Informationally Robust Trade Under Adverse Selection

Gabriel Carroll, Stanford University

gdc@stanford.edu

April 14, 2015

Abstract

A buyer and seller have the opportunity to exchange an indivisible good at a prespecified price. Each agent may be imperfectly informed, in an arbitrary way, about both his own value for the good and the other agent's value. In such environments, contagious adverse selection can potentially lead to breakdown of trade in complex patterns. Nonetheless, we show limits to this breakdown: an observer who knows only the distribution of values can predict some amount of (expected) trade that is attainable in equilibrium no matter what the information structure is. We show how to compute the sharpest such prediction.

Thanks to (in random order) Paul Milgrom, Alp Simsek, Stephan Lauermann, Alex Wolitzky, Matthew Jackson, Anton Tsoy, Benny Moldovanu, Nathan Hendren, Daron Acemoglu, Jean Tirole, Brett Green, Richard Holden, and anonymous referees for helpful comments.

1 Introduction

Imagine a buyer and seller who can potentially meet to exchange a single object at a fixed price. The value of the object to each party is uncertain. With high probability, both parties stand to gain from the trade; but each party also foresees some nonnegligible probability that the trade will hurt him. For example, suppose that the object can be traded at a price of 5, and there is an 80% chance that the world is in a "normal" state,

where the object is worth 6 to the buyer and 4 to the seller; but there is also a 10% chance of a "bad" state, where both parties' values are lower by 2 (so that ex post, the buyer loses from the trade), and a 10% chance of a "good" state, where both values are higher by 2 (so that the seller loses). Explicitly, the values are

$$\begin{array}{lll} (4,2) & \text{(bad)} & \text{with probability 10\%;} \\ (6,4) & \text{(normal)} & \text{with probability 80\%;} \\ (8,6) & \text{(good)} & \text{with probability 10\%.} \end{array}$$

If there is asymmetric information between the buyer and seller, this can lead to some breakdown of trade. A long literature in information economics since Akerlof [1] has emphasized the possibility of such a breakdown. Moreover, as more recent contributions have pointed out, this breakdown can be exacerbated by *contagion*. For example, if the seller has a private signal that the good state is likely, he may refuse to trade, since it is not in his interest. Then, if the buyer has a private signal that one of the extreme states has occurred (but he does not know which), he may worry that the seller is only willing to trade in the bad state, and this adverse selection makes the buyer in turn unwilling to trade. This behavior by the buyer can in turn make the seller unwilling to trade for some other realizations of his signal, and so forth. Thus, higher-order beliefs can play an important role in determining economic outcomes [17, 2, 14]. These patterns of contagion can be fairly subtle, as we shall momentarily illustrate with a more detailed example.

It seems, then, that we cannot predict the outcome of the interaction without knowing the details of the information structure: whether the seller is perfectly informed and the buyer completely uninformed, or vice versa, or each receives conditionally-independent noisy signals of the state, or perhaps something much more intricate. Unfortunately, information structures (and especially higher-order information) can be complex, and difficult for an outside observer to model. In this paper, we show how such an observer can nonetheless make *some* predictions about trade. In the above example, without knowing the information structure at all, the observer can predict that the parties will be able to realize the gains from trade with probability at least 60%. More precisely, no matter what the information structure is, as long as the buyer and seller share a common prior over it, the resulting Bayesian game between them has an equilibrium in which at least 60% of the gains from trade are realized (in expectation); and this 60% prediction is sharp.

The results of this paper will extend the example, and find the sharpest possible

prediction for attainable gains from trade that an observer can make, if the observer knows only the distribution of traders' values and not anything about the information structure. We will also describe the information structure that makes this prediction sharp. As it turns out, this worst-case information structure does not involve asymmetric information. Instead, both parties receive the same signal: either a "high-value" signal (in which case the seller does not want to trade at the posted price, because his expected value from keeping the good is higher); or a "low-value" signal (where the buyer does not want to trade); or a "normal" signal. (The fact that the worst case entails symmetric information is specific to the posted-price mechanism that we assume; this will be further discussed in the conclusion.)

On some level, our characterization of the worst case is no big surprise: If trade fails, it should be either because the seller expects his value is too high or the buyer expects his value is too low. But the conclusion is not trivial because of the interaction of the parties' information. It is *not* the case in equilibrium that the buyer simply refuses to trade when his expected value given his signal is below the price, and likewise for the seller. Instead, each conditions on the information content of the other's willingness to trade.

Here is a small example illustrating how equilibrium behavior can be subtle. We stick with the 80–10–10 distribution of values described above. Suppose that the information structure is as follows: The buyer's signal η_B may take one of three possible realizations, which we call A, B, C; the seller's signal η_S may take on realizations D, E, F. The probability of each pair of signals, and buyer's and seller's values for the good for each possible signal pair, are as shown in Table 1. (In this example, each possible pair of signals can occur for only one state of the world, but our general model will not assume this.)

$\eta_B \backslash \eta_S$	D	E	F
A	0.36 $6, 4$	0.03 $4, 2$	0.40 $6, 4$
B	0.04 $6, 4$	0.05 $4, 2$	
C		0.10 $8, 6$	0.02 $4, 2$

Table 1: Joint distribution of signals and values

The buyer and seller observe their respective signals, and then each decide whether to agree to trade or to stay out. If both agree, then they trade at the price of 5.

In this example, if the seller receives signal D or F, then for sure he benefits from trading, so we may as well assume he agrees to trade. Then, if the buyer receives signal A, his expected gain from agreeing to trade is positive (although its exact value depends how the seller with signal E behaves), so the buyer with signal A agrees as well.

Does the seller with signal E agree to trade? If he does, then we can check that the buyer with signal B prefers not to trade, while the buyer with signal C prefers to trade. Given that the buyer trades under signals A and C but not B, then the seller with signal E earns negative gains from trade, so prefers not to trade.

On the other hand, if the seller does not trade under signal E, then the buyer prefers to trade under signal B and not C. In this case, the seller's best reply under signal E is to trade.

So it must be in equilibrium that the seller mixes under signal E. With a little more calculation, we can check that the equilibrium is as follows: the buyer agrees to trade with probability $1/15$ following signal B, and probability 1 for signal C; and the seller agrees with probability $4/5$ under signal E. The resulting probability of both parties agreeing to trade is $667/750 \approx 0.89$.

This mixing gives us a clue that, in general, there is no easy recipe to compute the equilibrium for a given information structure, and suggests that there might perhaps exist more complex, email-game-like information structures [17] where contagion effects across signals lead to frequent breakdown of trade. Our results provide an answer to this concern, by showing that, from an ex-ante point of view, such contagion is limited.

Now that we have sketched out our results, it is tempting to discuss interpretations and possible applications. However, it will be easiest to give this discussion clearly after having given the full statement of the model and results, and indicating their limitations. So we leave the discussion of direct interpretations to the concluding Section 5 — with apologies to any hurried readers — and instead use the rest of this introduction to talk about the paper's methodological contribution and its context.

The broader question behind this paper is: In situations of uncertainty, what can we predict about economic interactions without knowing the details of the information structure? This question connects with the work of Bergemann and Morris, who take a similar approach at an abstract level to general static games [5] and apply it to games with a quadratic-normal structure [6], and with Bergemann, Brooks and Morris, who perform a similar analysis in a monopoly pricing problem [4] and a first-price auction [3]. Like the latter two papers in particular, we choose a relatively simple and common form of economic interaction and explore the possible information-free predictions.

A basic difference between our work and the others just mentioned is that the latter explore *all* possible equilibria, whereas we focus on the best equilibrium for each information structure. Indeed, in our setup, it is always an equilibrium for both parties to never agree to trade. Moreover, such bad equilibria cannot always be eliminated with a simple refinement (see Subsection 4.3). Hence if we allowed all equilibria, the observer could make no predictions about the realized gains from trade. Accordingly, our results are best interpreted not so much as a positive prediction about how much trade will actually happen, but rather as a study of the limits of informational contagion arguments.

This difference also means that we cannot use the same technical tools as in the Bergemann-Brooks-Morris work. They use linear programming methods, which are suitable for studying the set of all possible equilibria under different information structures, but not for identifying a particular equilibrium for each information structure, as we do here. Instead, we use a nonconstructive method, applying the Nash existence theorem to variant games.

Our contribution is also reminiscent of the work of Kajii and Morris [9] on ex-ante robustness to incomplete information. They consider Nash equilibria of complete-information games, and give conditions under which any nearby incomplete information game (whose payoffs equal those of the complete information game with high probability) must have a nearby equilibrium. Our results are connected if we think of the situation where both parties gain from trade ex-post (the "normal" state in the above example) as corresponding to the complete-information game, and both parties accepting trade as the complete-information equilibrium. In fact, the argument in Kajii and Morris leads to a sharp quantitative bound on how far the equilibrium can move when one introduces a given amount of incomplete information, if both the payoffs in the newly-introduced states and the information structure can be chosen arbitrarily. In contrast, for the particular trading game that we consider, we keep the payoff distribution fixed, and vary the information structure only, and give a corresponding sharp bound by different techniques.

2 Model

Let's now flesh out the formal model. The buyer's and seller's values for the good, b and s, are random variables whose joint distribution is given by a probability measure μ on $\mathbb{R}^2$, with compact support. This μ describes the prior belief, shared by the buyer, seller, and the outside observer. We assume $b \geq s$ with probability 1: it is common knowledge that there are (weak) gains from trade. (We will discuss later the consequences of relaxing

this assumption.)

We assume a very simple institution for trading. There is a known market price p, which is constant. Each of the two agents can either agree to trade at that price or decline to trade. If both agents agree, they trade, giving payoffs $b-p$ and $p-s$ to the buyer and seller respectively. If either declines, then both receive payoff 0.

We will assume that neither the buyer nor the seller is certain ex ante that trade is beneficial for him: the events $b-p<0$ and $p-s<0$ both have positive probability.

Both the buyer and seller may receive information prior to trading, via an *information structure* which is unknown to the observer. We restrict to finite information structures, to avoid complications with equilibrium existence. Thus, an information structure consists of two finite sets of signals, $\mathcal{I}_B$ and $\mathcal{I}_S$, and a joint probability measure ν on $\mathbb{R}^2 \times \mathcal{I}_B \times \mathcal{I}_S$, such that the marginal of ν on the $\mathbb{R}^2$ component coincides with μ. The signals received by the two agents will be denoted by $\eta_B \in \mathcal{I}_B$ and $\eta_S \in \mathcal{I}_S$.

Any information structure induces a Bayesian game, in which the two agents observe their signals and then decide whether to agree to trade. The buyer's possible (mixed) strategies are functions $\sigma_B : \mathcal{I}_B \to [0,1]$, denoting the probability of agreeing after each signal, and the seller's strategies are functions $\sigma_S : \mathcal{I}_S \to [0,1]$. The expected payoffs from a strategy profile are

$$u_B(\sigma_B,\sigma_S) = \int \sigma_B(\eta_B)\sigma_S(\eta_S)(b-p)\,d\nu, \qquad u_S(\sigma_B,\sigma_S) = \int \sigma_B(\eta_B)\sigma_S(\eta_S)(p-s)\,d\nu. \tag{2.1}$$

(Here and subsequently, all integrals are taken to be over the entire probability space unless indicated otherwise.) A strategy profile (σ_B,σ_S) is a (Bayesian Nash) *equilibrium* if

$$u_B(\sigma_B,\sigma_S) \geq u_B(\sigma'_B,\sigma_S) \qquad \text{and} \qquad u_S(\sigma_B,\sigma_S) \geq u_S(\sigma_B,\sigma'_S)$$

for any alternative strategies σ'_B, σ'_S.

