# Notes on the Economics of Information

April 17, 2024

*These notes are still a work in progress. If you spot any typos, or if anything is unclear, please let me know. They are light on examples, as they are written to supplement, not replace, lectures.*

## 1 Introduction

It goes without saying that information plays a fundamental role in almost all economic interactions. Most notably, Hayek proffered the view that the primary role and virtue of markets was their ability to aggregate diffuse information. And yet, the role of information was largely ignored in the initial decades of modern economics. It plays almost no role many of the canonical models you've seen in first year microeconomics. And perhaps, after completing Advanced Microeconomics 1 and 2, one may wonder, is this simplification really consequential? Surely, these models are smooth enough that the introduction of small informational asymmetries won't change much? Wouldn't most of this information be revealed through economic activity anyway? And if not, could we not simply treat information as an additional economic good, and incorporate it into our models as one would incorporate grain, or cloth, or widgets? Surely the logic which has led us to view markets as a powerful mechanism to allocate scarce resources will still hold with incomplete information.

This was largely the attitude of economists for much of the 20th century. As every overview of this topic points out, Stigler – while demonstrating a setting where information plays a non-trivial role – compared the information to subzero weather, its important but when properly handled its effects can be kept within "comfortable bounds".[1] But, as we'll see in this course, what Stigler uncovers is closer to the norm

[1]This comparison is drawn in a paper where Stigler shows that if consumers have to search for opportunities to buy a good, trade can occur above market price. If Stigler went a bit further with the analysis, but allowing the firms in his model to set prices, he would have uncovered the "Diamond Paradox." Once small search costs are introduced, even with many firms, the equilibrium price is the monopoly price.

than the exception. When met with incomplete information, very few of our insights from the complete information world survive unscathed.

In this course, I'll cover some of the canonical models of the economics of information. I focus on two fundamental frictions:

- **Adverse Selection:** "Hidden type", differences in information at the start of the interaction (e.g. I know I have a heart condition when I go to buy insurance)
- **Moral Hazard:** "Hidden Action", differences in information that arise as part of the interaction (e.g. I know how hard I worked, my boss does not).

Game theory provides us with a powerful tool to analyze the role of information. While some of these insights were initially developed in frameworks closer to general equilibrium, game theory will almost always enable a crisper analysis of the phenomena. I won't have time to provide an overview of all the important models and results in this literature. But, we'll have time to see some cool ones.

## 2 Screening

### 2.1 An Example

To begin our analysis of the role of information, let's start with a problem that is phenomenally boring without it. A monopolist can produce a single good of quality $q$ for cost $\frac{1}{2}q^2$ that it then sells to a single consumer. The consumer has utility $\theta q - t$ for buying a good of quality $q$ from the monopolist at price $t$, and the monopolist's profits from selling are $t - \frac{1}{2}q^2$. The consumer gets utility 0 from not buying. So formally the monopolist solves

$$\max_{t,q} t - \frac{1}{2}q^2 \text{ s.t. } \theta q - t \geq 0$$

Having taken math camp, our monopolist knows exactly what to do here. They should produce a good of quality $q = \theta$ and sell it for price $t = \theta^2$. We can make the following observations about this interaction:

- Trade is efficient.
- The monopolist extracts all surplus.

Now suppose that $\theta = 1$ or 3, where $Pr(\theta = 3) = \alpha$. If the monopolist could observe $\theta$, nothing really changes. I set up a shelf of $\theta = 1$ goods and a shelf of $\theta = 3$ goods at my store, and won't let 3s buy from the 1 shelf or 1s buy from the 3 shelf. But, it seems

clear that in most situations, I probably won't be able to tell 1s from 3s. What would happen if I offered the same menu?

Each of our consumers now has two options, they can either:

- Pay 1 for a good of quality 1. They then receive utility $\theta - 1$.
- Pay 9 for a good of quality 3. They then receive utility $3\theta - 9$.

Our 3s aren't idiots, they like the quality 3 good more, but it's so expensive. Why don't they just buy they quality 1 good, and get positive utility? The monopolist would then make an expected profit of 1/2.

Our monopolist thinks this through and says, "well shit, I can do better than this." Some possibilities:

- I could give the high type a discount. If I charge 8 for a good of quality 3, then the 3s have no reason not to buy that one. I make a profit of $\frac{7}{2}\alpha + (1-\alpha) > 1$.
- I could stop selling the low quality good. Then I make $\frac{9}{2}\alpha$.
- I could raise the quality of the high quality good. If I sold a good of $q = 10/3$, then the 3s would have no reason not to buy that. Then I make $\frac{31}{9}\alpha + (1-\alpha)$. Note that this is always worse than offering a discount.
- Something else.

Now our monopolists problem is much more complicated. The monopolist is now selling by offering a menu. One item on the menu they intend to be purchased by 1s, the other by 3s. Let's denote these as $(q_3, t_3)$ and $(q_1, t_1)$. Due to *adverse selection*, they have to make sure 1s don't want to pretend to be 3s and 3s don't want to pretend to be 1s. This leads us to the two new constraints, which capture the role of asymmetric information in this problem. These are the **Incentive Compatibility Constraints**.

The monopolist solves

$$\begin{aligned}
&\max_{t,q} \alpha(t_3 - \frac{1}{2}q_3^2) + (1-\alpha)(t_1 - \frac{1}{2}q_1^2) \\
&\text{s.t. } q_1 - t_1 \geq q_3 - t_3 && (IC_{1,3}) \\
&3q_3 - t_3 \geq 3q_1 - t_1 && (IC_{3,1}) \\
&q_1 - t_1 \geq 0 && (IR_1) \\
&3q_3 - t_3 \geq 0 && (IR_3)
\end{aligned}$$

We saw in class that $IR_3$ is redundant. It's a pretty good guess that $IC_{3,1}$ is the constraint that gives us trouble. Solving this like we did in class, we get the menu

$(q_1, t_1) = (\max\{0, (1-3\alpha)/(1-\alpha)\}, \max\{0, (1-3\alpha)/(1-\alpha)\})$ and $(q_3, t_3) = (3, 9-2q_1)$. We now see that

- The allocation is no longer efficient.
- Consumer surplus goes up. The high type is now collecting rents.
- The low type quality good is being produced at a lower than efficient quality.

Intuitively, our monopolist really wants to identify the high type here, as they have a much higher willingness to pay per unit of quality. But the high type isn't just going to reveal themselves to the monopolist. They need to be compensated for the information they give up. But, at the same time, the value of this information for them caused by their ability to mimic the low type. So, the monopolist has two levers, they can make mimicking the low type less appealing by the low type's option worse, or they could offer a discount to the high type. We see that if there is a high chance they are selling to the low type, they do both, while if they are pretty confident they are selling to a high type, they just make the low type good worthless, which makes it costless to get the high type's information.

How does this generalize to more then two types? We can already see that adding additional types makes the problem much more complicated. The number of IC constraints scales badly, and the whole problem looks like a big mess. So, let's do what we usually do when the finite world seems to complicated. Let's think about the infinite world.

Now suppose $\theta \in [0, 1]$ is drawn from a continuous distribution with density $f(\theta)$ with support $[0, 1]$. In principle, the monopolist could offer any number of selling schemes, formally they could commit to a arbitrary space $M$ of "messages" the agent could send, and a mapping between $M$ and quantities and prices. It's pretty obvious that we don't need to worry about this. The following is pretty obvious

**Theorem 1** (Revelation Principal)**.** *For any mechanism $\Gamma = (M, (q, t))$ and optimal strategy $\sigma^*_\Gamma$ there is an incentive compatible direct mechanism $\hat{\Gamma} = (\Theta, (\hat{q}, \hat{t}))$ with the same outcome as mechanism $\Gamma$.*

Then the monopolist solves

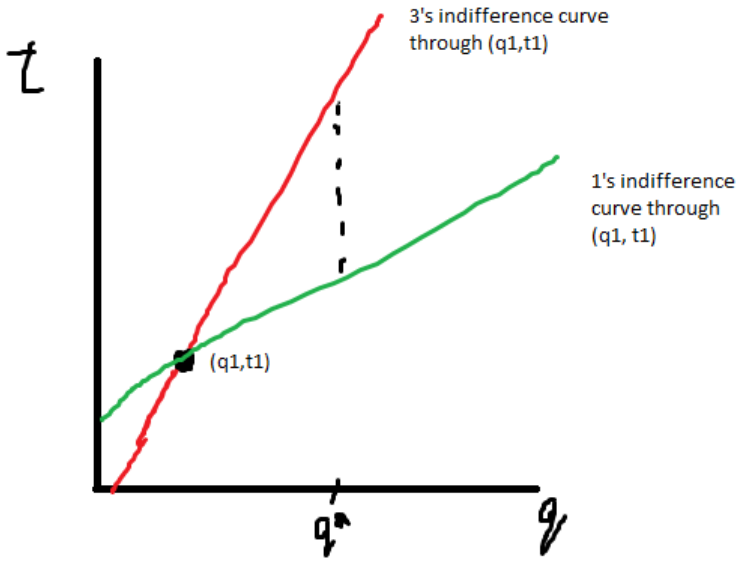
$$\begin{aligned} \max \int_{\underline{\theta}}^{\bar{\theta}} \left[ t(\theta) - \frac{1}{2} q(\theta)^2 \right] f(\theta) d\theta & \\ \text{s.t. } \theta q(\theta) - t(\theta) \geq \theta q(\theta') - t(\theta') & \qquad (IC_{\theta,\theta'}) \\ \theta q(\theta) - t(\theta) \geq 0 & \qquad (IR_\theta). \end{aligned}$$

Shit.

## 2.2 Dealing with IC constraints

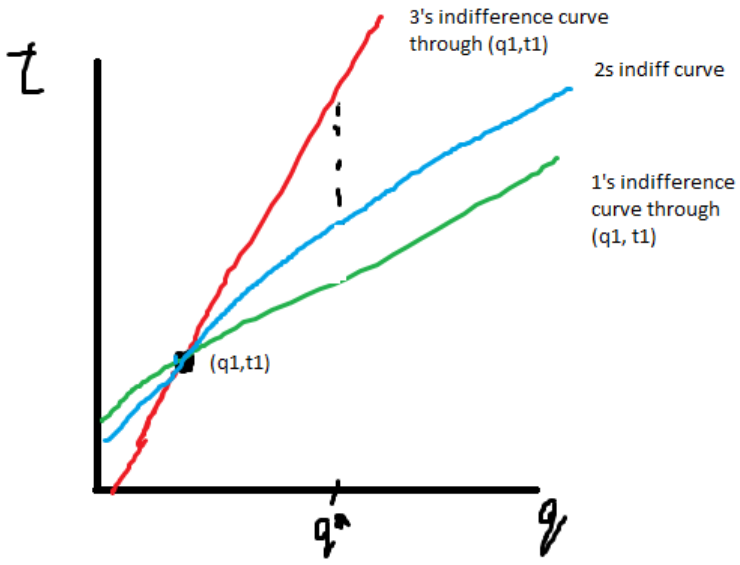
It seems like we have a big problem here. We are maximizing over the space of functions. We now have an unbelievable number of constraints. This seems terrible for us.

It turns out, this is actually pretty easy.[2] To see why, consider our two type problem. If we decide on an allocation for the 1 type, it immediately limited the space of $(t, q)$s we could offer the type 3 agent. Suppose the monopolists him to buy a specific $q^*$. Then they have to charge a price below the red indifference curve and above the green one.




What if I add an addition type, $\theta = 2$, and want them to buy, for simplicity $(q1, t1)$. Then I can charge even fewer prices.

[2] I promise, I'm telling the truth.




As you can sort of intuit, as I fill in types between 1 and 3, the number of options the monopolist has for what price to charge for the quality they want to sell is going to shrink. Hopefully it is clear that if I add all types between 1 and 3 and try to design a menu so they all buy $(q_1, t_1)$ then the only price I could charge for $q^*$ is the price that puts type 3 on their indifference curve through $(q_1, t_1)$. Maybe it's a bit less clear, but this example at least suggests that IC constraints for types closer to 3 (in both directions) "matter more" for determine what prices can be charged for $q^*$.[3]

Now let's go back to our constraints. The IC constraint is

$$\theta q(\theta) - t(\theta) \geq \theta q(\theta') - t(\theta')$$

for all $\theta, \theta'$. These seem hard to deal with, but also impose a lot of structure on the monopolist's choices. Consider the pair $IC_{\theta,\theta'}$ and $IC_{\theta',\theta}$.

$$\theta q(\theta) - t(\theta) \geq \theta q(\theta') - t(\theta')$$
$$\theta' q(\theta') - t(\theta') \geq \theta' q(\theta) - t(\theta)$$

Adding these up and rearranging terms, we can see that these imply that

$$(\theta - \theta')(q(\theta) - q(\theta')) \geq 0.$$

So in any incentive compatible menu, **quality must be increasing in type.**

[3]Try drawing this diagram for the case when $q_2$ allocation is not $q_1$. Then the indifference curves through what 2s $(q_2, t_2)$ determine what transfers can be offered for $q^*$.

