, \( \varepsilon \)-efficient trade is possible with continuous mechanisms and minmax IC for all \( \varepsilon > 0 \). However, proving this is not a straightforward adaptation of the proof of the positive direction of Theorem 1, since that proof is constructive and the required changes to the construction are not trivial. I therefore leave this result as a conjecture.

A final remark on minmax agents concerns the gap between the conditions for efficient trade under weak budget balance and exact budget balance. Under both the screening and reference rule mechanisms, a type \( \theta_i \) agent’s worst-case belief is \( \delta_{\max\{-\theta_i, \bar{\theta}_j\}} \bar{\theta}_j \) (which incidentally is also the belief that minimizes the probability that gains from trade exist). If we impose that agents have these beliefs (rather than letting beliefs depend on the mechanism), then it is actually fairly easy to see that efficient trade with exact budget balance is possible if and only if either \( \theta_1^* + \theta_2^* \leq 0 \) or \( \theta_1 + \theta_2 \geq 0 \)—that is, if and only if efficient trade is implementable with reference rules. In particular, incentive
compatibility for agent 1 implies that the transfer she receives against type \( \theta_2 \) cannot depend on her own type, so long as she expects to trade with positive probability: that is, 
\[
t_1(\bar{\theta}_1, \theta_2) = t_1(\max\{-\theta_2^*, \theta_1\}, \theta_2).
\]
Individual rationality of type \( \max\{-\theta_2^*, \theta_1\} \) now gives 
\[
t_1(\bar{\theta}_1, \theta_2) \geq \min\{\theta_2^*, -\theta_1\}.
\]
Symmetrically, 
\[
t_2(\bar{\theta}_1, \bar{\theta}_2) \geq \min\{\theta_1^*, -\theta_2\}.
\]
Strong budget balance now yields 
\[
t_1(\bar{\theta}_1, \theta_2) \geq \max\{\theta_2^*, -\theta_1\};
\]
\[\min\{\theta_1^*, -\theta_2\}.
\]
Finally, this interval is non-empty if and only if either \( \theta_1^* + \theta_2^* \leq 0 \) or \( \theta_1 + \theta_2 \geq 0 \). This argument suggests that efficient trade with exact budget balance and minmax agents is possible only if it is possible with reference rules. Of course, it is not a proof, because it is not true that a type \( \theta_i \) agent’s worst-case belief is \( \delta_{\max\{-\theta_i, \theta_2\}}(\theta) \) for every possible mechanism.

### 5.5 Maxmax Agents

As a last extension, it is interesting to compare the results of this paper with the case of optimistic (i.e., “maxmax”) agents. In particular, with optimistic agents, one can always implement efficient trade with the following mechanism: Let \( \mathbb{D} \) denote the dyadic numbers (i.e., the real numbers with terminating binary expansions), let \( I(\theta_i) \) denote the last digit of the binary expansion of \( \theta_i \) if \( \theta_i \in \mathbb{D} \), let

\[
z(\theta_1, \theta_2) = \begin{cases} 
0 & \text{if } \theta_1 \notin \mathbb{D}, \theta_2 \notin \mathbb{D} \\
I(\theta_1) & \text{if } \theta_1 \in \mathbb{D}, \theta_2 \notin \mathbb{D} \\
I(\theta_2) & \text{if } \theta_1 \notin \mathbb{D}, \theta_2 \in \mathbb{D} \\
(I(\theta_1) + I(\theta_2)) \mod 2 & \text{if } \theta_1 \in \mathbb{D}, \theta_2 \in \mathbb{D}
\end{cases}
\]

and consider the efficient mechanism with transfers

\[
t_i(\theta_i, \theta_j) = \begin{cases} 
-\theta_i & \text{if } z(\theta_i, \theta_j) = i \mod 2, \theta_i + \theta_j \geq 0 \\
\theta_j & \text{if } z(\theta_i, \theta_i) = j \mod 2, \theta_i + \theta_j \geq 0 \\
0 & \text{if } \theta_i + \theta_j \leq 0
\end{cases}
\]

for \( i = 1, 2 \).