The observer would like to make robust predictions about the best possible equilibrium. This could naturally be taken to mean the equilibrium that realizes the highest expected surplus; but we could also imagine other criteria, e.g. the highest probability of trade. We may as well give a formulation that allows for various such criteria, since it will not require much extra work. So, we assume the observer has an objective, represented by some bounded, measurable function of b,s, call it $w(b,s)$: the observer gets $w(b,s)$ when

trade occurs and 0 otherwise. Thus, the observer's criterion is

$$W(\sigma_B, \sigma_S) = \int \sigma_B(\eta_B)\sigma_S(\eta_S)w(b,s)\,d\nu.$$

For example, if we define $w(b,s) = b - s$ then this captures the expected gains from trade realized in equilibrium; if $w(b,s) = 1$ then we have the probability of trade. Other criteria might express the observer's placing more importance on trade in some states than in others. We do, however, need that the observer always prefers for trade to occur: $w(b,s) \geq 0$, for all (b,s) in the support of μ.

We then say that a value x for the observer's criterion is a *robust prediction* if, for every information structure $(\mathcal{I}_B, \mathcal{I}_S, \nu)$, there exists an equilibrium (σ_B, σ_S) satisfying $W(\sigma_B, \sigma_S) \geq x$.[1] It is immediate that there is some maximum robust prediction. We wish to characterize what this value is.

Our analysis will also lead us naturally to look at symmetric information structures, where both agents have the same information. Explicitly, we say the information structure is *symmetric* if $\mathcal{I}_B = \mathcal{I}_S$ and the measure ν places probability 1 on the event $\eta_B = \eta_S$. We say that a value x is a *robust prediction under symmetric information* if, for every symmetric information structure $(\mathcal{I}_B, \mathcal{I}_S, \nu)$, there exists an equilibrium (σ_B, σ_S) satisfying $W(\sigma_B, \sigma_S) \geq x$.

3 Results

3.1 Measures and decompositions

Let's jump to the punch line. To identify how good or bad an equilibrium outcome is (from the observer's point of view), it suffices to describe when the agents fail to trade. Our two main results describe these possible no-trade events. The first main result says that for any information structure, there exists an equilibrium in which the event of no trade is at most the union of two other events, one on which the buyer has a negative expected gain from trading at price p, and one on which the seller has a negative expected gain from trading. (Note that this result is just a characterization of the overall no-trade

[1] The term "prediction" is a bit of a misnomer, for two reasons: first, as already pointed out, we may not be confident in predicting that this equilibrium will actually occur; and second, "prediction" may suggest a point estimate, whereas in our language, if x is a robust prediction then any lower value is as well. A name like "robustly attainable value" or "robustly surpassable value" might be more descriptive. But we keep "prediction" for simplicity.

event: It does *not* say that in equilibrium, the buyer declines trade on the first sub-event and the seller declines on the second.) The second main result is a sort of converse: for any event that has such a decomposition into two sub-events, there is an information structure under which no trade can occur there. Thus, together, these two results characterize the maximal possible no-trade events. Moreover, in the second result, one can choose the information structure to be symmetric — that is, both players have identical information. We will give the results, and then, before proceeding to the proofs (Subsection 3.3), will first detail how they can be used to compute the maximum robust prediction (Subsection 3.2).

Although the above verbal description is in terms of events, it will actually be more convenient to work with (sub-probability) measures on $\mathbb{R}^2$. This will allow us to describe where the mass of (buyer, seller) values corresponding to an event is distributed, without needing to keep track of the underlying probability space.

For example, given an information structure, and given (mixed) strategies (σ_B, σ_S), we can define a measure μ_T on $\mathbb{R}^2$ by

$$\mu_T(E) = \int \sigma_B(\eta_B)\sigma_S(\eta_S)\, \mathbb{1}((b,s) \in E)\, d\nu,$$

for any measurable $E \subseteq \mathbb{R}^2$. So for any E, $\mu_T(E)$ gives the probability that the pair of values (b,s) is in E *and* the parties trade. This is in contrast to $\mu(E)$, which simply gives the probability that $(b,s) \in E$. Likewise, we can define a measure μ_{NT} by

$$\mu_{NT}(E) = \int (1 - \sigma_B(\eta_B)\sigma_S(\eta_S))\, \mathbb{1}((b,s) \in E)\, d\nu.$$

This is the probability that $(b,s) \in E$ and trade does not occur. We call μ_T and μ_{NT} the *trade measure* and *no-trade measure* associated to strategies (σ_B, σ_S), and note that $\mu_T + \mu_{NT} = \mu$. Note also that we can rewrite the observer's criterion as

$$W(\sigma_B, \sigma_S) = \int w(b,s)\, d\mu_T. \tag{3.1}$$

To describe the decompositions used in our results, we again consider pairs of measures on $\mathbb{R}^2$, describing how a portion of the probability mass of values is distributed. Given a pair (μ_B, μ_S) of such measures, call it a *negative-gains pair* if the following three conditions hold: $\int (b-p)\, d\mu_B < 0$, $\int (p-s)\, d\mu_S < 0$, and $\mu_B + \mu_S \leq \mu$ (that is, $\mu_B(E) + \mu_S(E) \leq \mu(E)$ for every event E). Then, say that a measure μ' has a *negative-gains decomposition* if it

can be written as $\mu' = \mu_B + \mu_S$ for some negative-gains pair (μ_B, μ_S).

The first main result is then as follows:

Proposition 3.1. *Let $(\mathcal{I}_B, \mathcal{I}_S, \nu)$ be any information structure. There exists an equilibrium (σ_B, σ_S), whose no-trade measure satisfies $\mu_{NT} \leq \mu'$, for some μ' that has a negative-gains decomposition.*

The converse proposition says that, given a measure that has a negative-gains decomposition, we can find an information structure where trade necessarily breaks down at least that often in any equilibrium.

Proposition 3.2. *Let μ' be a measure that has a negative-gains decomposition. Then there exists an information structure such that, in any equilibrium, the no-trade measure μ_{NT} satisfies $\mu_{NT} \geq \mu'$. Moreover, we can take this information structure to be symmetric.*

With these results stated, the next task is to show how they can be used to compute the maximum robust prediction for the observer's criterion, given the prior μ. This computation begins with the observation below:

Corollary 3.3. *The following are equivalent, for a real number x:*

(a) x is a robust prediction;

(b) x is a robust prediction under symmetric information;

(c) $x \leq \int w(b,s)\, d\mu - \sup_{\mu'} \int w(b,s)\, d\mu'$, where the supremum is over all measures μ' having a negative-gains decomposition.

Proof: Clearly (a) implies (b): if a prediction of x is valid for any arbitrary information structure, it is valid for any symmetric information structure.

For (b) implies (c), suppose that the conclusion (c) fails to hold; then $x > \int w(b,s)\, d\mu - \int w(b,s)\, d\mu'$ for some particular μ' that has a negative-gains decomposition. So, by Proposition 3.2, there exists a symmetric information structure where, in every equilibrium, the no-trade measure is at least as large as μ', and therefore the trade measure μ_T is $\leq \mu - \mu'$. So, by (3.1) (and the fact that $w \geq 0$ everywhere), in every equilibrium, the observer's criterion is $W(\sigma_B, \sigma_S) \leq \int w(b,s)\, d(\mu - \mu') < x$. Thus, x is not a robust prediction under symmetric information.

For (c) implies (a), suppose x satisfies the given condition. In any information structure, Proposition 3.1 gives an equilibrium (σ_B, σ_S), whose trade measure μ_T satisfies

$\mu_T \geq \mu - \mu'$, for some μ' that has a negative-gains decomposition. Then, using (3.1),

$$W(\sigma_B, \sigma_S) \geq \int w(b,s)\, d\mu - \int w(b,s)\, d\mu' \geq x$$

where the second inequality comes from the assumption in (c). So x is a robust prediction. □

3.2 Maximal no-trade measures

It remains, then, to calculate the supremum in (c) of Corollary 3.3 — the worst total value, as measured by the observer's criterion w, that can be stuck in a measure with a negative-gains decomposition. Or, equivalently, this is the worst total value of w that can be packed into two measures carved out of the prior μ, with one measure being negative-gain for the buyer and the other negative-gain for the seller. (Because of the strict inequalities, the worst case is approached but not actually attained; but we will disregard this detail in the informal description here.)

We first intuitively describe the worst possible such measures in the benchmark case where the observer is concerned with expected gains from trade, $w(b,s) = b - s$. The worst-case μ_B consists of as much total value of $b-s$ as possible, subject to the constraint that the total value of $b - p$ should be negative. This is constructed by grabbing the mass from μ that comes as "cheaply" as possible, i.e. for which the ratio $(b-p)/(b-s)$ is as low as possible, up until the $\int (b-p)\, d\mu_B < 0$ constraint becomes binding. That is, μ_B simply consists of the mass from μ lying in the region $(b-p)/(b-s) < \alpha_B$, for some threshold α_B. This region of (b,s)-space is shown by the horizontally-hatched area in Figure 1 (where the gray heat map represents the density of the prior distribution μ; note we ignore the upper-left half-space $b < s$ since values in that half-space never occur). How is the threshold value α_B determined? It is the value for which the integral of $b-p$ over this region is zero. Similarly, the worst-case μ_S consists of the mass from μ lying in the region $(b-p)/(b-s) > \alpha_S$ (the diagonally-hatched area in the figure), with the threshold α_S determined by the condition that the integral of $p - s$ over this region is zero.

Thus, the worst-case μ' consists of all the mass from μ lying in either of the two hatched regions in the figure. If these two regions were to overlap, then the worst-case μ' would equal μ, i.e. the worst-case prediction would be zero trade.

For more general criteria w, it will still hold that the worst-case μ_B and μ_S are sepa-

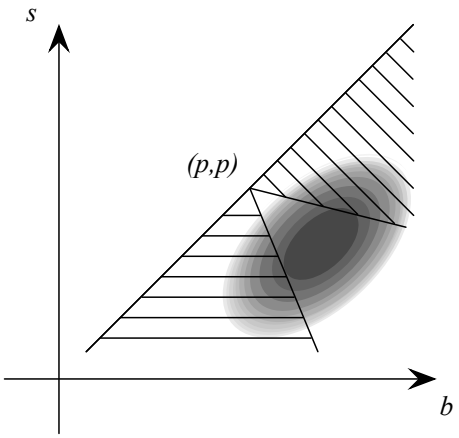

Figure 1: Worst-case negative-gains measures (criterion = gains from trade)

rated in terms of the ratio $(b-p)/(b-s)$, which we will call the *buyer-gains ratio* since it represents the share of gains from trade accruing to the buyer. In general, μ_B will place all its weight in a region of (b,s)-space where the buyer-gains ratio is low, and μ_S will place all its weight in a disjoint region where the buyer-gains ratio is high. But then, within these respective regions, μ_B will consist of mass from μ for which $(b-p)/w$ is as low as possible, and μ_S will consist of mass for which $(p-s)/w$ is as low as possible. All of this will be illustrated through examples in Section 4 ahead.