Now what about our observation that closer types seem to "matter more" for incentive constraints. Let

$$V(\theta) := \theta q(\theta) - t(\theta)$$

be the value function for the agent. This tells us the utility a type $\theta$ agent gets from revealing their type in the mechanism. We can reformulate incentive constraints as

$$V(\theta) \geq V(\theta') - \theta' q(\theta') + \theta q(\theta'),$$

which in turn, when we combine the pair $IC_{\theta,\theta'}$ and $IC_{\theta',\theta}$ gives us

$$(\theta - \theta')q(\theta) \geq V(\theta) - V(\theta') \geq (\theta - \theta')q(\theta').$$

Dividing through by $\theta - \theta'$ and taking the limit as $\theta$ approaches $\theta'$,[4] we get

$$V'(\theta) = q(\theta).$$

And so,

$$V(\theta) = \int_0^\theta q(s)\, ds + V(0).$$

We'll call this the *envelope condition*.[5]

So, we know that

$$V(\theta) = \int_0^\theta q(s)\, ds + V(0)$$

and

$$V(\theta) = \theta q(\theta) - t(\theta)$$

by the revelation principle. So for any quality schedule the monopolist wants to implement, the only transfer schemes that can possibly implement it are given by

$$t(\theta) = \theta q(\theta) - \int_0^\theta q(s)\, ds - V(0).$$

[4] You may be a bit worried that I'm fudging some math here. Since $q(\theta)$ is increasing and bounded, it is continuous almost everywhere, and the IC constraints imply $V(\theta)$ is Lipschitz, so is almost everywhere differentiable. So we're fine (modulo an a.e. that is irrelevant for the analysis that follows). If this doesn't make sense, don't worry about it. This note is for annoying nerds like me.

[5] To see why this really is a consequence of the envelope theorem we're used to, note that for incentive compatible $q(\theta), t(\theta)$

$$V(\theta) = \arg\max_{x \in [0,1]} \theta q(x) - t(x)$$

so $V'(\theta) = q(x^*(\theta))$ where $x^*(\theta) = \theta$ be the revelation principle. We've been a bit more careful than we were in math camp in establishing this, and are in some sense fixing a policy function and trying to find a transfer scheme that implements it. So we don't have to worry about all those weird details that caused us problems in math camp (e.g. differentiability of the policy function).

It turns out, that not only are these properties consequences of incentive compatibility, they are in fact necessary and sufficient for incentive compatibility. Formally

**Theorem 2.** *A direct mechanism $(q, t)$ is incentive compatible if and only if $q(\theta)$ is increasing and $t(\theta) = \theta q(\theta) - \int_0^\theta q(s)\,ds + K$ for some constant $K$.*

*Proof.* These notes have already shown that incentive compatibility implies these two properties. We'd like to show the reverse. Consider

$$\theta q(\theta) - t(\theta) - \theta q(\theta') + t(\theta'),$$

we need to show this is positive for any pair of $\theta, \theta'$. Let's substitute in our expression for $t(\theta)$ and do some algebra:

$$\begin{aligned}
&\theta q(\theta) - t(\theta) - \theta q(\theta') + t(\theta') \\
&= \int_0^\theta q(s)\,ds - \theta q(\theta') + \theta' q(\theta') - \int_0^{\theta'} q(s)\,ds \\
&= \int_{\theta'}^\theta q(s) - q(\theta')\,ds \\
&\geq 0.
\end{aligned}$$

□

This result – that the $q(\theta)$ the monopolist wants to implement pins down transfers (up to a constant) – is often called *revenue equivalence*. Any mechanism that implements allocation $q(\theta)$ and provides the same payoff to the lowest type must give the seller the same revenue. As we move to richer settings, this is going to prove to be a powerful implication of incentive compatibility.

### 2.3 The Optimal Mechanism

Now we understand a bit more about the structure of the set of feasible mechanisms. We can reformulate the principal's problem as

$$\begin{aligned}
&\max_{t,q} \int_0^1 \left[ t(\theta) - \frac{1}{2} q(\theta)^2 \right] f(\theta)\,d\theta \\
&\text{s.t. } t(\theta) = \theta q(\theta) - \int_0^\theta q(s)\,ds + V(0) \text{ for all } \theta \\
&V(\theta) \geq 0 \text{ for all } \theta \\
&q(\theta) \text{ is increasing}
\end{aligned}$$

We haven't talked much about the IR constraints. This is mostly because they are really boring. The logic we used in the example tells us that IC + lowest type IR implies IR for any higher type. And, given our envelope condition, it is clearly optimal for the designer to set $V(0)$ as low as possible, i.e. $V(0) = 0$.[6] Plugging our formula for $t(\theta)$ into the objective, we have

$$\max_q \int_0^1 \left[\theta q(\theta) - \int_0^\theta q(s)\,ds - \frac{1}{2}q(\theta)^2\right] f(\theta)\,d\theta$$
$$\text{s.t. } q(\theta) \text{ is increasing}$$

We have two problems here.

1. The objective depends on the entire function $q(s)$, as does the stuff inside the integral. It's not clear we can do something like integrate pointwise.

2. We have a really weird constraint.

We'll approach 1 by doing some dumb integration tricks. We'll approach 2 by ignoring it.

To deal with 1, note that everything in the integrand depends only $q(\theta)$ except for that one term from the transfer, which is

$$\int_0^1 \int_0^\theta q(s) f(\theta)\,ds\,d\theta.$$

If I just change the order of integration with Fubini's theorem, this becomes

$$\int_0^1 \int_s^1 q(s) f(\theta)\,d\theta\,ds = \int_0^1 q(s)(1 - F(s))\,ds.$$

So, finally we have

$$\max_q \int_0^1 \left[\theta q(\theta) - q(\theta)\frac{1 - F(\theta)}{f(\theta)} - \frac{1}{2}q(\theta)^2\right] f(\theta)\,d\theta.$$

Since each $q(\theta)$ only appears "once" in the integral, we can maximize this pointwise. This gives first order condition

$$q(\theta) = \theta - \frac{1 - F(\theta)}{f(\theta)}.$$

as the optimal $q(\theta)$. Let's go ahead and assume that at the start of this section we assume that this object was increasing, so I don't have to worry about monotonicity.

[6] This is mostly but not always the case in this course. Be careful!

We call $\theta - \frac{1-F(\theta)}{f(\theta)}$ the *virtual type*. Due to asymmetric information, a type $\theta$ agent receives the quality that was optimal for the agent with their virtual type, reducing the overall surplus created through trade. We make the following observations:

- The final allocation is inefficient. Agents receive lower quality goods than they would in the efficient allocation.
- Consumers are "better off", they now collect information rents.
- Higher type consumers collect higher rents and thus a higher share of the surplus (since $V(\theta) = \int_0^\theta q(s)\,ds$).

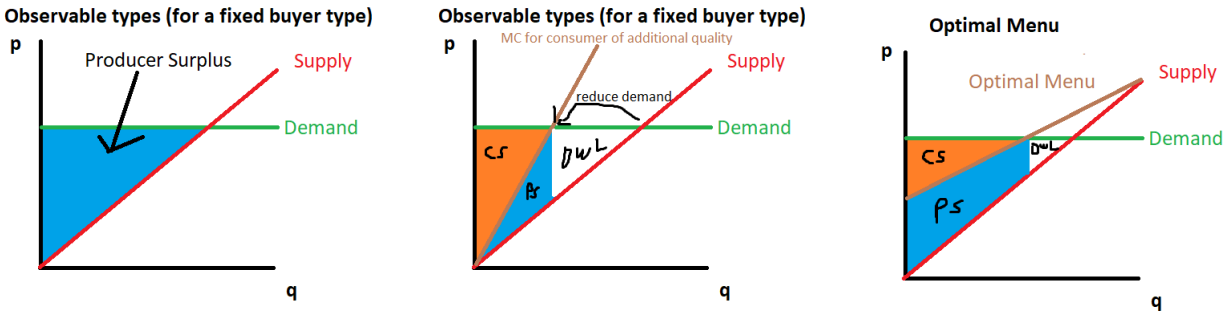



FIGURE 1. What supply and demand graphs look like for a fixed type.The monopolist would like sell $q(\theta) = \theta$ to the consumer, for a per-unit-of-quality price $\theta$. But, if the type was not observed directly, the consumer could reduce their demand. If they actually offer this price schedule, the consumer can then act as a monopsonist and extract rents. Since the firm is able to "design" the marginal cost curve the consumer faces, they can claw back more market power. Finally, observe that each quantity on the brown curve is demanded by some buyer type (and that CS must be equal to the info rent) gives us the envelope condition.

We can study other things as well. For instance, if $\theta$ is uniform on $[0, 1]$, then the virtual type is $2\theta - 1$ and it's straightforward to see that the marginal price of an additional unit of quality is increasing.

## 2.4 General Analysis

There's nothing here particularly special about the form of the consumers utility (beyond quasilinearity) or the firm's problem here. It's pretty obvious that if we don't change the agent's preference, we can redo this for any continuously differentiable, convex cost function (w/ some conditions on the 1st derivative) and get

$$\theta - \frac{1 - F(\theta)}{f(\theta)} = c'(q(\theta))$$

as the optimal quality menu.

The agent side stuff is really where all the cool stuff was in the first place. It turns out, we can do the same analysis for a pretty large class of preferences. Let's give agent's utility $u(q;\theta)$ and let $\theta$ be drawn from some continuous distribution with support $[\underline{\theta},\bar{\theta}]$ and assume $q$ is chosen from some set $Q$. Let's try to find a transfer scheme that implements an arbitrary incentive compatible $q(\theta)$. IC constraints can now be written as

$$V(\theta) - V(\theta') + (u(q(\theta');\theta') - u(q(\theta');\theta)) \geq 0$$

and thus imply that

$$u(q(\theta);\theta) - u(q(\theta);\theta') \geq V(\theta) - V(\theta') \geq u(q(\theta');\theta) - u(q(\theta');\theta').$$

Now if we want to do the same trick, we need a few assumptions.

- $u(q,\theta)$ sufficiently differentiable.
- Increasing differences, $u_{q,\theta} > 0$. This is essential for the sufficiency of the envelope conditions.
- And a more technical assumption, we need the derivatives to be uniformly bounded.

Then, by essentially the logic from before

$$V(\theta) - V(\underline{\theta}) = \int_{\underline{\theta}}^{\theta} \frac{\partial u}{\partial \theta}(q(s);s)\,ds.$$

Stated more abstractly (without e.g. quasilinearity)

**Theorem 3.** *Assume that $X$ is compact, and $\Theta = [\underline{\theta},\bar{\theta}]$ and $g : X \times \Theta \to \mathbb{R}$ is differentiable with uniformly bounded derivatives. Then if $q(\theta)$ solves*

$$V(\theta) := \max_{x \in X} g(x;\theta)$$

*then*

$$V'(\theta) = g_\theta(q(\theta),\theta) \text{ (a.e)}$$

*and furthermore*

$$V(\theta) = V(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} g_\theta(q(s),s)\,ds$$

So for large variety of problems, the IC constraints pin down the value function, and thus transfers in quasilinear settings. Cool![7] Applied to our setting

[7] I hope.

**Theorem 4** (Revenue Equivalence)**.** *Fix a function $q : \Theta \to Q$. Suppose that $\Theta = [\underline{\theta}, \overline{\theta}]$, $u : Q \times \Theta \to \mathbb{R}$ is differentiable with uniformly bounded derivatives and $Q$ compact. Any incentive compatible mechanism that implements $q(\theta)$ gives agents payoff*

$$V(\theta) = V(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} u_\theta(q(s), s)\, ds$$

*and transfers must satisfy*

$$t(\theta) = u(q(\theta); \theta) - V(\underline{\theta}) - \int_{\underline{\theta}}^{\theta} u_\theta(q(s), s)\, ds$$

So, the principal gets same payoff in any mechanism that implements $q(\theta)$, up to lowest type payoff.

It turns out that these conditions are necessary and sufficient.

**Theorem 5.** *Suppose the conditions for the previous theorem hold hold and $\frac{\partial^2 u}{\partial q \partial \theta} > 0$. Then $(q(\theta), t(\theta))$ is IC iff $q(\theta)$ is non-decreasing and*

$$t(\theta) = u(q(\theta); \theta) - V(\underline{\theta}) - \int_{\underline{\theta}}^{\theta} u_\theta(q(s), s)\, ds.$$

Let's quickly apply these to some different settings with price discrimination. Suppose that types are uniformly distributed on $[0, 1]$.

- Mussa Rosen: $c(q) = cq^2$, $u(q, \theta) = \theta q$.
  - We generally interpret $q$ as the quality of good. This gives us that the firm sells quality at a premium.
- Maskin Riley: $c(q) = cq$, $u(q, \theta) = \theta v(q)$, $v$ concave.
  - Can solve this using our tools

    $$v'(q(\theta)) = \frac{c}{2\theta - 1}$$

  - Moreover, if we express things as a function of $q$, $\theta v'(q) = t'(q)$

    $$t''(q) = \frac{1}{2} v''(q) \leq 0$$

  - We generally now interpret $q$ as quantity, this gives us a model that rationalizes quantity discounts.

It's going to be particularly useful for what comes next to think about the problem of a seller selling a single indivisible good that costs them $c$ to produce. Now interpret $q$ as

the probability of sale, and we get

$$\max \int_{\underline{\theta}}^{\overline{\theta}} [t(\theta) - cq(\theta)] f(\theta) d\theta$$

with IC constraints:

$$\theta q(\theta) - t(\theta) \geq \theta q(\theta') - t(\theta')$$

and IR:

$$\theta q(\theta) - t(\theta) \geq 0$$

We can solve this essentially exactly like our motivating example. The "relaxed problem" is

$$\max \int_{\underline{\theta}}^{\overline{\theta}} \left( \theta - \frac{1 - F(\theta)}{f(\theta)} - c \right) q(\theta) f(\theta) d\theta$$

Which tells us that the optimal mechanism is to simply post a price, and sell to everyone whose virtual type is above cost. There is no benefit from offering multiple alternatives, and the monopolist will never sell e.g. lotteries.

### 2.5 Wrap up

What have we learned about the role of information from this exercise?

**Information Rents.** Agents must be compensated for the information they reveal, and the level of this compensation depends on the what other types are receiving. We see this in the envelope expression $V(\theta) = \int_0^\theta q(s)\, ds + V(0)$. We call these *information rents*. While we studied this in a monopoly setting, these rents are present no matter what objective the principal is trying to optimize. We'll see in other settings that even with a more "benevolent" principal, information rents still create economically important distortions relative to the full information case.