Observe that if \( \theta_i + \theta_j \geq 0 \) then \( \theta_j \geq -\theta_i \), so an optimistic agent \( i \) always expects to receive \( \theta_j \) in the event of trade, and is therefore always willing to reveal \( \theta_i \) truthfully (formally, truthtelling is IC in the sense of maximizing \( \sup_{\phi_j \in \Delta_j} U_i(\bar{\theta}_i, \phi_j; \theta_i) \)).

This mechanism is of course not very appealing, and indeed there is a clear intuition for why such a mechanism is not IC with pessimistic (maxmin or minmax) agents: Ex post IR
and SBB imply that \( t_i(\theta_i, \theta_j) \in [-\theta_i, \theta_j] \) when trade occurs. Optimists are inclined to put weight on opposing types such that \( t_i(\theta_i, \theta_j) \) is closer to \( \theta_j \), while pessimists are inclined to put weight on types such that \( t_i(\theta_i, \theta_j) \) is closer to \(-\theta_i\). In particular, if fine changes in \( \theta_j \) flip \( t_i(\theta_i, \theta_j) \) between \( \theta_j \) to \(-\theta_i \) (as in the mechanism given above), optimists will expect transfer \( \theta_j \) and will report truthfully, while pessimists will expect transfer \(-\theta_i \) and will misreport. A more colorful way of putting this is that optimists think nature is acting on their behalf and are therefore happy to reveal their private information, while pessimists think nature will use their private information against them and consequently are more cautious.

6 Conclusion

This paper has revisited the classical problem of bilateral trade or bilateral public good provision under the assumption that agents are pessimistic about the distribution of opposing types. Exact characterizations of when efficient trade is possible are given for general mechanisms that are allowed to run a surplus and also for reference rules, a particularly simple class of mechanisms. The Myerson-Satterthwaite impossibility theorem sometimes continues to hold, despite the lack of a common prior or independent types. When the Myerson-Satterthwaite theorem fails, the mechanisms that implement efficient trade are economically plausible, and indeed in the case of reference rules have previously been recommended as market design solutions to certain problems. Similar results hold when weakly dominated strategies are disallowed, or when agents are Bayesian pessimists rather than maxmin optimizers.

I conclude by mentioning two directions for future research. First, it is natural in light of Theorem 1 to ask what the “second-best” mechanism is when Condition (*) fails. This question is ambiguous because there is no common prior. A first, trivial, observation is that is that, assuming \( \delta_{\theta^*_i} \in \Delta'_i \), a mechanism designer with an atomless prior can attain first-best expected utilitarian welfare with respect to her prior with the mechanism, “Do not trade if either \( \theta_1 = \theta^*_1 \) or \( \theta_2 = \theta^*_2 \); otherwise trade (whenever \( \theta_1 + \theta_2 \geq 0 \)) with transfers \( t_i(\theta_i, \theta_j) = \frac{1}{2} (\theta_j - \theta_i) \)”: this mechanism is IC, as \( \delta_{\theta^*_i} \) is a worst-case belief for any type. A more interesting question is how a mechanism designer with a given prior can maximize her
expectation of the agents’ expected welfare with respect to their own worst-case beliefs. An intriguing challenge here is that an agent’s worst-case beliefs depend on the mechanism, so the designer cannot fix beliefs and optimize over mechanisms.

Another natural direction is generalizing the results of this paper to the case of more than two agents, toward studying settings like those of Cramton, Gibbons, and Klemperer (1987) and Mailath and Postlewaite (1989). One observation is that it is not clear how reference rules can be generalized beyond the bilateral case. The reason in the following:

When \( n = 2 \), a reference rule simply specifies that when agent 2 with \( \theta_2 > -\theta_1^* \) trades with agent 1 with \( \theta_1 < -\theta_2^* \), agent 1 pays her value \( \theta_1 \) and agent 2 pays \(-\theta_1\). In contrast, suppose that \( n = 3 \) and that values satisfy \( \theta_2 > -\theta_1^* - \theta_3^* \), \( \theta_3 > -\theta_1^* - \theta_2^* \), and \( \theta_1 < -\theta_2^* - \theta_3^* \). If agent 1 is again asked to pay her value \( \theta_1 \), the budget-balancing payment of \(-\theta_1\) must be split between agents 2 and 3. However, this split must depend on the precise values of \( \theta_2 \) and \( \theta_3 \): for example, one of them may not be willing to pay \(-\theta_1^*/2\). This issue of cost-sharing among agents who expect to trade with positive probability does not arise when \( n = 2 \) (when \( \theta_1^* + \theta_2^* \leq 0 \)), but seems to present new reasons for misreporting when \( n > 2 \).
Appendix: Omitted Proofs

Proof of Proposition 1. Consider the mechanism

\[
y(\theta_i, \theta_j) = \begin{cases} 
1 & \text{if } \theta_i + \theta_j \geq 0 \\
0 & \text{if } \theta_i + \theta_j < 0 
\end{cases},
\]

\[
t_i(\theta_i, \theta_j) = \begin{cases} 
-\theta_i & \text{if } \theta_i + \theta_j \geq 0 \\
0 & \text{if } \theta_i + \theta_j < 0 
\end{cases},
\]

\[
t_j(\theta_i, \theta_j) = \begin{cases} 
\theta_i & \text{if } \theta_i + \theta_j \geq 0 \\
0 & \text{if } \theta_i + \theta_j < 0 
\end{cases}.
\]

This mechanism clearly satisfies E, IR, and SBB. For IC for agent \(i\), note that \(U_i(\theta_i, \theta_j; \theta_i) = 0\) for all \(\theta_i \in \Theta_i\) and \(U_i(\hat{\theta}_i, \hat{\theta}_j^*; \theta_i) \leq 0\) for all \(\theta_i, \hat{\theta}_i \in \Theta_i\) (as \(\hat{\theta}_i + \theta_j^* \leq 0\), so \(\delta_{\theta_j^*} \in \Delta_j\) implies that \(\inf_{\phi_j \in \Delta_j} U_i(\theta_i, \phi_j; \theta_i) = 0 \geq \inf_{\phi_j \in \Delta_j} U_i(\hat{\theta}_i, \phi_j; \theta_i)\). For IC for agent \(j\), note that

\[
U_j(\theta_j, \theta_i; \theta_j) = \max \{\theta_i + \theta_j, 0\} \geq U_j(\hat{\theta}_j, \theta_i; \theta_j)
\]

for all \(\theta_j, \hat{\theta}_j \in \Theta_j\) and \(\theta_i \in \Theta_i\), so \(\inf_{\phi_i \in \Delta_i} U_j(\theta_j, \phi_i; \theta_j) \geq \inf_{\phi_i \in \Delta_i} U_j(\hat{\theta}_j, \phi_i; \theta_j)\). ■

Proof of Lemma 1. The result is immediate when \(\theta_i \leq -\theta_j^*\) or \(\theta_i \geq -\theta_j\), as in both cases \(\alpha'_i(\theta_i) = 0\), which immediately implies that \(t_i(\theta_i, \theta_j)\) is non-increasing in \(\theta_i\).

When \(\theta_i \in (-\theta_j^*, -\theta_j)\), we have

\[
t_i(\theta_i, \theta_j) = (1 - \alpha_i(\theta_i)) \theta_j - \alpha_i(\theta_i) \theta_i,
\]

and therefore

\[
\frac{\partial}{\partial \theta_i} t_i(\theta_i, \theta_j) = -\alpha_i(\theta_i) - \alpha'_i(\theta_i) (\theta_j + \theta_i).
\]