The rest of this subsection will formalize these ideas in detail; it can be skipped on a casual reading. We first show that the worst-case negative-gain measures μ_B and μ_S can be separated by buyer-gains ratios as above. Explicitly, we show that there exists some $\alpha \in [0,1]$ such that μ_B puts weight only on value pairs (b,s) with $(b-p) \leq \alpha(b-s)$, and μ_S puts weight only on pairs with $(b-p) \geq \alpha(b-s)$. Moreover, if equality holds for a positive mass of value pairs (a detail omitted from the above description), then there exists $\beta \in [0,1]$ such that these pairs all contribute at most a share β of their mass to μ_B and at most $1-\beta$ of their mass to μ_S.

To state the separation lemma explicitly, given $\alpha, \beta \in [0,1]$, we define the following events in $\mathbb{R}^2$:

$$\begin{aligned} E^{\alpha}_{<} &= \{(b,s) \mid (b-p) < \alpha(b-s)\}, \\ E^{\alpha}_{=} &= \{(b,s) \mid (b-p) = \alpha(b-s)\}, \\ E^{\alpha}_{>} &= \{(b,s) \mid (b-p) > \alpha(b-s)\}, \end{aligned}$$

and define two measures $\overline{\mu}_B^{\alpha,\beta}, \overline{\mu}_S^{\alpha,\beta}$ by

$$\overline{\mu}_B^{\alpha,\beta}(E) = \mu(E \cap E_<^\alpha) + \beta\mu(E \cap E_=^\alpha), \qquad \overline{\mu}_S^{\alpha,\beta}(E) = \mu(E \cap E_>^\alpha) + (1-\beta)\mu(E \cap E_=^\alpha) \quad (3.2)$$

for any event E. Note that $\overline{\mu}_B^{\alpha,\beta} + \overline{\mu}_S^{\alpha,\beta} = \mu$. (We may write these without the superscript α, β.)

We will then say that a pair of measures (μ_B, μ_S) is (α, β)-*separated* if $\mu_B \le \overline{\mu}_B^{\alpha,\beta}$ and $\mu_S \le \overline{\mu}_S^{\alpha,\beta}$.

Lemma 3.4. *Let (μ_B, μ_S) be a negative-gains pair. Then there exists a negative-gains pair $(\widehat{\mu}_B, \widehat{\mu}_S)$ that is (α, β)-separated, for some α, β, and such that $\widehat{\mu}_B + \widehat{\mu}_S = \mu_B + \mu_S$.*

The proof is mechanical: With (μ_B, μ_S) given, any choice of parameters α, β specifies a way of redividing the mass $\mu' = \mu_B + \mu_S$ into $\widehat{\mu}_B$ and $\widehat{\mu}_S$. There is some range of (α, β) for which the needed inequality $\int (b-p)\, d\widehat{\mu}_B < 0$ is satisfied, and a corresponding range for $\widehat{\mu}_S$; we just need to show that these two parameter ranges overlap. The details are in Appendix A.

Lemma 3.4 shows that in our search for the supremum of $\int w(b,s)\, d\mu'$ over measures that have a negative-gains decomposition, we can restrict ourselves to decompositions that are (α, β)-separated for some α, β.

So, for any given α, β, define $Y(\alpha, \beta)$ to be the supremum of $\int w(b,s)\, d(\mu_B + \mu_S)$ over negative-gains pairs that are (α, β)-separated. We just need a way to compute $Y(\alpha, \beta)$ for given α and β, and then in a subsequent round we optimize over α, β.

It is evident that

$$Y(\alpha, \beta) = \sup_{\mu_B} \int w(b,s)\, d\mu_B + \sup_{\mu_S} \int w(b,s)\, d\mu_S,$$

where the first supremum is over all measures $\mu_B \le \overline{\mu}_B^{\alpha,\beta}$ satisfying $\int (b-p)\, d\mu_B < 0$, and the second is over all measures $\mu_S \le \overline{\mu}_S^{\alpha,\beta}$ satisfying $\int (p-s)\, d\mu_S < 0$: any two measures satisfying these bounds do indeed form a negative-gains pair. We denote these two separate suprema by $Y_B(\alpha, \beta), Y_S(\alpha, \beta)$.

These separate suprema can be calculated by the greedy algorithm that takes mass that (for μ_B) minimizes the ratio $(b-p)/w$, up until the point where the total integral of $b-p$ is zero; or (for μ_S) minimizes $(p-s)/w$, up until the integral of $p-s$ is zero.

Let us give a precise statement. For $\gamma > 0$ and $\delta \in [0,1]$, define

$$F_<^\gamma = \{(b,s) \mid b - p < \gamma w(b,s)\}, \qquad F_=^\gamma = \{(b,s) \mid b - p = \gamma w(b,s)\}.$$

Then $F^\gamma_<$ is increasing in γ, and so $\int_{F^\gamma_<}(b-p)\,d\overline{\mu}_B$ is also (weakly) increasing in γ, since the pairs that are in $F^\gamma_<$ but not in $F^{\gamma'}_<$ for $\gamma' < \gamma$ must satisfy $b-p \geq 0$. The integral is also left-continuous in γ. Let $\gamma^*_B \in (0,\infty]$ be the supremum of values such that $\int_{F^\gamma_<}(b-p)\,d\overline{\mu}_B < 0$. (This integral is negative for small enough $\gamma > 0$, so we are assured that $\gamma^*_B > 0$.) If $\gamma^*_B < \infty$ then the expression $\int_{F^{\gamma^*_B}_<}(b-p)\,d\overline{\mu}_B + \delta \int_{F^{\gamma^*_B}_=}(b-p)\,d\overline{\mu}_B$ is weakly increasing in $\delta \in [0,1]$, and is nonnegative at $\delta = 1$; let δ^*_B be the supremum of values for which it is < 0. The expression must be equal to 0 at $\delta = \delta^*_B$.

To optimize the seller's negative-gain measure μ_S, we perform a completely analogous computation, substituting $p - s$ for $b - p$ and $\overline{\mu}_S$ for $\overline{\mu}_B$, and defining events

$$G^\gamma_< = \{(b,s) \mid p - s < \gamma w(b,s)\}, \qquad G^\gamma_= = \{(b,s) \mid p - s = \gamma w(b,s)\}.$$

This gives values γ^*_S and δ^*_S.

Lemma 3.5. *If $\gamma^*_B = \infty$ then $Y_B(\alpha,\beta) = \int_{\mathbb{R}^2} w(b,s)\,d\overline{\mu}_B$. Otherwise,*

$$Y_B(\alpha,\beta) = \int_{F^{\gamma^*_B}_<} w(b,s)\,d\overline{\mu}_B + \delta^*_B \int_{F^{\gamma^*_B}_=} w(b,s)\,d\overline{\mu}_B.$$

Similarly, if $\gamma^*_S = \infty$ then $Y_S(\alpha,\beta) = \int_{\mathbb{R}^2} w(b,s)\,d\overline{\mu}_S$, and otherwise

$$Y_S(\alpha,\beta) = \int_{G^{\gamma^*_S}_<} w(b,s)\,d\overline{\mu}_S + \delta^*_S \int_{G^{\gamma^*_S}_=} w(b,s)\,d\overline{\mu}_S.$$

The proof is in Appendix A.

Finally, we can summarize our work in the following procedure to compute the observer's maximum robust prediction, given the prior distribution μ.

1. For each choice of $\alpha, \beta \in [0,1]$, split μ into $\overline{\mu}_B$ and $\overline{\mu}_S$ by (3.2).
2. Use the greedy algorithm on this $\overline{\mu}_B$ and $\overline{\mu}_S$ — taking the mass with the lowest buyer-gains ratio $(b-p)/w$ and $(p-s)/w$, respectively — to compute $Y_B(\alpha,\beta)$ and $Y_S(\alpha,\beta)$, as described in Lemma 3.5. This determines $Y(\alpha,\beta) = Y_B(\alpha,\beta) + Y_S(\alpha,\beta)$ for the given α and β.
3. Finally, as given by Corollary 3.3, the maximum robust prediction equals $\int w(b,s)\,d\mu - \sup_{\alpha,\beta} Y(\alpha,\beta)$.

We note that the brief description given earlier for the benchmark case $w(b,s) = b - s$ — where the measure μ_B is formed by restricting μ to the value pairs (b,s) with the lowest

ratio $(b-p)/(b-s)$, and μ_S is formed by restricting to the value pairs with the highest ratio — immediately follows as a special case.

3.3 Proofs of main results

We now turn to the proofs of the main results.

The proof of Proposition 3.1 — existence of a "good" equilibrium for any information structure — is nonconstructive. We consider a sequence of constrained games, where some of the possible signal realizations are "locked"; when a player receives a locked signal, we require him to agree to trade. Initially, all values of the signals are locked. We then gradually unlock the signal values one by one, and apply the Nash existence theorem to each such constrained game. As long as the equilibrium of the constrained game is not also an equilibrium of the unconstrained game, it must be that one player or the other wishes to decline trade at a signal that is currently locked. This fact can be written as an inequality. As we gradually unlock the signals, we obtain a succession of such inequalities, and combining these inequalities leads to our result.

Proof of Proposition 3.1: We successively define sequences of signal sets $\mathcal{J}_B^k \subseteq \mathcal{I}_B, \mathcal{J}_S^k \subseteq \mathcal{I}_S$ and functions $\lambda_B^k, \lambda_S^k : \mathcal{I}_B \times \mathcal{I}_S \to [0,1]$, for $k = 0, 1, \ldots$. These sets and functions will be made to satisfy the following conditions:

(a) $\lambda_B^k(\eta_B, \eta_S) = 0$ whenever $\eta_B \in \mathcal{J}_B^k$;

(b) $\lambda_S^k(\eta_B, \eta_S) = 0$ whenever $\eta_S \in \mathcal{J}_S^k$;

(c) if $(\eta_B, \eta_S) \notin \mathcal{J}_B^k \times \mathcal{J}_S^k$, then $\lambda_B^k(\eta_B, \eta_S) + \lambda_S^k(\eta_B, \eta_S) \geq 1$;

(d) if $\mathcal{J}_B^k \neq \mathcal{I}_B$, then $\int \lambda_B^k(\eta_B, \eta_S) \cdot (b-p)\, d\nu < 0$;

(e) if $\mathcal{J}_S^k \neq \mathcal{I}_S$, then $\int \lambda_S^k(\eta_B, \eta_S) \cdot (p-s)\, d\nu < 0$.