**Allocative Efficiency.** Without adverse selection, the monopolists problem was very simple. They maximized utility net costs and had consumers pay them all the utility they'd gained. Even with adverse selection, our monopolist could still offer a menu that lead to an allocation that would maximize utility net costs. But, it follows from the envelope theorem in our example that they would have to charge type $\theta$ a price of $\theta^2/2 + V(0)$ for an object of quality $q$, and thus would be unable to extract any surplus. We see that the monopolist, in the optimal mechanism, creates dead-weight loss in order to relax incentive constraints and increase the share of the surplus they extract.

**Distributional Consequences.** Consumer surplus is not evenly divided amongst types. For any quality schedule we can implement, higher types receive a higher share of the surplus. In the optimal mechanism, the lowest type receives the same utility that

they received in the complete information version of this problem, but receives a good of lower quality than they'd receive in the full information setting. In contrast, the highest type receives exactly the same quality good that they would have received without adverse selection, but at a greatly reduced price. This makes sense, the lowest types information is of no value to the monopolist, so they have no leverage to extract rents. In contrast, the highest type is the most valuable type for the monopolist to identify, as they provide the most utility per unit of quality.

**Assumptions.** We can see from the envelope theorem, that the intuition that adverse selection creates rents and reduces the monopolists ability to extract surplus is relatively general. The rest of the analysis hinges on a number of relatively important, and sometimes subtle assumptions.

1. Utility is quasilinear.
    - Gives us a clean expression for prices in terms of the quality allocation.
    - Imposes e.g. risk neutrality.
2. Types are single dimensional, and drawn from an interval in $\mathbb{R}$.
    - Gives us a lot of structure on the set of types.
    - Rules out e.g. information that is gradually revealed over time, different willingnesses to pay for different attributes, etc. Without this, harder to see what the natural analogue for the next two assumptions is.
3. Utility satisfies increasing difference/single crossing.
    - Allows us to only deal with local incentive constraints. Without this, we'd still have the envelope condition, but the solution to the relaxed problem could violate some of the IC constraints.
    - Higher types increase marginal utility for quality. Imposes a clear meaning on types (and how they are ordered). This is going to be a problem, for instance, if we try to find an isomorphism between multidimensional types, as described above, into a single dimensional type.
4. The type distribution satisfies a monotone hazard rate condition.
    - This allowed us to ignore the monotonicity constraint.
    - Without this, we can't maximize the principal's objective pointwise. This is technically annoying to resolve, but can be resolved using techniques from calculus of variations. In mechanism design, we refer to the technique to solve

this as ironing, the monopolist essentially smooths out the non-monotonic part of the solution. It's a huge pain in the ass, and we follow convention by assuming it away (until a referee asks about it).

A final assumption that may give you pause is the assumption that the monopolist knows the type distribution. In this setting, this it to some extent innocuous, one could view the type distribution as simply capturing the monopolists prior beliefs over the distributions of willingness-to-pays in the population. The agent's incentives within the mechanism are the same, regardless of what distribution types are drawn from. Similarly, while we've structured the optimal mechanism as having the agent report their type and the principal then gives them a product and transfer that correspond to their type, it's easy to see here how this would be implemented without relying on the agent reporting an abstract object like type. The principal could simply offer a menu of price-quality pairs, and allow the agent to select their post preferred one. This is called the *taxation principle*, and will hold in essentially any setting where we have the revelation principle.

A natural question arises from this analysis. In our model, we have one buyer and one seller, which gives the buyer quite a bit of market power when "selling" their information to the seller. What if we allowed the seller to choose between a number of potential buyers? Would competition amongst buyers mitigate the impact of adverse selection? What would the optimal selling mechanism look like in that case?

A second natural question that arises is, what if the private information is on the seller's side? This is a good question. If the seller can still commit to a mechanism, the problem becomes enormously complicated, as they sellers choice of mechanism now potentially conveys information. A more straightforward, and perhaps more important question is what happens in the game without commitment. It turns out this game is interesting, even if we assume there are always possible gains from trade (the buyer always values the item more than the seller's opportunity cost of selling it). This is the well known "market for lemons," you'll do a version of this on a problem set.

## 3 Bayesian Mechanism Design and the Optimal Auction

A seller wants to sell a single indivisible good that costs $c$ to produce. There are $N$ buyers, $i \in \{1, 2, \dots N\}$, each with a willingness to pay $\theta_i$ drawn from continuous distribution $F_i$ with pdf $f_i$, with support $[\underline{\theta}_i, \overline{\theta}_i]$. These types are independent, the buyer who receives the good and pays transfer $t$ receives utility $\theta_i - t$. The seller commits to a mechanism, which buyers then play. How should the seller sell this item?

This looks a lot like the problem we solved in the previous section. In that problem, we only had one buyer, and the principal solved

$$\max E(t(\theta) - cq(\theta))$$

$$\theta q(\theta) - t(\theta) \geq \theta q(\theta') - t(\theta') \text{ for all } \theta, \theta'$$
$$\theta q(\theta) - t(\theta) \geq 0$$

It seems like it would be straightforward to set this up for $N$ buyers. Go ahead and try it now, I'll wait.

Not as straightforward. Incentive constraints in the previous section were capturing the optimality of the agent's decision. It's no longer obvious how to do that. It's not even obvious what exactly the principal is now designing, as they are now designing a game for the agents to play, not an individual decision problem. We're going to think about designing games in a couple different settings, so it will be helpful to take a step back and think about these sorts of problems in the abstract.

### 3.1 Incentive Compatibility

We have $N$ agents, each with type $\theta_i \in \Theta_i$. Let $\theta \in \Theta$ denote the vector of types, and assume it is drawn from joint distribution $F(\theta)$. The set of possible allocations is $x \in X$. Agent $i$ knows their own type but not others' types. If the final allocation is $x$ and the transfer is $t$, an agent receives utility $u_i(x, \theta) - t$. We say this is a setting with independent, private values if types are independent and if $u_i(x, \theta)$ only depends on $\theta_i$. We have interdependent values if the entire vector of types enters preferences. We'll spend very little time in this course thinking about settings with correlation or interdependence.

A mechanism is now a set of messages $M_i$ for each player, with $M = M_1 \times M_2 \times \dots M_N$. an allocation rule $x : M \to X$ and a transfer scheme $t : M \to \mathbb{R}^N$. $M_i$ is essentially the strategy space of the game being played, and the designer can tailor payoffs using both the allocation rule and transfer scheme.

We'd like to simplify the principal's problem using the revelation principal. But, it's harder to see what this even means. The obvious candidate is for it to be a Nash equilibrium. But, unlike in Micro 3, here we are designing the mechanism. So, we could ask for a stronger solution concept that puts us on firmer epistemic ground. We don't have to worry about non-existence of dominant strategies, because we can construct the game to have a dominant strategy.[8] As before, we can specify incentive constraints which capture the requirements of different solution concepts. Let $(m_i^*(\theta_i))_{i=1}^N$ be a strategy profile

- Dominant strategy: $\forall \theta_i, \theta_{-i}, m_{-i}$

$$\begin{aligned} &u_i(q(m_i^*(\theta_i), m_{-i}), \theta) - t_i(m_i^*(\theta_i), m_{-i}) \\ &\geq u_i(q(m, m_{-i}), \theta) - t_i(m, m_{-i}) \end{aligned}$$

[8] We could also worry about uniqueness vs. non-uniqueness of the equilibrium outcome. This is called full implementation (as opposed to partial implementation). There is a literature on this, it's pretty goofy. Don't tell Hannu about this footnote.

- Ex-post equilibrium: $\forall \theta_i, \theta_{-i}, m$

$$\begin{aligned} & u_i(q(m^*(\theta)), \theta) - t_i(m^*(\theta)) \\ & \geq u_i(q(m, m^*_{-i}(\theta)), \theta) - t_i(m, m^*_{-i}(\theta)) \end{aligned}$$

- Bayes-Nash equilibrium: $\forall \theta_i$

$$\begin{aligned} & \int_{\Theta_{-i}} u_i(q(m^*(\theta)), \theta) - t_i(m^*(\theta))\, dF_{-i}(\theta_{-i}|\theta_i) \\ & \geq \int_{\Theta_{-i}} u_i(q(m, m^*_{-i}(\theta)), \theta) - t_i(m, m^*_{-i}(\theta))\, dF_{-i}(\theta_{-i}|\theta_i) \end{aligned}$$

We can apply the revelation principle here as well. Given an indirect mechanism, our designer can implement the same outcome as a direct mechanism by ask for types and committing to "play" the indirect mechanism for them. Stated somewhat formally

**Theorem 6.** *For any $\cdot$IC mechanism $(M, t, q)$ there exists an $\cdot$IC direct mechanism that implements $(t, q)$.*

For a direct mechanism, incentive constraints become

- Dominant strategy incentive compatibility: $\forall \theta_i, \theta'_i, \theta_{-i}, \theta'_{-i}$

$$\begin{aligned} & u_i(q(\theta_i, \theta'_{-i}), \theta) - t_i(\theta_i, \theta'_{-i}) \\ & \geq u_i(q(\theta'_i, \theta'_{-i}), \theta) - t_i(\theta'_i, \theta'_{-i}) \end{aligned}$$

- Ex-post incentive compatibility: $\forall \theta_i, \theta_{-i}, \theta'_i$

$$\begin{aligned} & u_i(q(\theta), \theta) - t_i(\theta)) \\ & \geq u_i(q(\theta'_i, \theta_{-i}), \theta) - t_i(\theta'_i, \theta_{-i}) \end{aligned}$$

- Bayes-Nash incentive compatibility: $\forall \theta_i, \theta'_i$

$$\begin{aligned} & \int_{\Theta_{-i}} u_i(q(\theta)), \theta) - t_i(\theta)\, dF_{-i}(\theta_{-i}|\theta_i) \\ & \geq \int_{\Theta_{-i}} u_i(q(\theta'_i, \theta_{-i}), \theta) - t_i(\theta'_i, \theta_{-i})\, dF_{-i}(\theta_{-i}|\theta_i) \end{aligned}$$

It's pretty clear that DIC implies EIC imples BIC. The difference between Bayesian and ex-post/dominant strategy incentive compatibility is pretty clear. The distinction between ex-post and dominant strategy may be a bit harder to see. Dominant strategy asks for optimality against any possible strategy, while ex-post only asks for the strategy

to remain optimal after types are revealed. In private values settings, these are essentially the same thing, as other players' types only enter payoffs through reports.[9] This is not the case in other settings.

We have both positive and normative goals with the different design exercises in this course, and the exact concept that's appropriate is going to change somewhat based on our goals. When we're exploring the limits of what we can do, impossibility results are more striking under a weaker solution concept, like BIC. At the same time, EIC and DIC both have clear conceptual advantages over the bayesian concepts, so understanding the limits of what can be done in dominant strategies is also important. When thinking about designing the optimal mechanism, there are similar trade-offs between different approaches. Understanding what can be done with dominant strategies is important, as is understanding how much we lose from what could be done if we only required BIC.

A similar issue lies in the timing of the IR constraint. We can ask for ex-post, interim or ex-ante IR.[10] The context of the problem will similarly provide guidance here for which is appropriate.

### 3.2 Back to the Optimal Auction

Let's go back to our sellers problem. Suppose the seller would like to design the optimal BIC mechanism to sell their good. Then their problem is (omitting some obvious constraints on $q_i$'s)

$$\begin{aligned}
&\max \int_\Theta \sum_{i=1}^N t_i(\theta) - c\sum_{i=1}^N q_i(\theta) dF(\theta) \\
&\text{s.t.} \int_{\Theta_{-i}} \theta_i q_i(\theta) - t_i(\theta)\, dF_{-i}(\theta_{-i}) \geq 0 \\
&\int_{\Theta_{-i}} \theta_i q_i(\theta) - t_i(\theta)\, dF_{-i}(\theta_{-i}) \geq \int_{\Theta_{-i}} \theta_i q_i(\theta', \theta_{-i}) - t_i(\theta', \theta_{-i})\, dF_{-i}(\theta_{-i}).
\end{aligned}$$

This actually looks a lot like our single agent problem. What if we let $Q_i(\theta_i) = E(q_i(\theta)|\theta_i)$ and $T_i(\theta_i) = E(t_i(\theta)|\theta_i)$. Then, with this new notation, the IC constraints become

$$\theta_i Q_i(\theta_i) - T_i(\theta_i) \geq \theta_i Q_i(\theta_i') - T_i(\theta_i')$$

[9] Even in a private values setting, a game with message profiles that are never played in equilibrium could be ex-post but not DIC. The designer can always eliminate these off-path messages to create an equivalent game that is dominance solvable.

[10] Respectively: I receive more than my reservation utility after any possible realization of type, I expect to receive more than my reservation utility after learning my type, I expect to receive more than my reservation utility before learning my type.

This is exactly what our IC constraints looked like in the single agent problem. So, applying the envelope argument from that problem, we know that

$$T_i(\theta_i) = \theta_i Q_i(\theta_i) - \int_{\underline{\theta}_i}^{\theta_i} Q_i(s)\, ds + V_i(\underline{\theta}_i),$$

and $Q_i(\theta_i)$ must be increasing. Just like before, incentive compatibility pins down transfers and payoffs. But, only up to the *interim* stage. This makes a lot of sense. In a Bayes-Nash equilibrium, the incentives to report truthfully are determined at the interim stage. As long as my expected payoff, conditional on my type, is the same, I don't distinguish between different possible things that can happen after all types are revealed.

It's again clear that we always want to set the lowest type's interim utility to 0. So, rewriting the objective and using the independence of types we end up with the relaxed problem

$$\begin{aligned} \max & \sum_{i=1}^{n} \int_{\Theta_i} [T_i(\theta_i) - cQ_i(\theta_i)] f_i(\theta_i) d\theta_i \\ & \text{s.t. } Q_i \text{ non-decreasing} \\ & T_i(\theta_i) = \theta_i Q_i(\theta_i) - \int_{\underline{\theta}}^{\theta_i} Q_i(s)\, ds. \end{aligned}$$

and following some algebra and a change of variables, we end with solving

$$\max \int_{\Theta} \sum_{i=1}^{n} \left[ \theta_i - \frac{1 - F_i(\theta_i)}{f_i(\theta_i)} - c \right] q_i(\theta) dF(\theta).$$

So the optimal mechanism will always sell to the bidder with highest *virtual type*, as long as that virtual type exceeds $c$.