In addition,

\[
\alpha'_i(\theta_i) = -\frac{1}{\theta_j^* + \theta_i} \left( \alpha_i(\theta_i) - \frac{\theta_j - \theta_j^*}{\theta_j^* + \theta_i} \right),
\]

and therefore

\[
\frac{\partial}{\partial \theta_i} t_i(\theta_i, \theta_j) = \frac{\theta_j - \theta_j^*}{\theta_j^* + \theta_i} \alpha_i(\theta_i) - \frac{\theta_i + \theta_j - \theta_j^*}{\theta_j^* + \theta_i} = \frac{\theta_j - \theta_j^*}{\theta_j^* + \theta_i} \left[ \frac{\theta_j - \theta_j^*}{\theta_j^* + \theta_i} \log \left( 1 + \frac{\theta_j^* + \theta_i}{\theta_j - \theta_j^*} \right) - \frac{\theta_j + \theta_i}{\theta_j + \theta_i} \right].
\]
Since $\theta_i > -\theta_j^*$, the sign of $\frac{\partial}{\partial \theta_i} t_i (\theta_i, \theta_j)$ equals the sign of the term in brackets. Using the fact that $\frac{1}{x} \log (1 + x) < 1$ for all $x > 0$, this term is less than

$$\frac{\theta_j - \theta_j^*}{\theta_j - \theta_j^* - \theta_j + \theta_i} = - \frac{\theta_j - \theta_j^*}{\theta_j - \theta_j^*} \frac{\theta_j^* + \theta_i}{\theta_j + \theta_i} \leq 0,$$

where the last inequality again uses $\theta_i > -\theta_j^*$, hence, $t_i (\theta_i, \theta_j)$ is non-increasing in $\theta_i$. □

**Proof of Proposition 4.** It is clear that the mechanism satisfies E, (ex post) IR, and SBB. I now verify IC for an arbitrary type $\theta_i$.

If $\hat{\theta}_i \leq -\theta_j^*$, then $U_i (\hat{\theta}_i, \theta_j^*; \theta_i) = 0$, so ex post IR and $\delta_{\theta_j^*} \in \Delta_j'$ yield $U_i (\theta_i) \geq \inf_{\phi_j \in \Delta_j} (\hat{\theta}_i, \phi_j; \theta_i)$. If $\hat{\theta}_i > \theta_i$, then ex post IR, budget balance, and the fact that $t_i (\theta_i, \theta_j)$ is non-increasing in $\theta_i$ in the region where $\theta_i + \theta_j \geq 0$ (which follows as in the proof of Lemma 1) imply that $U_i (\theta_i, \theta_j; \theta_i) \geq U_i (\hat{\theta}_i, \theta_j; \theta_i)$ for all $\theta_j$, as in the proof of Theorem 1. Hence, $U_i (\theta_i) \geq \inf_{\phi_j \in \Delta_j} (\hat{\theta}_i, \phi_j; \theta_i)$. This completes the proof for IC for $\theta_i \leq -\theta_j^*$, so assume henceforth that $\theta_i > -\theta_j^*$.

If $\hat{\theta}_i \in (-\theta_j^*, \theta_i)$, then $\hat{\theta}_i > -\theta_j^* > -t_j^*$. Hence, $U_i (\hat{\theta}_i, \theta_j; \theta_i) \leq \tilde{U}_i (\hat{\theta}_i, \theta_j; \theta_i)$, where $\tilde{U}_i$ is utility under a standard reference rule (with $\varepsilon = 0$). Recalling that $\tilde{U}_i (\theta_i) \geq \inf_{\phi_j \in \Delta_j} \tilde{U}_i (\theta_i, \phi_j; \theta_i)$ by Theorem 2, I complete the proof of IC by showing that $U_i (\theta_i) = \tilde{U}_i (\theta_i)$.

I claim that

$$\inf_{\phi_j \in \Delta_j} U_i (\theta_i, \phi_j; \theta_i) = \frac{\theta_j^* + \min \{ \theta_i, -\hat{\theta}_j \} (\theta_i + t_j^*)}{\theta_j + \min \{ \theta_i, -\hat{\theta}_j \}}$$