$\mathcal{J}_B^k$ will be the set of signal realizations for the buyer that are locked in the kth constrained game; similarly for the seller and $\mathcal{J}_S^k$. λ_B^k and λ_S^k will be weights derived from the deviation inequalities along the way.

For the base case, we take $\mathcal{J}_B^0 = \mathcal{I}_B$, $\mathcal{J}_S^0 = \mathcal{I}_S$, and λ_B^0, λ_S^0 identically zero. It is clear that (a) and (b) hold, and (c)–(e) are vacuous.

Now suppose these sets and functions have been defined for some k. Consider the Bayesian game where each player learns his signal according to ν, and agrees or declines to trade, with the constraint that the buyer must agree to trade whenever $\eta_B \in \mathcal{J}_B^k$, and

likewise the seller must agree whenever $\eta_S \in \mathcal{J}_S^k$. That is, the (mixed) strategy space of the buyer is the set of $\sigma_B : \mathcal{I}_B \to [0,1]$ such that $\sigma_B(\eta_B) = 1$ whenever $\eta_B \in \mathcal{J}_B^k$, and likewise for the seller; and the payoffs are given by (2.1). This game has a Bayesian Nash equilibrium, call it (σ_B, σ_S).

Suppose that (σ_B, σ_S) is not an equilibrium of the original, unconstrained game. In this case we will define $\mathcal{J}_B^{k+1}, \mathcal{J}_S^{k+1}, \lambda_B^{k+1}, \lambda_S^{k+1}$. One of the players has a profitable deviation, say the buyer (the argument if it is the seller is totally analogous). In particular, there is at least one signal η_B^* on which the buyer benefits from deviating. That is, there is σ_B' that agrees with σ_B for all signals except η_B^*, and such that

$$u_B(\sigma_B', \sigma_S) > u_B(\sigma_B, \sigma_S). \tag{3.3}$$

We must have $\eta_B^* \in \mathcal{J}_B^k$, because otherwise the deviation σ_B' would be allowed in the constrained game, and (3.3) contradicts the assumption that (σ_B, σ_S) was an equilibrium of the constrained game. Therefore, $\sigma_B(\eta_B^*) = 1$, and $\sigma_B'(\eta_B^*) < 1$. So (3.3) implies

$$\int_{\eta_B = \eta_B^*} \sigma_S(\eta_S)(b - p)\, d\nu < 0. \tag{3.4}$$

Define $\mathcal{J}_B^{k+1} = \mathcal{J}_B^k \setminus \{\eta_B^*\}$, and define

$$\lambda_B^{k+1}(\eta_B, \eta_S) = \begin{cases} \sigma_S(\eta_S) & \text{if } \eta_B = \eta_B^*, \\ \lambda_B^k(\eta_B, \eta_S) & \text{otherwise.} \end{cases}$$

Also define $\mathcal{J}_S^{k+1} = \mathcal{J}_S^k$ and $\lambda_S^{k+1} = \lambda_S^k$.

We check that (a)-(e) are satisfied for step $k+1$. It is straightforward to see that (a) for $k+1$ follows from (a) for k. For (c), we only need to check the cases where $\eta_B = \eta_B^*$. There are two possibilities. If $\eta_S \notin \mathcal{J}_S^k$, then

$$\begin{aligned} \lambda_B^{k+1}(\eta_B, \eta_S) + \lambda_S^{k+1}(\eta_B, \eta_S) &\geq \lambda_B^k(\eta_B, \eta_S) + \lambda_S^{k+1}(\eta_B, \eta_S) \\ &= \lambda_B^k(\eta_B, \eta_S) + \lambda_S^k(\eta_B, \eta_S) \\ &\geq 1. \end{aligned}$$

Here the first line is because $\lambda_B^k(\eta_B, \eta_S) = 0$ (by (a) for k); the second is because $\lambda_S^{k+1} = \lambda_S^k$; the third is by (c) for k. If on the other hand $\eta_S \in \mathcal{J}_S^k$, then $\lambda_B^{k+1}(\eta_B, \eta_S) = \sigma_S(\eta_S) = 1$

already. So (c) holds. For (d), we already know $\int \lambda_B^k(\eta_B, \eta_S)(b-p)\,d\nu \leq 0$. And

$$\begin{aligned}\int \lambda_B^{k+1}(\eta_B, \eta_S)(b-p)\,d\nu - \int \lambda_B^k(\eta_B, \eta_S)(b-p)\,d\nu \\ &= \int_{\eta_B=\eta_B^*} (\lambda_B^{k+1}(\eta_B, \eta_S) - \lambda_B^k(\eta_B, \eta_S))(b-p)\,d\nu \\ &= \int_{\eta_B=\eta_B^*} \sigma_S(\eta_S)(b-p)\,d\nu \\ &< 0\end{aligned}$$

by (3.4). Finally, (b) and (e) hold since $\mathcal{J}_S^{k+1} = \mathcal{J}_S^k$ and $\lambda_S^{k+1} = \lambda_S^k$.

Now, at each step k of this construction, the sets $\mathcal{J}_B^k, \mathcal{J}_S^k$ become weakly smaller, and one of them becomes strictly smaller. By finiteness, the process must stop at some k. This can only happen when the constrained equilibrium (σ_B, σ_S) is an equilibrium of the unconstrained game. This will be the equilibrium claimed in the proposition, so we focus now on this k and these strategies. We need to show that the no-trade measure is bounded above by $\mu_B + \mu_S$, for some negative-gains pair (μ_B, μ_S).

First, we can change λ_B^k and λ_S^k if necessary so that the inequality in condition (c) becomes an equality. To see this, consider any $(\eta_B^*, \eta_S^*) \notin \mathcal{J}_B^k \times \mathcal{J}_S^k$. At least one of

$$\int_{(\eta_B,\eta_S)=(\eta_B^*,\eta_S^*)} (b-p)\,d\nu, \qquad \int_{(\eta_B,\eta_S)=(\eta_B^*,\eta_S^*)} (p-s)\,d\nu$$

is nonnegative, since their sum is nonnegative. If the former, we can replace $\lambda_B^k(\eta_B^*, \eta_S^*)$ by the lower value $1 - \lambda_S^k(\eta_B^*, \eta_S^*)$ (keeping all other values of λ_B^k the same); this will make (c) hold with equality at this pair and will preserve (d) since the left side of the inequality there becomes weakly smaller. Likewise, in the latter case we replace $\lambda_S^k(\eta_B^*, \eta_S^*)$ by $1 - \lambda_B^k(\eta_B^*, \eta_S^*)$. Doing this for each signal pair, we ensure that (c) is an equality for each signal pair where it applies, without violating any of the other conditions.

Now suppose momentarily that $\mathcal{J}_B^k \neq \mathcal{I}_B$ and $\mathcal{J}_S^k \neq \mathcal{I}_S$. Define our measures μ_B and μ_S by

$$\mu_B(E) = \int \lambda_B^k(\eta_B, \eta_S)\,\mathbb{1}((b,s) \in E)\,d\nu,$$

$$\mu_S(E) = \int \lambda_S^k(\eta_B, \eta_S)\,\mathbb{1}((b,s) \in E)\,d\nu.$$

Then,

$$\int (b-p)\,d\mu_B = \int \lambda_B^k(\eta_B, \eta_S)(b-p)\,d\nu < 0$$

by (d), and

$$\int (p-s)\,d\mu_S = \int \lambda_S^k(\eta_B,\eta_S)(p-s)\,d\nu < 0$$

by (e). Moreover, for any event $E \subseteq \mathbb{R}^2$,

$$\mu_B(E)+\mu_S(E) = \int (\lambda_B^k(\eta_B,\eta_S)+\lambda_S^k(\eta_B,\eta_S))\mathbb{1}((b,s)\in E)\,d\nu \le \int \mathbb{1}((b,s)\in E)\,d\nu = \mu(E),$$

because the equality in (c), together with (a) and (b), ensures that $\lambda_B^k+\lambda_S^k \le 1$ everywhere. Thus, $\mu_B+\mu_S \le \mu$, and so (μ_B,μ_S) form a negative-gains pair.

We need to check that in our equilibrium, the resulting no-trade measure satisfies $\mu_{NT} \le \mu_B+\mu_S$. From the definitions, we can see that this is equivalent to checking

$$1-\sigma_B(\eta_B)\sigma_S(\eta_S) \le \lambda_B^k(\eta_B,\eta_S)+\lambda_S^k(\eta_B,\eta_S) \tag{3.5}$$

for all η_B,η_S. If $(\eta_B,\eta_S)\in \mathcal{J}_B^k\times\mathcal{J}_S^k$, then the left side is 0 by definition of the constrained game, and the right side is also 0 by (a–b). Otherwise, the left side is ≤ 1 and the right side is 1 by (c). So (3.5) indeed holds.

This proves the proposition if $\mathcal{J}_B^k \ne \mathcal{I}_B$ and $\mathcal{J}_S^k \ne \mathcal{I}_S$.

Finally, if $\mathcal{J}_B^k$ is all of $\mathcal{I}_B$ or $\mathcal{J}_S^k$ is all of $\mathcal{I}_S$, then we can use the same construction, but then we will have $\int(b-p)\,d\mu_B = 0$ or $\int(p-s)\,d\mu_S = 0$, respectively, instead of < 0 as needed. We can fix this using small adjustments. If $\int(b-p)\,d\mu_B = 0$ but $\int(p-s)\,d\mu_S < 0$, we can take a small amount of probability mass from μ where $b-p<0$, and either place it in μ_B or, if this mass already belongs to μ_S, then move it from μ_S to μ_B. If $\int(b-p)\,d\mu_B < 0$ and $\int(p-s)\,d\mu_S = 0$, then we similarly place a small probability mass in μ_S. Finally, if both of the gains integrals are zero, then $\mathcal{J}_B^k = \mathcal{I}_B$ and $\mathcal{J}_S^k = \mathcal{I}_S$, so that μ_B and μ_S as originally defined are identically zero; in this case we adjust them by just taking any small amount of probability mass with $b-p<0$ and placing it in μ_B, and likewise with $p-s<0$ for μ_S. In all cases, after the adjustment we will have both gains integrals $\int(b-p)\,d\mu_B$, $\int(p-s)\,d\mu_S$ strictly negative, and we still have $\mu_{NT} \le \mu_B+\mu_S \le \mu$, which is what we need. □

It now remains to prove Proposition 3.2, on existence of an information structure forcing some amount of no-trade. In contrast to the above proof, this one will consist of a very simple construction: Let (μ_B,μ_S) be the negative-gains decomposition of the given measure μ'. Then, for the signal structure, simply have both players observe whether they end up in μ_B, μ_S, or neither.