We still have a few loose ends. First, we have the same issues with monotonicity here that we had in the single agent problem. And we'll resolve them the same way, by assumption. We also haven't finished dealing with transfers. Recall that incentive compatibility only pins down transfers at the interim level, i.e.

$$T_i(\theta_i) = \theta_i Q_i(\theta_i) - \int_{0}^{\theta_i} Q_i(s)\, ds$$

The IR constraint similarly only cares about interim transfers and the interim allocation.

We need to show that a transfer scheme that induces that interim expected transfer exists. Here this isn't too hard, but even here there are some subtleties. An appealing

transfer scheme is

$$t_i(\theta) = \theta_i q_i(\theta) - \int_0^{\theta_i} q_i(s, \theta_{-i})\, ds.$$

which, by the law of iterated expectations satisfies our interim condition. In fact, the mechanism described by this $t_i$ and $q_i$ is dominant strategy incentive compatible![11] So here there is no gap between requiring BIC and DIC (and similarly no gap between interim and ex-post IR). Note that this is not the only transfer scheme that works. This question of whether or not there actually is a mechanism becomes a bit more subtle when we, for instance, have a mixture of interim and ex-post constraints.

Let's think a bit about the structure of the optimal mechanism. First, suppose that all the $F_i$'s are the same. Then the optimal mechanism assigns the good to the agent with the highest type. Let $\theta_{(2)}$ denote the second highest type in the vector of types. The highest type's transfer becomes

$$t_i(\theta) = \theta_i - \int_0^{\theta_i} 1_{s \geq \theta_{(2)}}\, ds = \theta_{(2)}$$

if that type's virtual type is above $c$ and $\theta_{(2)}$'s virtual type is above $c$. Similar calculations reveal that the optimal auction with symmetric types is a second price auction with a reservation price $\theta^*$, where $\theta^*$ solves $\theta^* - (1 - F(\theta^*))/f(\theta^*) = c$.

This result is I think, in general, what people think of when they think of this paper. It provides a clear justification for the use of auctions as a selling mechanism and provides an optimal mechanism that is commonly used, and an explanation for why reserve prices make sense.

But, we can see by looking at the mechanism that the optimal auction looks very different once types are asymmetric. The highest value good may not be allocated to the highest value buyer. The seller has to trade-off generating value through the allocation (which they can then extract) and the information rents they have to pay. They may prefer to allocate to a lower value bidder whose type was drawn from a distribution with very little "uncertainty" (as measured by the inverse hazard rate), as opposed to a much higher value bidder if that bidder also has a much larger inverse hazard rate.

We also begin to see a shortcoming of this literature. The optimal mechanism depends crucially on the details of the type distribution. In the symmetric case, this is only insofar as the type distribution determines the optimal reserve, while in the asymmetric case these distribution also enter transfers. As this is a dominant strategy mechanism, we don't need to worry about incentives changing if the designer is wrong about the type distribution, which in general could be a problem with Bayesian mechanisms. But,

[11] We could apply essentially the same envelope argument to the DIC constraints to get this as the envelope theorem for the optimal DIC auction.

the mechanism that maximizes the designers revenue changes quite a bit depending on the distributions.

Nevertheless, the result here remains striking. We get something that looks familiar and is used in practice in well-behaved settings. The optimal mechanism in general confirms our intuition that competition can alleviate, but not completely remove information rents. The presence of asymmetric information still distorts the final allocation away from the efficient one. We see once again that information changes the sellers selling decision, and that the ability of the buyers to extract rents through their information has a real impact on how the good is allocated.

There's a huge literature that builds on this, adding twists like endogenous entry, information acquisition, information that arrives over time, etc. We've stayed in independent private values settings for this whole analysis. Common values settings are interesting, and different (as are e.g. multiunit settings), but things like information rents are still going to show up and be distortionary.

Allowing for correlation feels like a natural way to extend this model but turns out to introduce a lot of weirdness. To see some intuition for why correlation would be fundamentally different than the independent values case, think about a model where all the agents know the entire vector $\theta$. Then if there are more than two agents, I can extract all information essentially for free, just ask everyone to report their type and charge a huge negative penalty if the reports don't match. If this penalty is big enough, then truthful reporting is going to be an equilibrium, and the designer pays no information rents.[12] It turns out, while mathematically more difficult, you can essentially do the same thing if types are correlated by constructing side-bets about the other players types. These bets are constructed in such a way that their expected value is hugely negative unless the player truthfully reports their type, in which case it is 0. Cremer Mclean (1988) and Mcafee Reny (1992) have details. Here the fine details of the setting are really important.

### 3.3 Revenue Equivalence

Before we continue, it's worth highlighting again just how strong (and useful) the envelope result is. We used the following theorem to construct the optimal auction

**Theorem 7** (Revenue Equivalence)**.** *Under standard conditions (compact spaces, uniformly bounded partial derivative), any Bayesian incentive compatible mechanism that*

[12] Much of the literature on full implementation thinks about settings like this, and essentially designs a cleverer version of this mechanism that has a unique equilibrium.

*implements $q(\theta)$ gives agents payoff*

$$V_i(\theta_i) = V_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \int_{\Theta_{-i}} \frac{\partial u_i}{\partial \theta_i}(q(s, \theta_{-i}), s, \theta_{-i})\, dF_{-i}(\theta_{-i})\, ds$$

*and transfers must satisfy*

$$T_i(\theta_i) = E_{\theta_{-i}}(u_i(q(\theta); \theta)|\theta_i) - V_i(\theta_i).$$

This result, which follows from the structure of the IC constraints, tells us that there is at most one interim transfer that makes the given allocation incentive compatible. In settings that satisfy single crossing, this envelope condition, along with the restriction to monotone $q(\theta)$, is equivalent to Bayesian incentive compatibility. Analogous results hold for DIC and EIC mechanisms.

**Example 1.** (Equilibrium of auctions) *So far we've used this in design settings, where a principal is trying to choose the optimal $q(\theta)$ and corresponding transfer scheme. But, these are also powerful tools for equilibrium analysis. Consider a first price auction. Suppose there are $N$ bidders with independent private values drawn from some distribution $F$ with pdf $f$ supported on $[0, 1]$. We can write a bidder $i$ with valuation $\theta_i$'s interim payoff from bidding $b$ as*

$$Pr(Winning|b)(\theta_i - b).$$

*Let's conjecture this has a symmetric equilibrium where the bidder with the highest type wins the auction, and the lowest value bidder always gets $0$. Let $b^*$ denote the equilibrium bidding (behavior) strategy. Then we know that in this equilibrium, the bidder must prefer bidding $b^*(\theta_i)$ over bidding $b^*(\theta_i')$ for any $\theta_i' \in [0, 1]$. If we let $X_i(\theta_i)$ be the equilibrium probability of winning given the bidder has submitted $b^*(\theta_i)$ then we know in equilibrium for any $\theta_i$ and $\theta_i'$ the following constraints must hold*

$$X(\theta_i)(\theta_i - b^*(\theta_i)) \geq X(\theta_i')(\theta_i - b^*(\theta_i')).$$

*These are IC constraints, and in turn this tells us that the interim expected transfer $X(\theta_i)b^*(\theta_i)$ must satisfy the envelope condition. Moreover, $X(\theta_i) = F(\theta_i)^{N-1}$. So, by revenue equivalence*

$$F(\theta_i)^{N-1} b^*(\theta_i) = \theta_i F(\theta_i)^{N-1} - \int_0^{\theta_i} F(s)^{N-1}\, ds$$

*and this pins down the equilibrium bidding function immediately. We can integrate by*

*parts to rewrite this equation as*

$$F(\theta_i)^{N-1}b^*(\theta_i) = \int_0^{\theta_i} s(N-1)f(s)F(s)^{N-2}\,ds.$$

*Rearranging this, the expected payment is the expected value of the second largest order statistic conditional on me winning. As our theorem implies, this is also the expected payment in e.g. the symmetric monotone equilibrium of the second price auction. We can apply similar analysis to other auction formats, as well as other independent private value incomplete information games.*[13]

## 4 Efficiency

Now let's apply our tools to ask a fundamental question, does asymmetric information necessarily lead to inefficiencies? If our goal is to design a game to implement an efficient allocation, is there one?

A few answers you may have already seen:

- Walrasian Equilibrium: With complete info, no externalities, etc. not only can we implement the efficient outcome, we can do it in a decentralized way through Walrasian prices.
- Lindhal Equilibrium: With the appropriate *personalized taxes* we can similarly implement the efficient outcome in settings with externalities, public goods, etc.
- Coase's Theorem: With correctly defined "property rights" and no "transaction costs" individuals can bargain to reach the efficient outcome.[14]
- Arrow's Theorem: If we want to define a procedure that aggregates every preferences that both selects a efficient allocation and satisfies unanimity and transitivity, then the only procedures that work are dictatorial.

Arrow's theorem seems like a bad sign for us. It asks for something that satisfy what seem like "good" properties in terms of incentives, and finds that essentially nothing works. Let's depart from quasilinear settings for a bit, and think about whether this intuition about Arrow's theorem is in fact true.

### 4.1 Gibbard Satterthwaite

First, let's remind ourselves of what Arrow's theorem says. This was covered in Micro 2.

[13] We may apply it to some communication games on a problem set later on in the course.

[14] Contrary to it's name, this is not a formal theorem, and is pretty close to a tautology. That said, it has guided a great deal of jurisprudence in the US.

**Arrow's Theorem.** $X$ is a set of alternatives. There are $N$ agents, each has a preference order $R_i \subseteq X^2$ which is complete and transitive. Let $R$ denote the preference profile, the vector $(R_1, R_2, \ldots R_N)$ and $\mathcal{R}$ be the set of all preference profiles. $B \subseteq X$ is a feasible set, and $\mathcal{B}$ is the set of all possible feasible sets.

A *social choice correspondence* is a non-empty valued correspondence $C : \mathcal{B} \times \mathcal{R} \rightrightarrows X$ that satisfies $C(B, R) \subseteq B$ for all $B \in \mathcal{B}$ and $R \in \mathcal{R}$. We assume $C$ satisfies the following properties:

1. Unrestricted domain: $\mathcal{B} = 2^X \setminus \{\emptyset\}$, $\mathcal{R}$ is the set of all possible preference profiles.
2. Rationality: For any preference profile $R$, there exists an $\tilde{R} \in X^2$ such that $R$ is transitive and for all $B \in \mathcal{B}$
$$C(B, R) = \{x : x\tilde{R}y \, \forall y \in B\}$$
3. Unanimity: For any $x, y \in X$, if $xP_iy$ for all $i$ then $C(x, y) = \{x\}$, where $P_i$ is the strict part of $R_i$.
4. Independence of Irrelevant Alternatives (IIA): For any preference profiles $R$ and $R'$ and any $B \in \mathcal{B}$ such that $R$ and $R'$ are identical when restricted to $B$ then $C(B, R) = C(B, R')$.

Of there axioms, one could argue with the desirability of probably all but unanimity. IIA is probably the one most often relaxed, but I don't see a particularly compelling reason to ask for rationality here either. There is a literature that is very concerned with formulating alternative axioms and proving new impossibility results.[15] Anyway, here's our big theorem

**Theorem 8.** *(Arrow's Theorem) Suppose $|X| \geq 3$. A social choice correspondence satisfies 1-4 if and only if it is dictatorial, i.e. there exists an $i$ s.t.*
$$C(R, B) \subseteq \{x : xR_iy \, \forall y \in B\}$$
*for all $R$ and $B$.*

Now let's turn to the question of incentives.

Keep the setting from before, but now let's work with social choice functions $f : \mathcal{R} \to X$, and focus on settings where $\mathcal{R}$ is the set of all strict preference profiles.[16] We work with functions instead of correspondences here, as conceptually our question is a

[15] This literature is also pretty goofy.

[16] See MWG for an extension to larger sets of preferences

bit different. In the Arrow's theorem world, preferences were known and $C$, the correspondence identified which allocations society should choose given those preferences. Here, our function $f$ is trying to assign an allocation in order to create incentives to reveal private information. We can define a concept like DIC here

**Definition 1.** *A social choice function is manipulable on $\mathcal{R}$ by player $i$ at profile $R$ iff there exists an $R_i'$ such that*

$$f(R_i', R_{-i})P_i f(R).$$

*A social choice function is non-manipulable if it is not manipulable for any $i$ at any $R$.*

Note that a social choice function is non-manipulable iff a truthfully reporting preferences is a dominant strategy equilibrium.[17] By the revelation principle, anything we can implement in dominant strategies some game, we can implement in dominant strategies in a direct revelation game. Unfortunately, it turns out that there's not much we can do here.

**Theorem 9.** *Suppose that $|f(\mathcal{R})| \geq 3$ and $f : \mathcal{R} \to X$ is non-manipulable. Then it is dictatorial, i.e. there exits an $i$ s.t. for all $P$ $f(P) \in \{x : xP_i y \, \forall y \neq x\}$.*

*Proof.* I doubt I'll do this in class. Here I follow Schmeidler and Sonnenschein who prove it by showing that $f$ must satisfy the assumptions of Arrows theorem.

For simplicity, let $X = f(\mathcal{R})$. First, let's establish unanimity. To do this, we first need to construct a social choice correspondence.