for $\theta_j > -\theta_j^*$, whenever $\varepsilon < \min_{i \in \{1, 2\}} \theta_j + \max \{ \theta_i - \hat{\theta}_j \}$. In particular, I show that the infimum of $U_i (\theta_i, \phi_j; \theta_i)$ over all of $\Delta_j$ (and hence over $\Delta_j'$) is attained at $\phi_j = \delta_{\max \{ -\theta_i, \theta_j \}} \hat{\theta}_j$ (which is in $\Delta_j'$). To see this, first note that any $\phi_j \in \Delta_j$ must put positive mass on the interval $[-\theta_i, \hat{\theta}_j]$. Now $\theta_i$ gets positive payoff against all $\theta_j \in (-\theta_i, \hat{\theta}_j]$ and gets zero payoff against all $\theta_j \leq -\theta_i$, so if $\phi_j \in \Delta_j$ puts positive mass on $\theta_j < -\theta_i$ then there exists $\phi_j' \in \Delta_j$ that shifts this mass to $-\theta_i$ and reduces the mass on $[-\theta_i, \hat{\theta}_j]$, thus reducing $\theta_i$’s payoff. Next, if $\phi_j$ puts positive mass on $[-t_j^*, \hat{\theta}_j]$ then shifting this mass to $\hat{\theta}_j$ decreases the probability of trade and again reduces $\theta_i$’s payoff. Finally, if $\phi_j$ puts positive mass on $(-\theta_i, -t_j^*)$, with $E^{\phi_j} [\theta | \theta \in (-\theta_i, -t_j^*)] = \theta_j$, then it is worse for $\theta_i$ to split this mass between $\max \{ -\theta_i, \theta_j \}$.
and \( \bar{\theta}_j \) in the proportions that preserve the mean (which is possible because \( \bar{\theta}_j > -t^*_j = t^*_j \)). To see this, note that type \( \theta_i \)'s payoff against \( \theta_j \in (-\theta_i, -t^*_j) \) is

\[
\theta_i + (1 - \varepsilon) \theta_j - \varepsilon \min \{ \theta_i, -\bar{\theta}_j \},
\]

while type \( \theta_i \)'s payoff from facing \( \bar{\theta}_j \) with probability \( \frac{\theta_j + \min \{ \theta_i, -\bar{\theta}_j \}}{\theta_j + \min \{ \theta_i, -\bar{\theta}_j \}} \) is

\[
\frac{\theta_j + \min \{ \theta_i, -\bar{\theta}_j \}}{\theta_j + \min \{ \theta_i, -\bar{\theta}_j \}} (\theta_i + t^*_i).
\]

(5) is less than (4) if and only

\[
\varepsilon < \frac{\theta_i + \theta_j}{\min \{ \theta_i + \theta_j, \theta_j - \bar{\theta}_j \}} - \frac{\theta_i + t^*_i}{\min \{ \theta_i + \theta_j, \bar{\theta}_j - \theta_j \}}.
\]

The right-hand side of this expression is non-increasing in \( \theta_j \), so it is minimized over \( \theta_j \) at \( \theta_j = \bar{\theta}_j \), where it equals \( \frac{\theta_j - \bar{\theta}_j}{\min \{ \theta_i + \theta_j, \bar{\theta}_j - \theta_j \}} \). Finally, this term is minimized over \( \theta_i \) at \( \theta_i = \bar{\theta}_i \).

Hence, a sufficient condition for (5) to be less than (4) is \( \varepsilon < \frac{\theta_i + t^*_i}{\min \{ \theta_i + \theta_j, \bar{\theta}_j - \theta_j \}} \). Thus, any measure \( \phi_j \) that puts positive mass on the intervals \([\bar{\theta}_j, \theta_i], [-t^*_j, \bar{\theta}_j] \), or \((-\theta_i, -t^*_j) \) can be improved upon, so it follows that the infimum of \( U_i (\theta_i, \phi_j; \theta_i) \) over \( \Theta_j \) is attained at a measure that puts mass only on \( \{-\theta_i, \bar{\theta}_j \} \), and the only such measure is \( \delta_{\max \{-\theta_i, \bar{\theta}_j \}} \).

The above claim gives \( U_i (\theta_i) = \frac{\theta_j + \theta_i}{\theta_j + \theta_i} (\theta_i + t^*_i) \). Since a standard reference rule corresponds to \( \varepsilon = 0 \), the same argument gives \( \tilde{U}_i (\theta_i) = \frac{\theta_j + \theta_i}{\theta_j + \theta_i} (\theta_i + t^*_i) \). Hence, \( U_i (\theta_i) = \tilde{U}_i (\theta_i) \), completing the proof of IC.