Proof of Proposition 3.2: Let (μ_B, μ_S) be the negative-gains decomposition of μ', and put $\mu_O = \mu - \mu'$. Let $\mathcal{I}_B = \mathcal{I}_S = \{O, B, S\}$ be the set of signals. Define the measure ν on $\mathbb{R}^2 \times \mathcal{I}_B \times \mathcal{I}_S$ as follows: for any event $E \subseteq \mathbb{R}^2$,

$$\nu(E \times \{(O,O)\}) = \mu_O(E), \qquad \nu(E \times \{(B,B)\}) = \mu_B(E), \qquad \nu(E \times \{(S,S)\}) = \mu_S(E),$$

and ν puts zero mass on all other signal pairs. Evidently, ν is a probability measure whose marginal on $\mathbb{R}^2$ is $\mu_O + \mu_B + \mu_S = \mu$. Thus, we have an information structure, and it is symmetric. For any equilibrium (σ_B, σ_S), we must have $\sigma_B(B)\sigma_S(B) = 0$: If $\sigma_S(B) > 0$, then the buyer will strictly prefer not to trade when he receives signal B, since his gain from agreeing is $\int \sigma_S(B)(b-p)d\mu_B < 0$. Likewise, $\sigma_B(S)\sigma_S(S) = 0$. It follows that the resulting no-trade measure satisfies $\mu_{NT} \geq \mu_B + \mu_S = \mu'$. □

3.4 Comments on sign criteria

Now that we have those proofs finished, we briefly discuss the consequences of relaxing some of the assumptions on signs.

Both parties' gains uncertain. We have assumed that the events $b - p < 0$ and $p - s < 0$ both have positive probability under μ. What happens when one has probability zero? If, say, $p - s \geq 0$ for certain, then the seller is always willing to accept trade. On any information structure, constraining the seller to always accept, and having the buyer choose a best response, gives an equilibrium. So Proposition 3.1 becomes simpler: there exists a measure $\mu_B \leq \mu$ for which the integral of $b - p$ is negative, and $\mu_{NT} = \mu_B$.

The converse, analogous to Proposition 3.2, says now that for any measure $\mu_B \leq \mu$ for which $\int (b-p)\, d\mu_B$ is negative, there is a (symmetric) information structure for which the no-trade measure is always at least as large as μ_B. Hence, to compute the maximum robust prediction, we just need to compute the supremum of $\int w(b,s)\, d\mu_B$ over measures with $\int (b-p)\, d\mu_B < 0$ — which we do by the greedy algorithm — and then subtract it from $\int w(b,s)\, d\mu$.

Of course, if both $b - p \geq 0$ and $p - s \geq 0$ for certain, then it is always an equilibrium for both agents always to trade.

Observer prefers trade. We have required the observer's criterion w to be non-negative. What if w could be negative — for example, the observer is concerned with the buyer's expected payoff in equilibrium? Then Corollary 3.3 no longer determines exactly the maximum robust prediction, because of a gap between Propositions 3.1 and 3.2. Proposition 3.1 says that the no-trade measure μ_{NT} is *bounded above* by μ' that has

a negative-gains decomposition. If w is nonnegative, then for any given μ', the worst case is to have μ_{NT} actually equal to μ', and Proposition 3.2 says there is an indeed an information structure that forces this. However, if w can have negative values, then the worst-case no-trade measure may be strictly smaller than μ', picking out only the realizations of (b, s) where w has positive values.

It is still possible to use Proposition 3.1 to give a nontrivial robust prediction for the observer's payoff, in many cases: an analogue to Corollary 3.3 implies the observer's criterion must be at least $\int w(b,s)\,d\mu - \sup_{\mu'} \int \max\{w(b,s), 0\}\,d\mu'$. But this robust prediction might not be best possible.

Aggregate gains from trade. We have also assumed common knowledge of gains from trade — $b-s \geq 0$ for sure. Nothing changes as long as we have the weaker assumption that $\max\{b-p, p-s\} \geq 0$ for sure (and continue to require $w(b,s) \geq 0$ everywhere, which may require the observer's criterion to be something other than gains from trade).

However, if it is possible that both $b-p$ and $p-s$ are negative, then we can no longer ensure $\mu_B+\mu_S \leq \mu$ in Proposition 3.1; instead we only have $\mu_B \leq \mu$ and $\mu_S \leq \mu$ separately. This is because condition (c) in the proof may be satisfied with strict inequality, and unlike before, we can no longer decrease one of the λ's to make it become an equality. This again gives us a gap between Propositions 3.1 and 3.2. So again, Proposition 3.1 may give us a nontrivial robust prediction, but Proposition 3.2 no longer ensures that this prediction is optimal.

4 Examples

Here we give a couple of examples illustrating the results of Section 3, as well as discussions exploring some interpretive issues.

4.1 Computing the Maximum Robust Prediction

We start with a simple (perhaps too simple) application of our results, showing how to compute the maximum robust prediction, in an example adapted from Morris and Shin [14]. It is common knowledge that the the good is worth $2c$ more to the buyer than it is to the seller. Most likely, it is worth $p+c$ to the buyer and $p-c$ to the seller. However, there is a small probability δ that the good is a lemon, with low value to both parties, and probability δ that it is a peach, with high value to both parties. Specifically, the common

prior distribution μ is that

$$(b,s) = \begin{cases} (p-M+c, p-M-c) & \text{with probability } \delta, \\ (p+c, p-c) & \text{with probability } 1-2\delta, \\ (p+M+c, p+M-c) & \text{with probability } \delta. \end{cases}$$

(Here $M > c > 0$.) We take $w(b,s) = 1$ everywhere, so we are interested in robustly predicting the probability of trade; predicted gains from trade are just $2c$ times this probability. Note that the numerical illustration at the beginning of the introduction is an instance of this setup.

Since the criterion $w(b,s) = 1$ and the criterion $w(b,s) = 2c$ are equivalent in this example, the shortcut at the beginning of Subsection 3.2 applies: we form the buyer's negative-gain measure μ_B by carving out probability mass from μ with the lowest possible buyer-gains ratios, up until the point where the buyer's conditional expected value equals the price p; and we form the measure μ_S by carving out probability mass with the highest possible buyer-gains ratios, until the seller's expected value equals p. If these regions end up overlapping, then the best robust prediction is zero trade.

Specifically, there are two cases depending on parameters:

- If $\delta M/c \leq 1/2$, then the maximal possible total mass of μ_B is $\delta M/c$ — consisting of the δ probability of lemon realizations, together with a $\delta(M-c)/c$ probability mass of normal realizations. Likewise the maximal μ_S consists of the δ probability of peach realizations and a $\delta(M-c)/c$ probability mass of normal realizations. (Again, these are really suprema, not maxima, but we glide over this distinction.) Therefore, by Corollary 3.3, the maximum robust prediction is $1-2\delta M/c$. That is, for any information structure, there is an equilibrium where trade occurs with probability at least $1-2\delta M/c$; and this bound is sharp, even with the restriction to symmetric information.

  To be fully explicit, we describe an information structure approaching the bound: Both parties receive the same signal, $\eta_B = \eta_S = \eta \in \{O, B, S\}$. The joint distribution of values and signals is as shown in Table 2. (Note that the formatting of this table is different from Table 1; here rows are values and columns are signals.) Here $\epsilon > 0$ is arbitrarily small. Thus, under the signal B — which is a noisy signal of the lemon state — trade cannot occur because the buyer's expected value is less than p. Under the peach signal S, trade cannot occur because the seller's expected value is greater than p. So trade occurs with probability at most $1-2\delta M/c+2\epsilon$.

Values	$\eta = B$	$\eta = O$	$\eta = S$
$(p - M + c, p - M - c)$	δ	0	0
$(p + c, p - c)$	$\delta\frac{M-c}{c} - \epsilon$	$1 - 2\delta\frac{M}{c} + 2\epsilon$	$\delta\frac{M-c}{c} - \epsilon$
$(p + M + c, p + M - c)$	0	0	δ

Table 2: Distribution of values and (symmetric) signals

- If $\delta M/c > 1/2$, then the best possible robust prediction is 0: the information may be structured so that no trade can occur in equilibrium.

  One possible information structure that yields no trade (not the only one) is to have a shared signal $\eta \in \{B, S\}$, jointly distributed with the values as shown in Table 3. Under signal B, the buyer's expected value is less than p; under signal S, the seller's expected value is more than p.

Values	$\eta = B$	$\eta = S$
$(p - M + c, p - M - c)$	δ	0
$(p + c, p - c)$	$\frac{1}{2} - \delta$	$\frac{1}{2} - \delta$
$(p + M + c, p + M - c)$	0	δ

Table 3: Distribution of values and (symmetric) signals

4.2 A More Complicated Illustration

The example in the previous subsection was rather minimal. Here we briefly walk through a more involved example, featuring a continuous and asymmetric distribution of values. To avoid a flood of notation, we use specific numbers.

Let μ be the uniform distribution on the pentagon

$$\mathcal{P} = \{(b, s) \in \mathbb{R}^2 \mid 3 \leq b \leq 8; 1 \leq s \leq 6; b \geq s\},$$

and let $p = 5$. This is shown in Figure 2(a), where the pentagon is shaded.

We first consider the gains-from-trade criterion, $w(b, s) = b - s$. In this case, again following the method described in Subsection 3.2, the worst-case μ_B consists of all the mass from μ lying in the region $(b-p)/(b-s) < \alpha_B$, for some threshold α_B. The portion of the pentagon $\mathcal{P}$ satisfying this inequality is shown horizontally hatched in Figure 2(b).

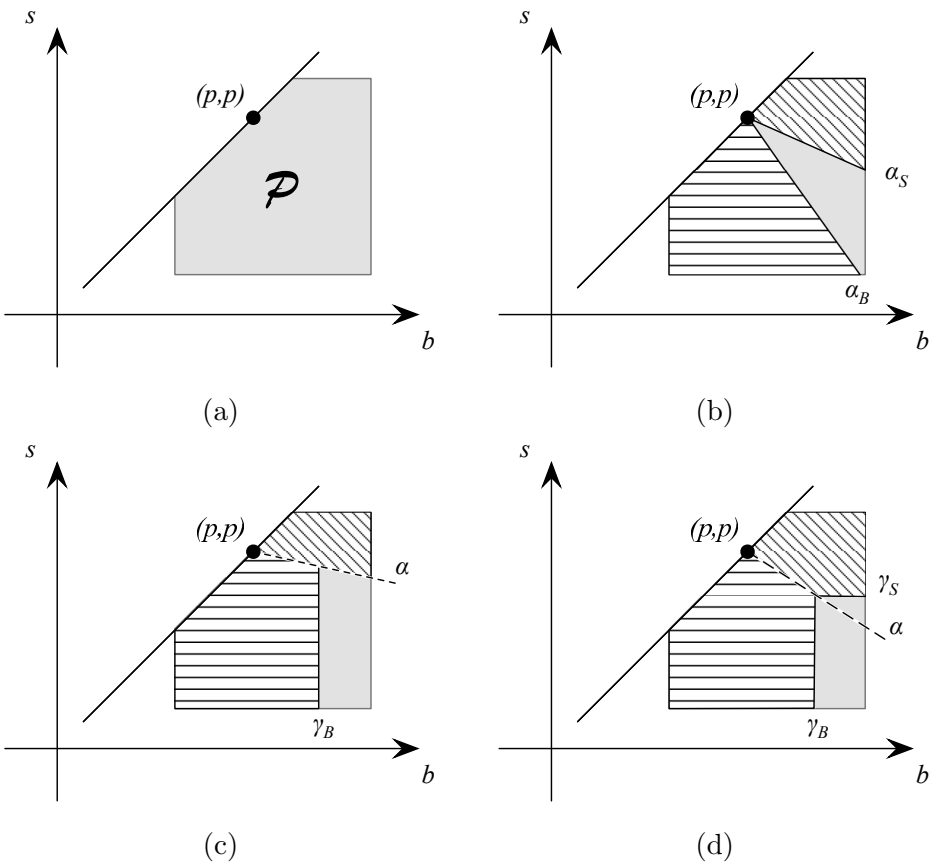

Figure 2: Worst-case computation, in an example. (a) The observer's μ. (b) Worst case for the gains-from-trade criterion. (c) Computing $Y(\alpha)$ for the probability-of-trade criterion. (d) Worst case for the probability-of-trade criterion.