**Lemma 1.** *Take any non-empty $A$, $B$ such that $A \cup B = X$ and $A \cap B = \emptyset$. Let $P$ be the preference profile for which all individuals prefer any element in $B$ to any element in $A$. Then $f(P) \in B$.*

*Proof.* Suppose not, $f(P) \in A$. There must exist a profile $P'$ s.t. $f(P') \in B$. Consider the sequence of profiles:

$$\begin{aligned} P^0 &= P \\ P^1 &= (P_1', P_2, \dots P_N) \\ P^2 &= (P_1', P_2', \dots P_N) \\ &\dots \\ P^N &= P' \end{aligned}$$

Since $f(P') \in B$, there must be a $k \leq N$ s.t. $f(P^k) \in B$ and $f(P^{k-1})$ is not. But then truthful reporting is not dominant for individual $k$, as if they lied and reported $P_k'$

[17] We are working directly with preferences, instead of utility as is standard in textbook game theory. This makes the link with Arrow's theorem clear, and is a bit more general.

instead of $P_k$, they would receiver $f(P^k)P_k f(P^{k-1})$. □

Now for each $\{x, y\}$ define $C(P, \{x, y\})$ as $f(P')$ where $P'$ is the preference order we get when we move $\{x, y\}$ to the top of each agent's preference order, but leave the order of alternatives otherwise unchanged. Then use the induced preference order $F(P)$ to extended $C$ to all feasible sets.[18] By the previous lemma, this satisfies unanimity and unrestricted domain.

**Lemma 2.** *$C$ satisfies IIA if $f(p)$ is non-manipulable.*

*Proof.* Suppose not. Then there exists a $B$ and a $P$ and $P'$ s.t. $x = C(B, P)$, $y = C(B, P')$, $x \neq y$. Define $\hat{P}$ and $\hat{P}'$ as the profiles we get from moving $x$ and $y$ to the top of $P$ and $P'$ respectively, and then define sequence of preferences from $P^0 = \hat{P}$ to $P^N = \hat{P}'$ as before. By Lemma 1 and the construction of $C$ $C(B, \hat{P}) = x$ and $C(B, \hat{P}') = y$. But, since the preference order is strict, at the first $k$ where the choice switches from $x$ to $y$ $f$ is manipulable (either at $P^{k-1}$ or $P^k$). □

**Lemma 3.** *$C(P, B)$ is rational.*

*Proof.* Suppose not. Then for some $P$ the induced preference order $F(p)$ is not transitive. So there exist $x, y, z$ s.t. $xF(p)y$, $yF(P)z$, and $zF(P)x$. Let $P'$ be the preference order that has $x, y, z$ at the top but otherwise is unchanged. WLOG suppose $f(P') = x$. Now let $P''$ be the preference order we get from taking $P'$ and moving $y$ to third place. Then by IIA and the observation that $zF(P)x$, it must be that $f(P'') = z$. Now consider the sequence of preferences going from $P'$ to $P''$. This is, by similar logic manipulable at the first $k$ where it stops choosing $x$. Since $F(p)$ is complete and transitive, it represents $C(B, P)$ by construction. □

So $C(B, P)$ satisfies the axioms from Arrows theorem, so it is dictatorial. Therefore $f(P)$ must also be dictatorial. □

So, if we'd like to have a dominant strategy incentive compatible mechanism that works for all preferences, we're in general out of luck. The natural approach to receive is to restrict the domain. Two restrictions are pretty common. In the literature on voting, we tend to restrict to single-peaked preferences (which you may remember from micro 2). Here, we'll continue with our restriction to quasilinear preferences.

[18] $xF(p)y$ iff $C(P, \{x, y\}) = x$. Always choose the $F(p)$ maximal elements. Note that it's not clear that this always results in a non-empty choice, or a single choice.

## 4.2 Ex-Post Efficiency

Now let's think about whether we can find a dominant strategy mechanism that implements an ex-post efficient allocation in quasilinear settings. Let $X$ be a set of possible allocations, and assume an agent's utility from allocation $x \in X$ and transfer scheme $t_i \in \mathbb{R}^N$ is $v_i(x) - t_i$. Let the utility function $v_i$ be the agent's private information. This is a private value setting.

For any vector of utility function $v = (v_1, \dots v_n)$ we'd like to implement the ex-post efficient allocation

$$x^*(v) = \max_{x \in X} \sum_{i=1}^{N} v_i(x).$$

Since we are in a quasilinear setting, any ex-post efficient outcome must select $x^*$. Is there a DIC mechanism that implements this.

This is a much more general private values setting than we've been dealing with. It captures a lot of different environments where adverse selection is a natural friction. For instance

- Public goods allocation,
  - $x \in \Delta(\{0, 1\})$ probability of producing public good.
  - Known per-capita cost of production $c$, private value $v_i$
- Auctions
  - Set of indivisible goods (single or multiple, finite or infinite, etc.), $X$ is set of possible (randomized) assignments.
- Bilateral Trade/Incomplete Info Bargaining
  - Buyer and seller, each have private value for a single indivisible good.

Can we implement the ex-post efficient allocation in these settings?

Sure (sort of). There are two ways we could approach this problem. We could either (i) put more structure on the space of preferences and use the envelope theorem or (ii) observe that the following mechanism clearly works without any further assumptions. Let $t_i$ be

$$t_i(v) = h_i(v_{-i}) - \sum_{j \neq i} v_j(x^*(v))$$

for some arbitrary function $h(\cdot)$. Why does this work? If I deviate, I receive some other

$x \in X$. Then my payoff is

$$\sum_{i=1}^{n} v_i(x) - h(v_{-i}) \leq \sum_{i=1}^{n} v_i(x^*(v)) - h_i(v_{-i})$$

since if I truthfully report, I receive the allocation that maximizes the sum. I'm paying everyone the ex-post welfare and then subtracting at term that is independent of their report, which doesn't impact incentives.

This family of mechanisms are called Vickery-Clark-Groves (VCG) mechanisms. The following, somewhat imprecisely stated, is a consequence of the envelope theorem.

**Theorem 10.** *Suppose that the set of possible $v$'s is sufficiently "rich", then any DIC mechanism that implements the efficient allocation is a VCG mechanism.*

As a quick sanity check, we know that the second price auction in private values settings is one such mechanism. Let $x$ be a vector s.t. $x_i$ is the probability that $i$ is allocated the good, and in a slight abuse of notation $v_j(x) = v_j$ if $x_j = 1$, and 0 otherwise. The second price auction is a VCG mechanism where $h_i(v_{-i}) = \max_{x \in X} \sum_{j \neq i} v_j(x) = \max_{j \neq i} v_j$.

At first, this theorem looks encouraging. Now if we want to find an efficient DIC mechanism, the problem collapses to choosing nice $h_i$'s. Let's think about this more carefully.

Suppose that $v_i(x) > 0$ for all $x \in X$ and all utility functions. Also suppose that $v_i(x) = 0$ for all $x$. Suppose we are worried about our agent walking away from the mechanism, so we introduce an IR constraint

$$v_i(x^*(v)) - t_i(v) \geq 0.$$

Plugging in our VCG transfers, we get

$$\sum_{j=1}^{n} v_j(x^*(v)) - h_i(v_{-i}) \geq 0$$

So, largest $h_i(v_{-i})$ is $h_i(v_{-i}) = \max_{x \in X} \sum_{j \neq i} v_j(x)$. This is called Pivot mechanism, it is the most commonly used form of the mechanism.

**Example 2.** (Public Goods). *A government is deciding whether or not to produce a public good*

- $v_i \geq 0$ *is value of the public good.*
- $c$ *is a (known) per-capita cost of production.*

- 0 *is value of no public good.*
- *Recast utility as* $v_i - c$ *(e.g. additional transfer of costs* $c$ *whenever public good gets produced).*

*The pivot mechanism is a DIC tax scheme that always covers costs:*

$$t_i(v) = \begin{cases} -\sum_{j\neq i} v_j + (N-1)c \text{ if } \frac{\sum_{j\neq i} v_j}{N-1} \leq c \text{ and } \frac{\sum_{j=1}^{N} v_j}{N} \geq c \\ \sum_{j\neq i} v_j - (N-1)c \text{ if } \frac{\sum_{j\neq i} v_j}{N-1} \geq c \text{ and } \frac{\sum_{j=1}^{N} v_j}{N} < c \\ 0 \text{ o.w.} \end{cases}$$

*By construction, are transfers are always positive, so we are always able to afford the public good. But, what if we were worried that the players in this game could move to Sweden and not contribute to the public good. We can modify pivot mechanism to deal with this new participation constraint,*

$$t_i(v) = \max_{x\in[0,1]} x[(0-c) + \sum_{j\neq i}(v_j - c)] - x^*(v)\sum_{j\neq i}(v_j - c)$$

*instead welfare without agent* $i$ *for* $h_i(\cdot)$*, we're now using the welfare if they were the lowest type. This new "pivot" mechanism (now including the tax) is*

$$t_i(v) + cx^*(v) = \begin{cases} -\sum_{j\neq i} v_j + Nc \text{ if } \frac{\sum_{j\neq i} v_j}{N} \leq c \text{ and } \frac{\sum_{j=1}^{N} v_j}{N} \geq c \\ 0 \text{ o.w.} \end{cases}$$

*So we can design a mechanism that builds the public good only when it's benefit exceeds its costs. But, there's something fishy here. Our optimal mechanism doesn't cover costs, as only types that are pivotal in the decision pay taxes. If the government wants to run this public goods mechanism, they are going to have to also pay for some of the costs themselves.*

This example highlights a potential issue. We haven't talked at all about the numeraire good. Budget balance, $\sum t_i(v) = 0$ for all $v$ seems like a natural requirement to impose here. But, we've already seen that if we also require IR, all transfers are positive. In some settings, this isn't a problem. For instance, if we introduce a "player 0" with no private information, these transfers can be paid to that player. That mechanism provides all players with positive utility, is IC and is budget balanced. In some settings (e.g. auctions), this makes a lot of sense. In others (e.g. public goods), this makes less sense.

The IR constraint made it clear that budget balance is a problem in the example, but it's clear from working out the combinatorics that this is a problem in general. For

each preference profile, budget balance requires that

$$\sum_{j=1}^{n} h_j(v_{-j}) - (N-1)\sum_{i=1}^{n} v_i(x(v)) = 0$$

This is an equation with $n$ unknowns, and ex-post budget balance requires that this holds for every preference profile. Consider any pair of preferences for each type $i$, $v_i$ and $\hat{v}_i$. There are $2^n$ preference profiles where each type has one of these two preferences. So, we have $2^n$ equations of the above form that must be satisfied. But, at the same time, there are only $n2^{n-1}$ different $h$'s.[19] So we have a system of way more equations than unknowns, so unless we are in a really pathological setting we're in bad shape.[20]

VCG mechanisms have a number of other weaknesses. Beyond being incompatible with budget balance, they are also susceptible to shill bidding, collusion between bidders, etc. and perform badly if bidders have for instance limited budgets (so can't pay large transfers).

**A BIC, Budget balanced, and Ex-post Efficient Mechanism.** We can relax incentive compatibility and resolve some of these issues. By similar logic to above, any transfer scheme of the form

$$t_i(v) = -E(\sum_{j\neq i} v_j(x^*(v))|v_i) + h_i(v_{-i})$$

is Bayesian incentive compatible. But now, the first term in each transfer is independent of the other types reports. We can leverage this to get budget balance, by dividing the payment we need to make to each bidder to generate incentives amongst the other bidders, i.e.

$$t_i(v) = -E(\sum_{j\neq i} v_j(x^*(v))|v_i) + \frac{1}{n-1}\sum_{j\neq i} E(\sum_{k\neq j} v_k(x^*(v))|v_j).$$

It's easy to see that this is both budget balanced and Bayesian incentive compatible. Note that it is not necessarily interim individually rational. Unlike the VCG mechanism, this mechanism is also sensitive to the fine details of the setting. In order to design the transfer scheme, the designer needs to understand the distribution types are drawn from. More on both these points later.

[19] If $n = 2$, we have to consider triples instead of pairs, but the same heuristic argument applies

[20] We can formulate this as solving for the $h$ that satisfies a linear equation of the form $Ah = b$. No matter which pairs of preferences we choose, the matrix $A$ is the same, as long as there is some profile that gives us ex-post efficient allocations that don't make every $b$ end up in the column space of $A$, we're done. This is going to be very hard to satisfy for any setting where for instance, there are a continuum of possible preference profiles.

### 4.3 Bilateral Trade and Myerson Satterthwaite

Let's think about a very simple market model. Suppose there are a buyer and a seller, each with a private value for a single good. Is there a bargaining procedure that allows them to efficiently bargain for this good? In the complete information case, we know the answer to this is yes (recall alternating offer bargaining from micro 3). It seems reasonable to conjecture that even in this setting, the gains from trade could compensate for any information rents due to the buyer or the seller. It turns out, this is not the case, and any bargaining protocol must introduce some inefficiencies.

Consider the following setting. There is a buyer $B$ with valuation $v \in [\underline{v}, \bar{v}]$ for a good. There is a seller $S$ who can produce exactly one unit of the good by paying cost $c \in [\underline{c}, \bar{c}]$. Assume there are potentially gains from trade, so $\bar{v} > \underline{c}$ but the possibility of these gains is not unambiguous, so $\underline{v} \leq \bar{c}$.

The ex-post efficient allocation here is as follows

$$x(v,c) = \begin{cases} 1 \text{ if } v > c \\ 0 \text{ if } v < c \end{cases}$$

where $x(v,c) = 1$ denotes that the good is produced and given to the buyer, and $x(v,c) = 0$ denotes the case where the good is not produced. We'd like to find a BIC mechanism that implements this allocation, is interim individually rational and is ex-post budget balanced. So we have the following constraints (for all $v, c, v', c'$)

$$\begin{aligned} E(vx(v,c) - t_B(v,c)|v) &\geq E(vx(v',c) - t_B(v',c)|v) \\ E(t_S(v,c) - cx(v,c)|c) &\geq E(t_S(v,c') - cx(v,c')|c) \\ E(vx(v,c) - t_B(v,c)|v) &\geq 0 \\ E(t_S(v,c) - cx(v,c)|c) &\geq 0 \\ t_S(v,c) - t_B(v,c) &= 0 \end{aligned}$$

In fact, we can weaken the budget balance requirement quite a bit without changing the result. Instead of budget balance, let's require that the mechanism requires no expected subsidy

$$E(t_S(v,c) - t_B(v,c)) \leq 0,$$

so in expectation the buyer is paying a larger transfer than the seller is receiving.