Finally, to see that truth-telling is not weakly dominated for any type when \( \varepsilon > 0 \), first note that reporting \( \hat{\theta}_i > \theta_i \) cannot dominate truth-telling because \( U_i \left( \hat{\theta}_i, \theta_j; \theta_i \right) \leq U_i (\theta_i, \theta_j; \theta_i) \) for all \( \theta_j \), as noted above. Next, if \( \theta_i > -\bar{\theta}_j \), then reporting \( \hat{\theta}_i < \theta_i \) cannot dominate truth-telling because \( U_i \left( \theta_i, \frac{\theta_i + \hat{\theta}_i}{2}; \theta_i \right) > 0 = U_i \left( \hat{\theta}_i, \frac{\theta_i + \hat{\theta}_i}{2}; \theta_i \right) \). Finally, if \( \theta_i \leq -\bar{\theta}_j \), then reporting \( \hat{\theta}_j < \theta_i \) cannot dominate truth-telling because \( U_i (\theta_i, \theta_j; \theta_i) = U_i \left( \hat{\theta}_j, \theta_j; \theta_i \right) = 0 \) for all \( \theta_j \in \Theta_j \).

**Proof of Proposition 5.** Suppose there exists an \( \varepsilon \)-efficient and continuous mechanism \( x \) satisfying minmax IC, IR, and WBB. Continuity of the mechanism implies that \( U_i \left( \hat{\theta}_i, \phi_j; \theta_i \right) \) is continuous in \( \hat{\theta}_i \) and \( \phi_j \). By the maximum theorem, the “best-response correspondences” \( \arg \max_{\hat{\theta}_i} U_i \left( \hat{\theta}_i, \phi_j; \theta_i \right) \) and \( \arg \min_{\phi_j} U_i \left( \hat{\theta}_i, \phi_j; \theta_i \right) \) are nonempty, convex-valued, and upper hemicontinuous. Since \( \Theta_i \) and \( \Delta_j \) are nonempty, compact, and convex, the Kakutani-Fan-Glicksberg fixed point theorem now implies that a saddle point exists. Therefore, minmax IC and maxmin IC coincide.
The rest of the proof adapts the proof of the negative direction of Theorem 1. By \( \varepsilon \)-efficiency,
\[
\begin{align*}
\int_{t_i}^{\theta_i} \int_{t_j}^{\theta_j} y(\theta_i, \theta_j) (\theta_i + \theta_j) \, d\phi_j \, d\phi_i - \sum_{i=1}^{2} \int_{t_i}^{\theta_i} \int_{t_j}^{\theta_j} y(\theta_i, \theta_j) (1 - F_i(\theta_i)) \, d\phi_j \, d\theta_i \\
&\leq - (1 - \varepsilon) \int_{t_i}^{\theta_i} (1 - F_i(\theta)) (1 - F_j(-\theta)) \, d\theta + \varepsilon \\
&\leq \int_{t_i}^{\theta_i} (1 - F_i(\theta)) (1 - F_j(-\theta)) \, d\theta + 2\varepsilon.
\end{align*}
\]
Similarly, letting
\[
\hat{y}_i(\theta_i) = \inf_{\phi_j \in \Delta_j'} y(\theta_i, \phi_j),
\]
\( \varepsilon \)-efficiency implies that
\[
\hat{y}_i(\theta_i) \geq (1 - \varepsilon) \bar{y}_i(\theta_i - \varepsilon) \geq \bar{y}_i(\theta_i) - 2\varepsilon,
\]
where \( \bar{y}_i(\theta_i) \) is as in the proof of Theorem 1. Combining these observations with the proof of Theorem 1, it follows that if Condition (*) fails by more than \( 6\varepsilon \) (i.e., if the left-hand side of Condition (*) is less than \( 1 - 6\varepsilon \)), then the right-hand side of (3) is negative, which gives a contradiction. Hence, if Condition (*) fails by \( \Delta > 0 \), taking \( \varepsilon = \frac{\Delta}{6} \) completes the proof.

References