Thus, μ_B consists of mass with the same density as μ, uniformly distributed on this hatched region. (Unlike in the previous example, we need not worry about what happens to mass lying exactly on the boundary line $(b-p)/(b-s) = \alpha_B$, since this line has measure zero.) The relevant value of α_B is determined by the condition that the integral of $b-p$ over this hatched region should equal zero. Likewise, the worst-case μ_S consists of all the mass from μ lying in the region $(b-p)/(b-s) > \alpha_S$, shown diagonally hatched in Figure 2(b). The value of α_S is determined by the condition that the integral of $p-s$ over this region is zero.

Thus, in the worst case (ignoring the ϵ adjustments needed to make the inequalities strict), both players receive an (identical) signal telling them whether (b, s) is in the horizontally-hatched, diagonally-hatched, or the gray region in Figure 2(b). Only in the gray region does trade occur.

To identify the value of the threshold α_B, compute that the corresponding line intersects the lower edge of $\mathcal{P}$ at point $\left(\frac{5-\alpha_B}{1-\alpha_B}, 1\right)$. (This is assuming that the line does indeed intersect the lower edge of $\mathcal{P}$, as drawn, rather than the right edge; we later check that this is indeed the case.) Thus the range of possible values of b in the horizontally-hatched region is $\left[3, \frac{5-\alpha_B}{1-\alpha_B}\right]$, and for each such b, the range of corresponding values of s is $\left[1, \min\{b, \frac{b(\alpha_B-1)+5}{\alpha_B}\}\right]$. Hence the condition pinning down α_B is

$$\int_3^{\frac{5-\alpha_B}{1-\alpha_B}} \int_1^{\min\{b, \frac{b(\alpha_B-1)+5}{\alpha_B}\}} (b-5)\, ds\, db = 0.$$

The left side is a rational function of $\alpha_B \in [0,1]$, and the resulting polynomial equation has a unique solution in this range, which we compute to be $\alpha_B = \sqrt{2}-1$. From this we check that $\frac{5-\alpha_B}{1-\alpha_B} = 5+2\sqrt{2} < 8$, so that the line does indeed intersect the lower edge of $\mathcal{P}$ as shown.

By a similar argument, the value of α_S is given by the condition

$$\int_{8-\frac{3}{\alpha_S}}^{6} \int_{\max\{s, \frac{5-\alpha_S s}{1-\alpha_S}\}}^{8} (5-s)\, db\, ds = 0$$

which gives the solution $\alpha_S = \frac{27}{20} - \frac{3}{20}\sqrt{21}$. We check that the intersection of the line $(b-p)/(b-s) = \alpha_S$ with the vertical $b=8$ is $\left(8, 5-\frac{1}{3}\sqrt{21}\right)$, whose vertical coordinate is > 1, so that the line does indeed intersect the right edge of $\mathcal{P}$ as shown.

This identifies the worst case for gains from trade. The maximum robust prediction is then given by integrating $b-s$, multiplied by the density of μ, over the gray region. This integral comes to approximately 0.772. Thus, we can robustly predict that for any information structure, there is an equilibrium that realizes expected gains from trade of at least 0.772, or 29.1% of the first-best gains from trade; and this bound cannot be improved.

We now take the same μ and p, but consider the probability-of-trade criterion, $w(b,s) = 1$. In this case, we proceed as described at the end of Subsection 3.2. For each $\alpha \in [0,1]$, we separate μ into the mass below the line $(b-p)/(b-s) = \alpha$ and the mass above it. (Our earlier description of the process also involves a choice of β, which tells us how to divide up the mass exactly on the line; again, this is irrelevant in the present example since this mass is zero.) Within the portion below the line, we form μ_B by taking all the mass with sufficiently low values of $(b-p)/w = b-p$ — that is, with $b-p < \gamma_B$ for some γ_B. This is shown by the horizontally-hatched region in Figure 2(c). The value of

γ_B is determined by the constraint that the integral of $b - p$ over this region should equal 0. Similarly, μ_S consists of all the mass above the α line with sufficiently low values of $p - s$, as shown by the diagonally-hatched region; the integral of $p - s$ over this region should equal 0. If there is no choice of the threshold γ_B (or γ_S) that makes the integral equal zero, then we come as close as possible by taking all of the available mass to form μ_B (respectively, μ_S); this is shown for μ_S in the figure. Once μ_B and μ_S are constructed, $Y(\alpha)$ consists of the sum of the integrals of w with respect to these two measures — that is, the total area of the two hatched regions (times a constant, the density of μ). Finally, whichever α maximizes $Y(\alpha)$ gives the worst case.

The thresholds γ_B and γ_S are functions of α that cannot be conveniently written in closed form; they are solutions to cubic polynomials whose coefficients depend on α. However, we can compute them, and maximize $Y(\alpha)$, numerically. The resulting α is approximately 0.614, and the corresponding worst-case μ_B, μ_S consist of the mass shown in the horizontally- and diagonally-hatched regions of Figure 2(d).[2] The maximum robust prediction is then given by integrating w, multiplied by the density of μ, over the remaining, gray region. This gives us 0.165: thus, for any information structure, there exists an equilibrium where trade occurs with probability at least 0.165.

4.3 No-Trade Equilibria

As mentioned in the introduction, it is hard to interpret our results as giving a positive prediction about how much trade will happen, because there is always at least one other equilibrium, in which neither agent ever accepts trade. One might try to get rid of such bad equilibria using a standard refinement, such as elimination of weakly dominated strategies, or more generally trembling-hand perfection. Unfortunately, this refinement does not help. We now demonstrate this with an illustration, building on the simple example from Subsection 4.1, in which there can be trembling-hand perfect equilibria with no trade, even though the good equilibrium outcome involves trade most of the time.

Let μ be as given in Subsection 4.1, for some parameter values with $\delta M/c$ small, so that for any information structure there is an equilibrium with a high probability of trade. Now consider the following information structure. The signal sets are $\mathcal{I}_B = \mathcal{I}_S = \{L, N, P\}$. The letters stand for "lemon, normal, peach," and the first and last signals are perfectly

[2]Note that, as depicted, the lines given by γ_B and γ_S happen to intersect on the line $(b-p)/(b-s) = \alpha$. This is not coincidence: in general, with continuous distributions and the probability-of-trade criterion, one can show that when this concurrency occurs, the first-order condition for maximizing $Y(\alpha)$ is satisfied.

informative while the middle signal is imperfectly informative. Specifically, conditional on the values (b, s), both players' signals are independently drawn from the same distribution, which is given by Table 4.

Values	$Pr(L)$	$Pr(N)$	$Pr(P)$
$(p - M + c, p - M - c)$	1/2	1/2	0
$(p + c, p - c)$	0	1	0
$(p + M + c, p + M - c)$	0	1/2	1/2

Table 4: Distribution of each player's signal, conditional on values

There are some signal realizations for which the players have (weakly) dominant actions: If the buyer receives L, he knows the values are $(p - M + c, p - M - c)$ for sure, so he does not accept trade, in any trembling-hand perfect equilibrium. Similarly, if the seller receives L, he does accept. If the buyer receives P, he accepts; if the seller receives P, he does not accept.

Let (σ_B, σ_S) be the following strategy profile: the buyer accepts only when his signal is P, and the seller accepts only when his signal is L. To check that this is a trembling-hand perfect equilibrium, it suffices to check that each player is playing a strict best reply to the other's strategy when his own signal is N, since it follows that each player's strategy is a best reply to any sufficiently small tremble. Consider the buyer's strategy when his signal is N. From his point of view, any of the three value pairs — and any of the seller's signals — can occur with positive probability. But if he accepts trade, the trade will only occur if the seller's signal is L, in which case trade is definitely bad for him. So the buyer strictly loses by agreeing to trade on signal N. Similarly for the seller.

In this equilibrium, trade only occurs if the buyer receives signal P and the seller receives L; but this can never happen.

4.4 Alternative Mechanisms

As mentioned in the introduction, our results are dependent on the particular posted-price mechanism we have assumed. In particular, a peculiar feature of this setup is that the worst-case information structure is symmetric; this would not hold in general for other trading mechanisms.

For example, consider instead a double auction mechanism: the buyer names a price p_B, and the seller names a price p_S; if $p_B < p_S$ then no trade takes place, and if $p_B \geq p_S$

then trade happens at price $(p_B + p_S)/2$. For any μ, and any *symmetric* information structure, there is an equilibrium in which the parties always trade: For each realization of the signal η, pick any price $p(\eta)$ lying in between the buyer's and seller's expected values conditional on η; then it is an equilibrium for both parties, after observing η, to name the price $p(\eta)$. This mechanism realizes all gains from trade.

In view of this observation, one might ask: is it possible that, no matter what the information structure is, the buyer and seller can always come up with *some* suitable mechanism — and an equilibrium of it — that realizes all (or at least most) of the gains from trade? After all, we have assumed it is common knowledge that $b \geq s$, so that, say, the classic impossibility result of Myerson and Satterthwaite [15] does not apply.

However, there exist variants of this result in the literature that do apply, showing that for certain value distributions and (asymmetric) information structures, no (budget-balanced, individually rational) mechanism guarantees efficient trade. For example, Fieseler, Kittsteiner, and Moldovanu [7] consider an information structure in which each party receives a one-dimensional signal, and each party's value is a function of both signals; they give a necessary and sufficient condition for the nonexistence of a mechanism guaranteeing efficient trade, generalizing Myerson-Satterthwaite. Moreover, Samuelson [18] considers a generalization of Akerlof's [1] lemons model, in which one party is perfectly informed about both parties' values and the other is completely uninformed; Samuelson shows that for some parameterizations, no mechanism can achieve *any* trade in equilibrium, even though there is common knowledge of gains from trade.

In view of all this discussion, it is natural to ask what happens when we allow the parties to choose the best mechanism, instead of simply assuming a posted-price mechanism. In general, what would the maximum robust prediction then look like? And what would the best mechanism (and the corresponding worst-case information structure) look like? These questions, however, seem substantially more difficult than the analysis we have given here for the posted-price mechanism.

5 Closing Discussion

5.1 Interpretation

Now it's time to fulfill the promise in the introduction, to discuss possible economic interpretations of our main results. We stress, however, that the question of economic interpretation is basically separate from the methodological purpose of the paper, which

has already been discussed.