**Theorem 11.** *There is no mechanism that is ex-post efficient, interim IR, BIC, and no expected subsidy*

*Proof.* I'll prove this in the case where $\underline{v} \leq \underline{c}$ and $\bar{c} \geq \bar{v}$. Nothing really changes if we

don't assume this, but it avoids some messy stuff.

There are a number of natural ways to approach this. We could for instance, try to solve for the mechanism that maximizes ex-ante surplus, because if an efficient mechanism existed, then every solution to that principal agent problem would be ex-post efficient. We are instead going to use the envelope theorem to pin down interim transfers, and argue that they are then incompatible with no expected surplus.

Let $s^*(v,c) = (v-c)x(v,c)$ be the surplus generated in the efficient allocation. Fix $V_S(\underline{v})$ and $V_B(\bar{c})$, the IR constraint requires that both of these are positive. We know from the envelope theorem that any ex-post BIC mechanism that gives these lowest types the same interim utility also gives all other types the same interim utility and has the same interim transfers. We also know that every DIC mechanism is BIC, so instead of working with those integrals we get out of the envelope theorem, we can work with the VCG mechanism

$$\begin{aligned} t_S(v,c) &= vx(v,c) + V_S(\bar{c}) \\ t_B(v,c) &= cx(v,c) - V_B(\underline{v}) \end{aligned}$$

which give ex-post utilities

$$\begin{aligned} V_B(v,c) &= s^*(v,c) + V_B(\underline{v}) \\ V_S(v,c) &= s^*(v,c) + V_S(\bar{c}). \end{aligned}$$

Even if these transfers don't satisfy no expected subsidy (which they pretty clearly don't), that doesn't immediately rule out the existence of a mechanism that does. But, by the envelope theorem, we know that any BIC mechanism provides interim utility

$$\begin{aligned} V_B(v) &= E(s^*(v,c)|v) + V_B(\underline{v}) \\ V_S(c) &= E(s^*(v,c)|c) + V_S(\bar{c}) \end{aligned}$$

which in turn means that any BIC mechanism has interim transfers

$$\begin{aligned} T_B(v) &= E(vx(v,c) - s^*(v,c)|v) - V_B(\underline{v}) \\ T_S(c) &= E(cx(v,c) + s^*(v,c)|c) + V_S(\bar{c}). \end{aligned}$$

Finally, the ex-ante subsidy required is

$$\begin{aligned} E(T_S(c) - T_B(v)) &= E((c-v)x(v,c) + 2s^*(v,c)) + V_S(\bar{c}) + V_B(\underline{v}) \\ &= E(s^*(v,c)) + V_S(\bar{c}) + V_B(\underline{v}) \\ &> 0 \end{aligned}$$

where the last inequality comes from the possibility of gains from trade (the overlapping supports) and the IR constraints which force $V_B$ and $V_S$ to be positive. So any BIC, IR mechanism must be subsidized in expectation. □

This result tells us that, even though there is private information on both sides, the surplus generated by trade is not enough to compensate for information rents. Unlike in complete information settings, incomplete information bargaining is necessarily inefficient. Under any bargaining protocol, there are $v$'s and $c$'s where trade is profitable, and yet agreement is reached late, if at all.

This is an important result in information economics. It tells us that appealing arguments like Coase's theorem don't necessarily hold in settings with adverse selection. We get a natural source of delays in bargaining. And we have another result where with adverse selection there's no way decentralization (or even centralization) can lead us to a fully efficient outcome. Beyond our already discussed assumptions, this result takes advantage of continuous support a bit more aggressively. It's not obvious that as striking a result holds in discrete settings. One might imagine (correctly) that adding more buyers and/or sellers, or making the initial ownership of the thing being bargained over more symmetric would also alleviate some of the inefficiencies.

### 4.4 Wrap up

As we've seen, asymmetric information can present an obstacle to implementing the efficient allocation, even under the restriction to quasilinear preferences. The VCG mechanism is essentially the unique class of mechanisms that implement efficient outcomes in dominant strategies, but it is hard to make this mechanism line up with natural constraints like IR constraints or budget balance.[21] In some settings, like auctions, the lack of budget balance is perhaps a less severe problem, as the seller serves as a natural third party who can absorb the excess transfer. And in-fact we do see VCG-like mechanisms used in practice, not only the second price auction, but also more complex multi-unit auctions are often run with a VCG inspired pricing scheme.[22] But, when used in practice, a number of the issues we haven't really discussed like issues with budget constraints, collusion, shill bidding, and whether or not the winner determination problem can be efficiently solved all become very real problems.

The Myerson-Satterthwaite theorem is another place where we can really see how adverse selection distorts things that would have been easy to do without the informa-

[21] There is a BIC mechanism that satisfies budget balance and ex-ante IR called the expected externality mechanism. We may do this in class or on a problem set. Interim IR, BIC and budget balance pose similar problems to those posed by ex-post IR and DIC.

[22] Spectrum auctions for instance tend to use something like VCG. The combinatorial clock auction for instance selects the prices closest to the VCG prices that lie in the core (as defined by bids), the deferred acceptance auction selects something like the VCG prices, etc.

tion asymmetry. Without this two sided private information, not only can we implement the efficient outcome, but an efficient outcome is selected by one of the most natural bargaining protocols. Myerson-Satterthwaite tells us that introducing asymmetric information makes this impossible, as the information rents required exceed the surplus that is generated. A number of natural questions arise from this analysis. Can we implement the "constrained efficient" outcome through a simple bargaining procedure?[23] What if we add correlation?[24] can adding a third party somehow help? What if allowed both players to own a share of what's being bargained over?

A final question we haven't addressed is what happens in settings like our General Equilibrium setting from micro 2 with adverse selection. The answer is that it's in general a problem, the analysis is sort of a mess and not worth going into here. It turns out that the equilibrium in these settings is not only not efficient, it often isn't even constrained efficient (i.e. the best outcome that can be implemented by any mechanism). Depending on time, and how I feel, we may talk more about this at the end of the course.

## 5 Moral Hazard

We've now spent a lot of time studying adverse selection, settings where differences in information due to hidden types led to natural distortions. Now let's turn to another natural source of asymmetric information, moral hazard. Moral hazard describes situations where the difference in information comes from hidden actions; my manager doesn't know if I worked hard or not, an insurance company doesn't know if I'm drinking a few pineapple long drinks before getting behind the wheel.

Like with adverse selection, there is a massive literature documenting the many ways moral hazard can distort economic outcomes. Unfortunately, the technical details of this class of models quickly become dense. In this class, we'll make some somewhat stark modeling assumptions that allow us to avoid these technical complications while delivering some of the main messages of this literature. There's a lot more out there, but it requires a lot more tedious heavy lifting than I want to do for this course.

### 5.1 The Principal Agent Model

There's a firm and the firm employs a worker. The worker can exert effort $e \in \{0, 1\}$, with increasing cost of effort is $c(1) > 0, c(0) = 0$ and this effort produces output $y \in Y$ drawn from a distribution described by pdf/pmf $f(y|e)$ where higher levels of effort lead to distributions that FOSD the distribution for lower levels of effort. The firm offers the worker a wage contract $w(e, y)$, which the worker can always choose to reject and receive utility 0.

[23] Yes for some of them, this is Ausubel and Deneckere 1993.

[24] We can do stupid things in BIC mechanisms, as in the auction setting. DIC is still clearly a problem.

Unlike in our adverse selection settings, let's think about settings with risk aversion here (more on why in a few pages). The principal has utility $v(y - w(e, y))$ and the agent has utility $u(w(y, e)) - c(e)$, where $u$ is concave, increasing, twice continuously differentiable, and that there is a wage $w$ s.t. $u(w) - c(e) < 0$ for all $e$, as well as a wage where $u(w) - c(e) \geq 0$ for all $e$. The principal solves for an incentive compatible wage contract that maximizes their payoff, i.e.

$$\begin{aligned}
&\max_{e,w} \int_Y v(y - w(y, e)) f(y|e)\, dy \\
&\text{s.t.} \int_Y u(w(y, e)) f(y|e)\, dy - c(e) \geq \int_Y u(w(y, e')) f(y|e')\, dy - c(e') \text{ for all } e' \\
&\int_Y u(w(y, e)) f(y|e)\, dy - c(e) \geq 0
\end{aligned}$$

This is hopefully an unsurprising set of constraints. The first constraint makes sure the principals choice of action is incentive compatible. The second makes sure it is individually rational.

This problem does not have moral hazard as effort $e$ is directly contractible. And it's pretty easy to solve. First, the incentive constraint isn't a problem for us. If the principal sees a level of effort they don't like, they just set the wage put the agent's utility below their reservation value. So, the problem can be simplified to

$$\begin{aligned}
&\max_{e,w} \int_Y v(y - w(y, e)) f(y|e)\, dy \\
&\int_Y u(w(y, e)) f(y|e)\, dy - c(e) \geq 0
\end{aligned}$$

Let $\lambda$ be the multiplier on the IR constraint. Then if we relax the constraint into the objective and maximize pointwise we get first order condition

$$v'(y - w(y, e)) f(y|e) = \lambda u'(w(y, e)) f(y|e)$$

which gives us the "Borch condition for optimal risk sharing. For any pair of states $y, y'$, we set the wage to equalize the marginal rates of substitution.

$$\frac{v'(y - w(y, e))}{v'(y' - w(y', e))} = \frac{u'(w(y, e))}{u'(w(y', e))}.$$

This immediately tells us that when the principal is risk neutral and the agent is risk averse, the principal takes on all risk in the optimal contract, and pays a constant wage. We know the best wage to induce either level of effort, now to solve for the optimal contract

Let's continue analyzing this by focusing on the case with a risk neutral principal. But now, let's assume that $e$ is not contractible. Now the principal's problem becomes

$$\begin{aligned}
&\max_{e,w} \int_Y (y - w(y)) f(y|e)\, dy \\
&\text{s.t.} \int_Y u(w(y)) f(y|e)\, dy - c(e) \geq \int_Y u(w(y)) f(y|e')\, dy - c(e') \text{ for all } e' \\
&\int_Y u(w(y)) f(y|e)\, dy - c(e) \geq 0
\end{aligned}$$

Implementing $e = 0$ is still easy. Since it's the cheaper level of effort, the same wage scheme from before works. So let's think about what the wage that implements $e = 1$ looks like. This sub problem is

$$\begin{aligned}
&\max_{w} \int_Y (y - w(y)) f(y|1)\, dy \\
&\text{s.t.} \int_Y u(w(y))(f(y|1) - f(y|0))\, dy - c(1) \geq 0 \\
&\int_Y u(w(y)) f(y|1)\, dy - c(1) \geq 0
\end{aligned}$$

This problem looks a bit more challenging. First, it's not clear (at least to me) whether both constraints bind. It's clear that IC must always bind. If it didn't, the optimal wage would be constant (we know that from the analysis when $e$ was contactable), but a constant wage always induces low effort. It's a bit less clear whether IR also binds.

To think about this, let's try to reason through that in a simple example before taking derivatives. Suppose there are two levels of output $y_H$ and $y_L$

We can reformulate the principal's problem as (after rearranging some terms)

$$\begin{aligned}
&\max_{\tilde{u}} \sum_Y (y - u^{-1}(\tilde{u}(y))) f(y|1)\, dy \\
&\text{s.t. } (\tilde{u}(y_H) - \tilde{u}(y_L))(f(y_H|1) - f(y_H|0)) - c(1) \geq 0 \\
&(\tilde{u}(y_H) - \tilde{u}(y_L)) f(y_H|1) + \tilde{u}(y_L) - c(1) \geq 0
\end{aligned}$$

The principal now chooses the level of utility the agent receives for each level of output. This problem is clearly equivalent. Suppose IC binds but IR was slack at the optimum. What if we lower $u(y_H)$ by $\varepsilon$ and $u(y_L)$ by $2\varepsilon$ for $\varepsilon$ small enough that the IR constraint still holds. Then the left hand side of the IC constraint becomes

$$(\tilde{u}(y_H) - \tilde{u}(y_L) + \varepsilon)(f(y_H|1) - f(y_H|0)) - c(1) > (\tilde{u}(y_H) - \tilde{u}(y_L))(f(y_H|1) - f(y_H|0)) - c(1)$$

and this contract is more profitable for the principal as she has to pay a lower wage

at both levels of output. Intuitively, we are reducing the wage in both states, but at the same time making low effort relatively riskier, so incentives are easier to satisfy. Therefore, we know the IC and IR constraint both must bind at the optimum.

Now back to the general problem. If we look at the first order conditions, letting $\lambda$ be the multiplier on IR and $\mu$ be the multiplier on IC. Maximizing pointwise, we get

$$\begin{aligned} f(y|1) &= \mu u'(w(y))(f(y|1) - f(y|0)) + \lambda f(y|1) \\ \frac{1}{u'(w(y))} &= \mu\left(1 - \frac{f(y|0)}{f(y|1)}\right) + \lambda \end{aligned}$$

As we noted before, the optimal wage is not constant. Somewhat surprisingly, it's not necessarily monotone either. For monotonicity, we need $f(y|0)/f(y|1)$ to be decreasing. This is called the *monotone likelihood ratio property* (MLRP). MLRP implies first order stochastic dominance but first order stochastic dominance does not imply MLRP.

The optimal wage is always, in expectation, higher than the first best wage, but it is also riskier for the agent. The principal can no longer both absorb all risk and provide incentives simultaneously.

If we allowed for more $N$ levels of effort $(e_i)_{i=1}^N$, a similar analysis would give a condition that looked like

$$\frac{1}{u'(w(y))} = \sum_{i=1}^{N} \mu_i \left(1 - \frac{f(y|e_i)}{f(y|e^*)}\right) + \lambda$$

to implement effort level $e^*$, suggesting that outputs with higher $f(y|e^*)/f(y|e_i)$ across effort levels are given more weight in the wage scheme. This ratio captures the effectiveness of each $y$ at distinguishing effort $e^*$ from other levels of effort (more on this on the problem set). The multipliers in turn capture which deviations are more "valuable", loading incentives on outputs that have higher likelihood ratios and higher multipliers.

With more than two levels of effort, this condition is somewhat misleading. It's not clear which of these constraints actually bind, and for any incentive constraints that are slack the corresponding likelihood ratio plays no role in the optimal wage. It may even be that a desired level of effort is impossible to implement. Analyzing this problem quickly becomes a big mess. One might hope that if we move an interval of possible levels of effort, this would become easier. But, unlike in the adverse selection settings we've studied, simplifying the IC constraints there is challenging, and introduces a host of technical complications.[25] I leave further analysis to future courses.

[25]It's tempting to replace the IC constraints with the first order conditions of the agents problem for a given wage. This approach, called the first order approach does appear to make the problem simpler. But, the solution we get out of it, even for relatively well behaved distributions, is a wage that makes the optimal level of effort satisfy the agent's first order conditions but the optimal level of effort is not

Finally, while we interpreted $y$ as output, we could do a similar analysis where the principal sees other signals in addition to, or instead of, actual output. Not much changes, the principal still wants to load incentives onto the signals that are best at identifying the desired level of effort.

### 5.2 Risk Neutrality and Limited Liability

We've focused on the role of moral hazard in distorting risk sharing away from the optimal. What if both the principal and the agent are risk neutral.

This problem is a bit more straightforward. Suppose the agent is choosing a level of effort $e \in [0,1]$ for convex, twice continuously differentiable cost $c(e)$ with $c(0) = 0$. Effort produces output $y = v > 0$ with probability $e$, i.e. $Pr(y = v|e) = e$, and $y = 0$ otherwise. Now the principal's problem becomes

$$\begin{aligned}
&\max_{e,w(y)} e(v - w(v)) + (1-e)w(0) \\
&e \in \arg\max_{x\in[0,1]} x(w(v)) + (1-x)w(0) - c(x) \\
&ew(v) + (1-e)w(0) - c(e) \geq 0.
\end{aligned}$$

Let $e^*$ denote the first best effort, which maximizes $ev - c(e)$. So the first best level of effort satisfies

$$0 = v - c'(e).$$

The principal is in luck here. Since the agent doesn't care about risk, I can just sell them the firm. I set $w(v) = v - (e^*v - c(e^*))$ and $w(0) = c(e^*) - e^*v$.The agent takes on all the risk, so they have the right incentives to work, but then the pay the principal all the surplus they generate in expectation. Conceptually principal asks the agent to pay them the firm's expected value, and then gives them the firm.

This is a bit of an unsatisfying answer. One might worry that asking a worker to pay me their expected output net costs and then paying them back their realized output is not really a feasible contract as it will often, for instance, result in the worker paying the firm. So, let's add the constraint that the principal must always pay the worker and not the other way around. This is called limited liability. Then our principal solves

$$\begin{aligned}
&\max_{e,w(y)} e(v - w(v)) + (1-e)w(0) \\
&e \in \arg\max_{x\in[0,1]} x(w(v)) + (1-x)w(0) - c(x) \\
&ew(v) + (1-e)w(0) - c(e) \geq 0
\end{aligned}$$

a maximizer. This is because the concavity of the agent's problem is endogenously determined by the wage chosen by the principal.

$$w(v), w(0) \geq 0.$$

It's clear that $w(0) = 0$ in the optimal contract. Let's simply denote $w(v)$ as $w$. The agent's first order condition is

$$w - c'(e) = 0$$

Since all wages are positive, IR isn't an issue here, since agent could always choose 0 effort. So the principal simply solves

$$\max e(v - c'(e))$$

and we get that the optimal level of effort solves

$$0 = v - c'(e) - ec''(e).$$

Recall that the first best was the solution to

$$0 = v - c'(e)$$

so in order optimally deliver incentives, the principal must lower the level of effort they ask for. This makes sense, in order to get the first best effort, we can see from the first order condition that they would have to pay $w = v$, the entire output generated by the agent. Without limited liability, they could regain a share of the output without messing up incentives by lowering the wage after high output and simultaneously increasing the penalty for low output. Now that they no longer have this lever, they choose to provide the agent with rents, in the form of lower required effort, in order to decrease the wage necessary to create incentives.

### 5.3 Wrap up

We've studied a natural source of differences in information, unobservable actions. Ideally, to combat this, our principal would like to design a wage scheme that aligns incentives but, as we've seen, this is easier said than done.

The text already discusses the many technical headaches that exist even in this setting, one can imagine moving to more complicated settings doesn't make things much easier. One thing to note here is that the optimal contract is in general quite complex, carefully tailoring incentives to the informational content of every signal. Contracts we see in practice in general seem to lack this complexity. There are a number of papers that posit explanations for this and try to find a model that rationalizes linear contracts. This includes models that take into account dynamic concerns or robustness concerns, I'm not sure any provide what feels like "the answer".

Don't let my relatively brief treatment of moral hazard fool you, there's a lot of cool stuff here. Some natural things to cover here that I'm not include moral hazard in teams, relational contracts and incomplete contracts. These are all cool directions to go in. There is now also a large literature on dynamic contracts in countinuous time that takes advantage of some of the tools you may have learned about in macro, a nice literature on behavioral contract theory, and an interesting literature on contracting with unawareness, where you might be worried that a specific wage scheme will call attention to deviations the agent hasn't considered, that is still sort of searching for a good model.

## 6 Communication Games

Let's think some more about hidden types. We are often in situations where we have private information that we'd like to communicate to someone who has incentives to use the information in a way that is different than we would like. To make that bad sentence clearer, think about some examples:

- A student is trying to convey to an employer that they are talented.
- A banker is trying to persuade a consumer into making a particular investment.
- A politician want to convince voters that their policies will be effective.
- blah blah more examples.

When is communication between individuals with conflicts of interest effective? To study this, let's study the following class of games:

- Two players, $s$(ender) and $r$(eceiver).
- $s$ has unknown type $\theta \in \Theta$, $\theta \sim \mu_0$.
- First stage, $s$ chooses a message $m \in M$.
- Second stage, $r$ sees message, chooses action $x \in X$.
- Payoffs $u_i(m, x, \theta)$.

Many games fall into this paradigm. Some of the leading examples:

- Cheap talk: $u_i(m, x, \theta) = u_i(m', x, \theta)$.
- Signaling: $u_s(m, x, \theta) \neq u_s(m, x, \theta')$.
- Disclosure: $M$ depends on type, e.g. $M(\theta) = \{\theta, \emptyset\}$.
- Bayesian Persuasion: Cheap talk with commitment.

We won't discuss all of these, let's start by thinking about signaling.

## 6.1 Signaling

The classic signaling story is as follows, you'd like to tell an employer that you are a high skilled worker to get a higher salary. But, obviously, even if you were a low skilled worker you'd like to tell them that. What can I do? Well, there are other activities that are relatively less costly for a high skilled worker than a low skilled worker. Even if these activities aren't productive or useful, I still benefit from say attending the University of Helsinki, as employers understand that it's less costly for a high skilled worker to waste their time getting this degree than it would be a low skilled worker.

To formalize this story, consider the following sender-receiver game:

- Let $\theta \in \{1, 2\}$, $M = \mathbb{R}_+$, prior $\mu_0 = Pr(\theta = 2)$
- Lets now call sender action $e$(ffort), receiver action $w$(age).
- Assume agent utility is $w - c(e, \theta)$.
- Receiver payoffs, $u_r(w, e, \theta) = -(w - f(e, \theta))^2$.

Assumptions:

- Standard:
  - $c(e, \theta) \geq 0$, $c_e(e, \theta) \geq 0$, $c_e(0, \theta) = 0$, $c_{ee}(e, \theta) > 0$ and $\lim_{e \to \infty} c_e(e, \theta) = \infty$.
  - $f(e, \theta) \geq 0$, $f_e(e, \theta) \geq 0$, $f_\theta(e, \theta) > 0$, $f_{ee}(e, \theta) \leq 0$.
- Single Crossing (standard for us):
  - $\partial^2 c(e, \theta)/\partial e \partial \theta < 0$
  - $\partial^2 f(e, \theta)/\partial e \partial \theta \geq 0$

In the first best, effort maximizes $f(e, \theta) - c(e, \theta)$ for each type. How close can we get to the first best in equilibrium?

What does an equilibrium look like here. This is a dynamic game of incomplete information, so Perfect Bayesian Equilibrium (PBE) is a natural concept to apply. Given the simple structure of the game, it's easy to see that all our different versions of PBE require the same things.

What do we need to solve for here? In a pure strategy PBE we need to specify the receivers strategy $w : M \to \mathbb{R}$, the senders strategy $\sigma : \{1, 2\} \to M$, and a system of beliefs following every message $\mu_M \in \Delta(\{1, 2\})$, which are consistent with the sender's strategy.

The sender's strategy is pretty boring here. They simply pay a wage equal to the expected $f(e, \theta)$ following each level of effort. This expectation is pinned down by

$\mu_M$, which in turn is pinned down on path by $\sigma$. There are only so many things that can happen in a pure strategy equilibrium. The two types could *separate* by choosing different levels of effort, the could *pool* by sending the same level of effort, or they could do something *semi-separating* in-between.

Let's think about separation. In a separating equilibrium type 1 chooses some effort level $e_1$ and type 2 chooses some effort level $e_2$. For $e_1$ and $e_2$ to be possible in equilibrium, they must satisfy the following constraints

$$\begin{aligned}
f(e_1, 1) - c(e_1, 1) &\geq f(e_2, 2) - c(e_2, 1) \\
f(e_2, 2) - c(e_2, 2) &\geq f(e_1, 1) - c(e_1, 2) \\
f(e_1, 1) - c(e_1, 1) &\geq \max_{e \in M} f(e, 1) - c(e, 1) \\
f(e_2, 2) - c(e_2, 2) &\geq \max_{e \in M} f(e, 1) - c(e, 2)
\end{aligned}$$

The first two constraints are IC constraints. The second two may seem a bit strange. These reflect that the sender can always choose an off-path level of effort. If they do, then the receiver still has to form some belief about their type, since we are considering PBEs. The worst belief the receiver can form is that they are type 1, so the benefits from deviation to an off path action in deviation can always be set to the right hand side of those inequalities, and this is the maximal possible punishment we can make for a deviation.

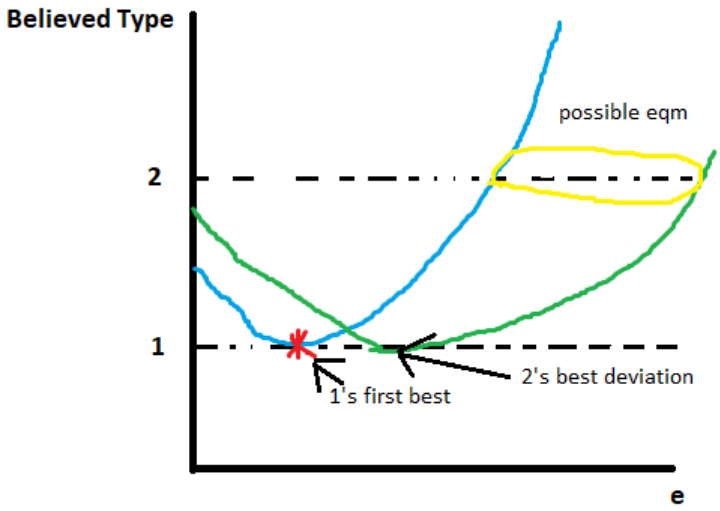
Because of these off-path actions, the lowest type in any separating equilibrium receives the first best level of effort, since $e_1$ must satisfy

$$f(e_1, 1) - c(e_1, 1) \geq \max_{e \in M} f(e, 1) - c(e, 1).$$

Moreover, as the worst that can happen is to the type 2 agent is that they are perceived as type 1, their IC constraint is subsumed by the constraint that they have no off-path deviation. We're left with the following two constraints

$$\begin{aligned}
f(e_1, 1) - c(e_1, 1) &\geq f(e_2, 2) - c(e_2, 1) \\
f(e_2, 2) - c(e_2, 2) &\geq \max_{e \in M} f(e, 1) - c(e, 2)
\end{aligned}$$

where $e_1$ solves $\max f(e_1, 1) - c(e_1, 1)$. If I draw these, they look something like




Because of single crossing, 1's indifference are always "steeper" than 2s. This allows us to conclude that a separating equilibrium exists. The lowest $e_2$ solves

$$f(e_1, 1) - c(e_1, 1) = f(e_2, 2) - c(e_2, 1)$$

and the highest solves

$$f(e_2, 2) - c(e_2, 2) \geq \max_{e \in M} f(e, 1) - c(e, 2)$$

**Example 3.** *Suppose that $f(e, \theta) = \theta$, and $c(e, \theta) = e^2/\theta$. Then $e_1 = 0$ and the set possible $e_2$'s in a separating equilibrium solve*

$$1 \geq 2 - e_2^2$$
$$2 - \frac{1}{2}e_2^2 \geq 1$$

*So $e_2 \in [1, \sqrt{3}]$. We can see that $e_1 = 0, e_2 = 1$ is closest to the first best, while $e_1 = 1, e_2 = \sqrt{3}$ asks for our high type to put in a lot of useless effort (the rMSC?).*

We are going to have a lot of separating equilibrium in this game. The separating equilibrium with the smallest distortion is called the *Riley Outcome*, and it seems like a pretty appealing outcome. We're also going to have a lot of pooling equilibrium, as any level of effort that satisfies

$$\mu_0 f(e, 2) + (1 - \mu_0) f(e, 1) - c(e_1, 1) \geq \max_{e \in M} f(e, 1) - c(e, 1)$$
$$\mu_0 f(e, 2) + (1 - \mu_0) f(e, 1) - c(e_1, 2) \geq \max_{e \in M} f(e, 1) - c(e, 2)$$

is a pooling equilibrium.