A key assumption is that the observer knows the distribution of buyer's and seller's values for the good, but does not know the information structure and does not directly observe the trading outcomes. Thus, it makes sense to think of the observer not as an econometrician who has past trading data, but perhaps as a planner trying to orchestrate future trades, with limited foresight of the relevant environment.

For example, one might imagine that a buyer and seller are considering contracting on a specialized widget, which they may or may not actually wish to trade in the future, but which requires some capital investment today in order to be able to trade later. Our model applies if they can currently foresee the physical circumstances that affect each party's value for the widget, but cannot anticipate what each party will know when the time comes to trade. A lower bound for the attainable gains from trade can potentially provide an immediate guarantee that the investment is worthwhile.[3]

A related application might be to a regulator designing a financial market, in which agents might be able to trade some security whose value depends on future events. If the regulator can anticipate how the events will affect the security's value but not the details of what information the traders will have, a lower-bound result can potentially provide assurance that there will still be enough trade in the market to warrant the fixed costs of opening the market.

A different perspective is to fit our work in with the literature on design of information structures [10, 16, 12], taking the worst-case information structure literally as a description of how an adversary might best prevent two parties from trading. This might describe, for example, a firm that tries to prevent its rival from successfully trading with a supplier by putting in place a signalling system that reveals to them information relevant to their trade.

Finally, one more economic interpretation of our results is as a counterpoint to the literature on how trade breaks down in lemons markets. As discussed in the introduction, recent work such as Morris and Shin [14] points to contagion in adverse selection, emphasizing the role played by higher-order beliefs. Although higher-order beliefs can indeed be important for particular information structures, our results suggest that they are not

[3]In our model, there is common knowledge of gains from trade. In this case, our analysis seems unnecessary: the parties simply could agree up front to trade with probability 1, at a price that splits the ex-ante gains from trade. However, the model fits the following variant: The buyer will find out tomorrow whether he wants the widget (in which case gains from trade are positive) or doesn't want the widget (gains are negative), and there may be additional information as well, of unknown structure. Ex ante, the buyer is unlikely to want the widget, so that simply contracting to sell is inefficient. The model then describes what happens conditional on the buyer wanting the widget.

needed to tell a story about breakdown of trade. That is, given the known distribution over values, the probability of trade breakdown that can be explained using higher-order beliefs is no worse than may occur with very simple and indeed symmetric information structures. This finding builds in a natural way on the earlier work of Kessler [11] and Levin [13] showing that the extent of trade breakdown in lemons markets is generally non-monotone in the amount of information asymmetry. However, an important caveat to this interpretation is that it depends on our assumption of a posted-price mechanism for trade. As discussed in Subsection 4.4 above, a different mechanism could lead to different predictions.

More generally, a few words should be said about our assumption of a posted-price mechanism and its importance. As we have seen, this assumption is limiting, both in terms of the sharpness of our characterization — we show how to find the highest possible robust prediction for gains from trade, but this is no longer sharp if the agents are allowed to choose a different mechanism — and our observations about the nature of the worst-case information structure. One modest defense is that we simply follow the literature — e.g. [1, 8, 14] — in adopting this simple trading mechanism, in order to better focus attention on the question of information structure. Another point is that our main result is a lower bound on the attainable gains from trade; it would continue to hold *a fortiori* if the parties were also allowed to use other mechanisms, instead of being restricted to a posted price. In particular, imagine a double auction mechanism as in Subsection 4.4. Any equilibrium of our posted-price mechanism can be translated into an equilibrium of the double-auction mechanism: reinterpret "accepting price p" as a bid of p in the double auction, and reinterpret "rejecting price p" as making an unacceptable bid in the double auction (a bid outside the support of values, which the other party would never want to accept). This produces the same outcome as the original equilibrium of the posted-price mechanism. Thus our sharp lower bound on attainable trade in the posted-price mechanism is also a valid lower bound for the double auction mechanism, which has the advantage of being "parameter-free," unlike the posted price mechanism which has the pesky p exogenously given.

5.2 Future directions

We wrap up by quickly surveying directions for future exploration. On the technical side, the sharp characterization of robust predictions of trade calls out to be extended to allow for $b < s$, and more generally to allow negative values of the observer's criterion

w. The other major direction, already pointed out in Subsection 4.4, would be to ask about the best equilibrium outcome of the best trading mechanism, rather than a specific posted-price mechanism. Alternative extensions could keep the restriction to a very simple trading mechanism, but consider trade in multiple units of a good, or multiple goods.

From the methodological point of view, the role of this paper is to ask what predictions can be made about economic interactions without knowing the details of the information structure. We have begun to think about this question by focusing on one of the simplest possible economic transactions — exchange of a single indivisible good, between one buyer and one seller. We should note, however, that our model already has other interpretations beyond the exchange setting: more generally it describes any situation where two agents can each approve or veto some proposal, which passes only if both approve, and where there is common knowledge that at least one agent benefits from the proposal. (In our exchange interpretation, the agents' values for the proposal are $b - p$ and $p - s$.) In any case, it will be natural for future work to take the same question of informationally robust prediction to other workhorse economic models — production, moral hazard, coordination games, public good provision — and see where there are interesting answers.

A Omitted Details

Proof of Lemma 3.4: In keeping with the main text, write $\mu' = \mu_B + \mu_S$. As α ranges over $[0, 1]$, the event $E^\alpha_<$ is increasing in α. (This depends on the fact that $b - s \geq 0$ everywhere.) Moreover, any pair (b, s) contained in one $E^\alpha_<$ but not another satisfies $b - p \geq 0$, since pairs with $b - p < 0$ are in every $E^\alpha_<$. Therefore $\int_{E^\alpha_<}(b - p)\, d\mu'$ is weakly increasing in α. Also, it is negative for small enough $\alpha > 0$, and is left-continuous. Let $\overline{\alpha} \in (0, \infty]$ be the supremum of values for which $\int_{E^\alpha_<}(b - p)\, d\mu' < 0$.

Similarly, $\int_{E^\alpha_>}(p - s)\, d\mu'$ is weakly decreasing in α, negative at $\alpha = 1$, and right-continuous. Let $\underline{\alpha}$ be the infimum of values for which $\int_{E^\alpha_>}(p - s)\, d\mu' < 0$.

We show that $\overline{\alpha} \geq \underline{\alpha}$. Suppose not. Then $E^{\overline{\alpha}}_\leq = (E^{\overline{\alpha}}_< \cup E^{\overline{\alpha}}_=)$ is disjoint from $E^{\underline{\alpha}}_\geq = (E^{\underline{\alpha}}_= \cup E^{\underline{\alpha}}_>)$. We must have $\int_{E^{\overline{\alpha}}_\leq}(b - p)\, d\mu' \geq 0$, otherwise the maximality of $\overline{\alpha}$ would be violated. Similarly, $\int_{E^{\underline{\alpha}}_\geq}(p - s)\, d\mu' \geq 0$.

Define two new signed measures by

$$\widetilde{\mu}_B(E) = \mu_B(E) - \mu'(E \cap E^{\overline{\alpha}}_\leq), \qquad \widetilde{\mu}_S(E) = \mu_S(E) - \mu'(E \cap E^{\underline{\alpha}}_\geq).$$

Note that $\widetilde{\mu}_B$ is nonpositive on $E^{\overline{\alpha}}_{<}$ and nonnegative on $E^{\overline{\alpha}}_{>}$, hence

$$\int ((b-p) - \overline{\alpha}(b-s))\, d\widetilde{\mu}_B \geq 0.$$

Similarly

$$\int ((p-s) - (1-\underline{\alpha})(b-s))\, d\widetilde{\mu}_S \geq 0.$$

Then we have

$$0 > \int_{\mathbb{R}^2} (b-p)\, d\mu_B - \int_{E^{\overline{\alpha}}_{\leq}} (b-p)\, d\mu' = \int_{\mathbb{R}^2} (b-p)\, d\widetilde{\mu}_B \geq \overline{\alpha} \int_{\mathbb{R}^2} (b-s)\, d\widetilde{\mu}_B,$$

$$0 > \int_{\mathbb{R}^2} (p-s)\, d\mu_S - \int_{E^{\underline{\alpha}}_{\geq}} (p-s)\, d\mu' = \int_{\mathbb{R}^2} (p-s)\, d\widetilde{\mu}_S \geq (1-\underline{\alpha}) \int_{\mathbb{R}^2} (b-s)\, d\widetilde{\mu}_S.$$

So $\int_{\mathbb{R}^2} (b-s)\, d\widetilde{\mu}_B < 0$ and $\int_{\mathbb{R}^2} (b-s)\, d\widetilde{\mu}_S < 0$, and therefore

$$\int_{\mathbb{R}^2} (b-s)\, d(\widetilde{\mu}_B + \widetilde{\mu}_S) < 0.$$

However, $\widetilde{\mu}_B + \widetilde{\mu}_S$ is a nonnegative measure since

$$(\widetilde{\mu}_B + \widetilde{\mu}_S)(E) = \mu'(E) - \mu'(E \cap E^{\overline{\alpha}}_{\leq}) - \mu'(E \cap E^{\underline{\alpha}}_{\geq}) = \mu'(E \setminus (E^{\overline{\alpha}}_{\leq} \cup E^{\underline{\alpha}}_{\geq})) \geq 0$$

for any event E. Since $b - s \geq 0$ μ'-almost everywhere, we have a contradiction.

So indeed we have $\overline{\alpha} \geq \underline{\alpha}$. If $\overline{\alpha} > \underline{\alpha}$, we can take α to be any number in between and β to be arbitrary. Then define

$$\widehat{\mu}_B(E) = \mu'(E \cap E^{\alpha}_{<}) + \beta \mu'(E \cap E^{\alpha}_{=}), \tag{A.1}$$

$$\widehat{\mu}_S(E) = \mu'(E \cap E^{\alpha}_{>}) + (1-\beta) \mu'(E \cap E^{\alpha}_{=}). \tag{A.2}$$

Now

$$E^{\alpha}_{<} \subseteq E^{\alpha}_{<} \cup E^{\alpha}_{=} \subseteq E^{\alpha'}_{<}$$

for any $\alpha' \in (\alpha, \overline{\alpha})$ readily implies

$$\int_{\mathbb{R}^2} (b-p)\, d\widehat{\mu}_B = \int_{E^{\alpha}_{<}} (b-p)\, d\mu' + \beta \int_{E^{\alpha}_{=}} (b-p)\, d\mu' < 0,$$

and by a similar argument

$$\int_{\mathbb{R}^2} (p-s)\, d\widehat{\mu}_S < 0.$$

Thus, $(\widehat{\mu}_B, \widehat{\mu}_S)$ is a negative-gains pair. Since $\mu' \leq \mu$, we can see from the definitions (A.1–A.2) that this pair is (α, β)-separated; and evidently $\widehat{\mu}_B + \widehat{\mu}_S = \mu' = \mu_B + \mu_S$. So we are finished in this case.

We are left with the case $\overline{\alpha} = \underline{\alpha}$. In this case, we fix $\alpha = \overline{\alpha} = \underline{\alpha}$ and repeat the argument with β.

Since $b - p, p - s \geq 0$ everywhere on $E^{\alpha}_{=}$, the expression

$$\int_{E^{\alpha}_{<}} (b-p)\, d\mu' + \beta \int_{E^{\alpha}_{=}} (b-p)\, d\mu' \tag{A.3}$$

is weakly increasing in $\beta \in [0,1]$. Let $\overline{\beta}$ be the supremum of such values for which it is < 0. (If it is ≥ 0 already at $\beta = 0$ then take $\overline{\beta} = 0$.) Note that by continuity in β, (A.3) is in fact ≥ 0 at $\overline{\beta}$, except in the corner case where $\overline{\beta} = 1$ and $\alpha = 1$. But in this corner case, the lemma is easily proven. Indeed, we can then take $(\alpha, \beta) = (1,1)$, and define $\widehat{\mu}_B$ and $\widehat{\mu}_S$ by (A.1–A.2), and the conclusion of the lemma holds: $(b-p)\, d\widehat{\mu}_B < 0$ by assumption, $\int (p-s)\, d\widehat{\mu}_S$ must be < 0 because $\widehat{\mu}_S$ only places weight on $E^1_{>}$, where $p - s < 0$ for sure, and the rest follows as before. Thus, we may assume that the expression (A.3) is ≥ 0.

Similarly, the expression $\int_{E^{\alpha}_{>}} (p-s)\, d\mu' + (1-\beta) \int_{E^{\alpha}_{=}} (p-s)\, d\mu'$ is decreasing in β; let $\underline{\beta}$ be the infimum of values for which it is < 0, or $\underline{\beta} = 1$ if no such values exist. The expression is ≥ 0 there except if $\underline{\beta} = 0$ and $\alpha = 0$, and again this corner case can be disposed of separately.

Now we show that $\overline{\beta} > \underline{\beta}$. Suppose not. Then take any β with $\overline{\beta} \leq \beta \leq \underline{\beta}$. Define

$$\begin{aligned}
\widetilde{\mu}_B(E) &= \mu_B(E) - \mu'(E \cap E^{\alpha}_{<}) - \beta \mu'(E \cap E^{\alpha}_{=}), \\
\widetilde{\mu}_S(E) &= \mu_S(E) - \mu'(E \cap E^{\alpha}_{>}) - (1-\beta) \mu'(E \cap E^{\alpha}_{=}).
\end{aligned}$$

As before, $\widetilde{\mu}_B$ is nonpositive on $E^{\alpha}_{<}$ and nonnegative on $E^{\alpha}_{>}$, hence

$$\int ((b-p) - \alpha(b-s))\, d\widetilde{\mu}_B \geq 0,$$

and similarly

$$\int ((p-s) - (1-\alpha)(b-s))\, d\widetilde{\mu}_S \geq 0.$$

Now

$$0 > \int_{\mathbb{R}^2} (b-p)\, d\mu_B - \left(\int_{E^\alpha_<} (b-p)\, d\mu' + \beta \int_{E^\alpha_=} (b-p)\, d\mu' \right)$$

(since the first integral is negative by assumption, and the expression in parentheses is just (A.3) at β, which is ≥ 0 because we have assumed we are not in the corner case)

$$= \int_{\mathbb{R}^2} (b-p)\, d\widetilde{\mu}_B \geq \alpha \int_{\mathbb{R}^2} (b-s)\, d\widetilde{\mu}_B.$$

Thus, $\int_{\mathbb{R}^2} (b-s)\, d\widetilde{\mu}_B < 0$. By a similar argument, $\int_{\mathbb{R}^2} (b-s)\, d\widetilde{\mu}_S < 0$. Adding, $\int_{\mathbb{R}^2} (b-s)\, d(\widetilde{\mu}_B + \widetilde{\mu}_S) < 0$. But $\widetilde{\mu}_B + \widetilde{\mu}_S = \mu' - \mu' = 0$ identically — a contradiction.

Thus, $\overline{\beta} > \underline{\beta}$. So we can choose $\beta \in (\underline{\beta}, \overline{\beta})$. Now let $(\widehat{\mu}_B, \widehat{\mu}_S)$ be defined by (A.1–A.2). It is immediate that $\int_{\mathbb{R}^2} (b-p)\, d\widehat{\mu}_B$, which is just (A.3), is < 0, and similarly $\int_{\mathbb{R}^2} (p-s)\, d\widehat{\mu}_S < 0$. Thus the new pair is a negative-gains pair, and the rest is checked as before. □

Proof of Lemma 3.5: We only prove the formula for Y_B; the Y_S case is analogous.

First suppose $\gamma^*_B = \infty$. Then $\int_{\mathbb{R}^2} w(b,s)\, d\overline{\mu}_B$ is clearly an upper bound for $Y(\alpha,\beta)$. From the definition of γ^*_B, we have $\int_{F^\infty_<} (b-p)\, d\overline{\mu}_B \leq 0$, where $F^\infty_<$ is the event ($w(b,s) > 0$ or $b-p<0$). If the inequality is strict, we can take $\mu_B = \overline{\mu}_B|_{F^\infty_<}$ (that is, the measure defined by $\mu_B(E) = \overline{\mu}_B(E \cap F^\infty_<)$ for any E). Otherwise, since there is a positive probability of $b-p<0$ under μ (by assumption) and so also under $\overline{\mu}_B|_{F^\infty_<}$ (note that this equals μ for events where $b-p<0$), then there is also a positive probability of $b-p>0$ under $\overline{\mu}_B|_{F^\infty_<}$. So we can form μ_B from $\overline{\mu}_B|_{F^\infty_<}$ by removing an arbitrarily small probability mass on such an event. In either case, we obtain μ_B with $\int_{\mathbb{R}^2} (b-p)\, d\mu_B < 0$ strictly, and $\int_{\mathbb{R}^2} w(b,s)\, d\mu_B$ arbitrarily close to $\int_{\mathbb{R}^2} w(b,s)\, d\overline{\mu}_B$.

Now suppose γ^*_B is finite. Define the measure $\widehat{\mu}_B$ by

$$\widehat{\mu}_B(E) = \overline{\mu}_B(E \cap F^{\gamma^*_B}_<) + \delta^*_B \overline{\mu}_B(E \cap F^{\gamma^*_B}_=).$$

So the expression given as the value of $Y(\alpha,\beta)$ in the lemma statement is simply $\int w(b,s)\, d\widehat{\mu}_B$. Also, we know that $\int (b-p)\, d\widehat{\mu}_B = 0$.

We first show that this value is an upper bound on $Y(\alpha,\beta)$. Otherwise, let μ_B be a measure with higher value of $\int w(b,s)\, d\mu_B$, still satisfying $\int (b-p)\, d\mu_B < 0$ and $\mu_B \leq \overline{\mu}_B$. Define a signed measure by $\widetilde{\mu}_B = \mu_B - \widehat{\mu}_B$. Then $\widetilde{\mu}_B$ is nonpositive on $F^{\gamma^*_B}_<$, and

nonnegative on $F_{>}^{\gamma_B^*}$ (which we define in the obvious way). Therefore,

$$\int_{\mathbb{R}^2} \left((b-p) - \gamma_B^* w(b,s)\right) \, d\widetilde{\mu}_B \geq 0.$$

This implies

$$\int_{\mathbb{R}^2} (b-p) \, d\mu_B - \int_{\mathbb{R}^2} (b-p) \, d\widehat{\mu}_B \geq \gamma_B^* \left(\int_{\mathbb{R}^2} w(b,s) \, d\mu_B - \int_{\mathbb{R}^2} w(b,s) \, d\widehat{\mu}_B \right).$$

But here the left side is negative, while the right side is positive — a contradiction.

So $\int_{\mathbb{R}^2} w(b,s) \, d\widehat{\mu}$ is indeed an upper bound on $Y(\alpha,\beta)$. For the reverse direction, note that, as in the $\gamma_B^* = \infty$ case, the measure $\widehat{\mu}_B$ places some positive probability on the event $b - p < 0$ (which is contained in $F_{<}^{\gamma_B^*}$), and so it must also place positive probability on $b - p > 0$. By removing an arbitrarily small amount of probability mass with $b - p > 0$, we get a new measure μ_B such that $\int_{\mathbb{R}^2} (b-p) \, d\mu_B < 0$ and $\mu_B \leq \overline{\mu}_B$, and $\int_{\mathbb{R}^2} w(b,s) \, d\mu_B$ is arbitrarily close to $\int_{\mathbb{R}^2} w(b,s) \, d\widehat{\mu}_B$. □

References

[1] George A. Akerlof (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* 84 (3), 488–500.

[2] George-Marios Angeletos and Jennifer La'O (2013), "Sentiments," *Econometrica* 81 (2), 739–779.

[3] Dirk Bergemann, Benjamin Brooks, and Stephen Morris (2013), "Extremal Information Structures in the First Price Auction," Princeton Economic Theory Center Working Paper #055–2013.

[4] Dirk Bergemann, Benjamin Brooks, and Stephen Morris (2013), "The Limits of Price Discrimination," *American Economic Review* 105 (3), 921–957.

[5] Dirk Bergemann and Stephen Morris (2014), "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," Princeton Economic Theory Center Working Paper #054–2013.

[6] Dirk Bergemann and Stephen Morris (2013), "Robust Predictions in Games with Incomplete Information," *Econometrica* 81 (4), 1251–1308.

[7] Karsten Fieseler, Thomas Kittsteiner, and Benny Moldovanu (2003), "Partnerships, Lemons, and Efficient Trade," *Journal of Economic Theory* 113 (2), 223–234.

[8] Oliver Hart and John Moore (1988), "Incomplete Contracts and Renegotiation," *Econometrica* 56 (4), 755–785.

[9] Atsushi Kajii and Stephen Morris (1997), "The Robustness of Equilibria to Incomplete Information," *Econometrica* 65 (6), 1283–1309.

[10] Emir Kamenica and Matthew Gentkow (2011), "Bayesian Persuasion," *American Economic Review* 101 (6), 2590–2615.

[11] Anke S. Kessler (2001), "Revisiting the Lemons Market," *International Economic Review* 42 (1), 25–41.

[12] Anton Kolotilin (2014), "Experimental Design to Persuade," *Games and Economic Behavior* 90, 215–226.

[13] Jonathan Levin (2001), "Information and the Market for Lemons," *RAND Journal of Economics* 32 (4), 657–666.

[14] Stephen Morris and Hyun Song Shin (2012), "Contagious Adverse Selection," *American Economic Journal: Macroeconomics* 4 (1), 1–21.

[15] Roger B. Myerson and Mark A. Satterthwaite (1983), "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* 29 (2), 265–281.

[16] Luis Rayo and Ilya Segal (2010), "Optimal Information Disclosure," *Journal of Political Economy* 118 (5), 949–987.

[17] Ariel Rubinstein (1989), "The Electronic Mail Game: Strategic Behavior under 'Almost Common Knowledge,'" *American Economic Review* 79 (3), 385–391.

[18] William Samuelson (1984), "Bargaining under Asymmetric Information," *Econometrica* 52 (4), 995–1005.