**Example 4.** *Consider the previous setup. If $\mu_0 = 1/2$, the pooling equilibrium is any $e$ that satisfies*

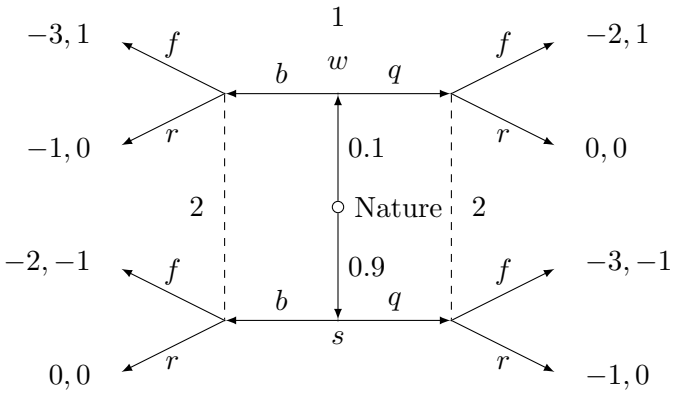
$$3/2 - e^2 \geq 1$$
$$3/2 - e^2/2 \geq 1$$

*so any $e \in [0, 1]$ works. We can see that this set shrinks as the prior becomes smaller, and grows as it becomes bigger. We can see that the $e = 0$ pooling equilibrium makes both types better off than any seperating equilibrium if $\mu_0 > 1/2$.*

There are also, unsurprisingly a bunch of semi-separating equilibrium, but I'm not going to waste your time characterizing those.

This multiplicity of equilibria has been a frequent source of annoyance in this literature. We can see that some separating and pooling equilibria are relatively mild in terms of the distortion induced by asymmetric information, and thus have relatively efficient communication. Others are quite wasteful. We have separating equilibrium where the high types works way more than the first best, pooling equilibrium where no information is transmitted and yet inefficient actions are taken. At the same time, off path beliefs here seem a little funny. The receiver is drawing pretty extreme inferences from off-path messages. In the example, it seems a bit strange to draw the inference that you must be the low type when, for instance, you choose a level of effort that costs more than you could possibly hope to gain from any interpretation the firm could have.

Let's think about this more carefully. Before we analyze it here, let's try thinking about some simpler games. Consider the following game, called the beer-quiche game.[26]




[26]This is a reference to a once popular humor book from before I was born.

The story of this game is as follows. The sender has decided to eat breakfast at a restaurant. Unfortunately, they live in a pretty rough town and have decided to eat at a pretty bad restaurant. The restaurant has two possible breakfasts: Beer and Quiche. A strong sender (type s) prefers beer to quiche, a weak sender (type w) prefers quiche to beer. The sender is worried about the receiver, who is sitting at the next table and is deciding whether to fight them or to retreat (f or r).[27] They'll win any fight against the weak type and lose against the strong type. Neither sender wants to get into a fight What should the sender eat for breakfast?

If we look at the game tree, pooling on beer and pooling on quiche are both consistent with some equilibrium. Since it's pretty likely that the sender is the strong type ex-ante, on-path the receiver gets -.8 from fighting and 0 from retreating. Off path, any belief that induces the receiver to fight prevents deviations (e.g. placing probability 1 on the weak type).

The outcome where they pool on quiche perhaps seems more implausible. Here the sender receives $-1$ if strong and 0 if weak. They don't want to deviate, because any deviation is interpreted as weakness, and will give the strong type $-2$ and the weak type $-3$.

Now, suppose the strong type looks at this game and decides to make the following speech:[28] "I'm having beer for breakfast. You should believe I'm the strong type, because in equilibrium the weak type is getting a payoff of 0. They have no reason to deviate and drink beer, since the most they could possibly get from this is $-1$. Therefore I must be the strong type, so you shouldn't fight me."

This leads us to a refinement called the "intuitive criterion," which ask for our equilibrium to be robust to speech like the one outlined in the previous paragraph. More generally,

**Definition 2.** *An equilibrium $(m, x, \mu)$ with sender value $V(\theta)$ fails the intuitive criterion if:*

- *Given beliefs $\mu \in \Delta(\Theta)$, let $BR(\mu)$ be the receivers best response correspondence to that belief.*
- *Let $D(m) = \{\theta : V(\theta) > \max_{\mu \in \Delta(\Theta)} \max_{x \in BR(\mu)} u_s(m, x, \theta)\}$.*
- *An equilibrium fails the intuitive criterion if there exists an $m$ such that*

$$V(\theta) < \min_{\mu \in \Delta(\Theta \setminus D(m))} \min_{x \in BR(\mu)} u_s(m, x, \theta).$$

[27] If some of our Finnish faculty are to be believed, this was what all of Finland used to be like in the good old days.

[28] So maybe this game isn't taking place in Finland.

Note, that while this feels like a restriction on beliefs, this technically only refers to behavior, not beliefs.

Some things might make you uncomfortable about this refinement. In particular, in our example the strong type is saying "look at what the weak type expects to get in equilibrium, you should interpret my deviation in this way based on that payoff." But, equilibrium beliefs already say how that deviation should be interpreted. And the fact that the weak type things deviations will be interpreted according to those beliefs is exactly what keeps them from making the same speech as the strong type, in which case perhaps the original equilibrium does make sense. There's something sort of circular here. Nevertheless, this concept, and it's close cousins D1 and D2 are probably the most commonly used refinements for signaling games.

Let's return to our job market signaling game. In any separating equilibrium, type 1 will be dominated for any level of effort to the right of the $e^*$ that solves

$$f(e^*, 2) - c(e^*, 1) = f(e_1, 1) - c(e_1, 1)$$

(i.e. $1 \in D(e)$ for $e > e^*$). In order to pass the intuitive criterion, in equilibrium $V(2) \geq \max_{e \geq e^*} f(e, 2) - c(e, 2)$. This implies that any equilibrium other than the Riley outcome fails the intuitive criterion, while the Riley outcome passes it. We can do a similar exercise for pooling equilibrium. Consider a pooling equilibrium with effort level $e^p$.

Let $\bar{e}$ solve

$$\arg\max_e e \text{ s.t. } f(e^p, E(\theta)) - c(e^p, 1) = f(e, 2) - c(e, 1).$$

So $1 \in D(e)$ for all $e > \bar{e}$. And by single crossing

$$f(e^p, E(\theta)) - c(e^p, 1) < f(\bar{e}, 2) - c(\bar{e}, 2)).$$

So pooling fails the intuitive criterion. Semi-separating equilibrium have similar issues.

The intuitive criterion gives a very stark prediction in this example. It always selects the Riley outcome. In general, the Riley outcome will always survive the intuitive criterion, although with more than two types the intuitive criterion no longer rules out all other separating outcomes. These facts led to a relatively large and mostly pointless literature that mostly focuses on finding refinements that select only the Riley outcome. Many, but not all, use essentially the logic of the intuitive criterion. In-spite of this, it is not immediately obvious that it makes sense to try to jump through hoops to find logic that selects the Riley outcome. Some pooling equilibrium, especially when the prior is very concentrated on one type, seem pretty reasonable. If the receiver is

almost certain that the sender is type 2, it seems natural to select the most efficient pooling equilibrium, which both sender types also prefer to the Riley outcome. Some nice refinements don't use intuitive criterion logic and sometimes rule out the Riley outcome, for instance "undefeated equilibrium" which formalizes the idea that off path messages are interpreted by the receiver as indications that the sender is playing a different equilibrium.[29]

### 6.2 Wrap up

This sort of Spence/job market signaling is one of the more impactful models in economics. It provides us with a somewhat subtle, but intuitive story for a lot of behavior that seems on it's face a bit strange. The multiplicity of equilibria and the prominent role of off-path beliefs make this a bit cumbersome to work with, and we'll see quite a few models that make somewhat conceptually odd modeling assumptions to sidestep having to worry about including an embedded signaling game (think about our lemons example from the problem set, for instance). The large literature on refinements seems to have sort of petered out with a shrug, for the most part people make a loose appeal to D1 (a sort of fancier intuitive criterion) is made to justify focusing on the Riley outcome.

As before, a number of assumptions were consequential here. As in everything we've done in this course, single crossing is incredibly important. Without it, it's unclear which constraints matter, and the existence of e.g. a separating equilibrium becomes a tough. It is the assumption we need to place a clear order on types in terms of incentives, and seems natural in these simple settings but may become a bit trickier to formulate if we have e.g. multidimensional types. We've stuck to the analysis with finitely many types (in fact, mostly to the analysis with two types). Having a continuum of types would be helpful here, as we could apply the envelope theorem to solve for a separating equilibrium (and this equilibrium is often unique in the class of separating equilibrium). The Mailath textbook spends a bit of time on this, and discussing this here gives me an excuse to call your attention to the fact that there's a paper called "On the behavior of separating equilibria of signaling games with a finite set of types as the set of types becomes dense in an interval" (JET 1988), which shows that if we add enough types every equilibrium is well approximated by that limiting equilibrium.

I'll conclude this set of notes by discussion two other prominent models of communication. There's an additional model, called "Bayesian Persuasion" that is currently in vogue, but it's much less cool so I'm not doing it here.

[29] This is formalized in Mailath, Okuno-Fujiwara, Postlewaite (1991).

### 6.3 Cheap Talk

What if we don't have a signaling technology? Is there scope for informative communication? Consider the following model.

- $\Theta = X = M = [0, 1]$.
- $u_s = -(x - \theta - b)^2$ and $u_r = -(x - \theta)^2$.
- $\theta \sim U[0, 1]$.

The sender wants $x = \min(1, \theta + b)$, receiver wants $x = \theta$. Clearly if $b = 0$, an equilibrium of this game is for the sender to report the state, and the receiver to play their jointly optimal action. There are other equilibria. For instance, an equilibria where the sender draws a message from $[0, 1]$ uniformly at random and the receiver plays $1/2$ after every message is an equilibrium. We call the latter a babbling equilbrium, no information is transmitted because the sender sends nothing informative.[30]

Now suppose $b > 0$. Informative communication is no longer possible, as the sender would always be willing to deviate to send the message that induces their most preferred action. Babbling on the other hand is still an equilibrium. Is there any equilibrium where any information is transmitted?

Suppose there's an equilibrium where two messages $m_1$ and $m_2$ that induce different receiver actions are sent. Let $x(m_1)$ and $x(m_2)$ the the corresponding receiver actions, WLOG assume $x(m_1) < x(m_2)$. What would the sender's strategy $\sigma : [0, 1] \to \{m_1, m_2\}$ look like? The following lemma hods for any pair of messages in any equilibrium.

**Lemma 4.** *If type $\theta$ prefers message $m_2$ to $m_1$ where $x(m_2) > x(m_1)$, then all higher types strictly prefer $m_2$ to $m_1$.*

*Proof.* This is a consequence of single crossing again. Suppose $\theta$ prefers $m_2$ to $m_1$. Then

$$-(x(m_2) - \theta - b)^2 + (x(m_1) - \theta - b)^2 \geq 0$$

Since $\partial^2 u / \partial x \partial \theta > 0$, this satisfies increasing differences. So for $\theta' > \theta$

$$-(x(m_2) - \theta' - b)^2 + (x(m_1) - \theta' - b)^2 > -(x(m_2) - \theta - b)^2 + (x(m_1) - \theta - b)^2 \geq 0.$$

□

This means that any equilibrium is monotone, higher types send messages that induce higher actions. So, an equilibrium with two message divides the type space into two intervals.

[30] A similar equilibrium without mixing can be constructed by just having the sender send a constant message and the receiver play $1/2$ after every message.

Since we have this monotonicity, the cutoff type must be indifferent between the two messages. So at the cutoff

$$-(x(m_2) - \theta - b)^2 = -(x(m_1) - \theta - b)^2$$

and $x(m_2) = (1 + \theta)/2$, $x(m_1) = \theta/2$. So, putting this all together we get (be careful with positive and negative roots)

$$\begin{aligned} -(x(m_2) - \theta - b)^2 &= -(x(m_1) - \theta - b)^2 \\ -(1/2 - \theta/2 - b)^2 &= -(-\theta/2 - b)^2 \\ 1/2 - \theta/2 - b &= \theta/2 + b \\ \theta &= 1/2 - 2b \end{aligned}$$

so for $b$ between 0 and 1/4th, such an equilibrium exists. In this equilibrium, the sender tells the receiver whether the state is "high" or "low" and the receiver responds optimally to that message. Since the sender is biased up relative to the receiver, it is in some sense easier for them to convey precise information about very low states, while more high states are being pooled together to make misreporting "high" less appealing.

We can add more cutoffs and repeat this procedure. Suppose there are $k$ messages $m_1, m_2 \ldots m_k$ and $0 = \theta_0 < theta_1 < \theta_2 < \ldots \theta_k = 1$ corresponding cutoffs. The logic from above tells us that the cutoffs must solve

$$-\left(\frac{\theta_{i-1} + \theta_i}{2} - \theta_i - b\right)^2 = -\left(\frac{\theta_i + \theta_{i+1}}{2} - \theta_i - b\right)^2$$

so

$$\theta_{i+1} - \theta_i = \theta_i - \theta_{i-1} + 4b$$

So we have a 2nd order difference equation with an initial and terminal condition. Working forward from $\theta_0 = 0$ we get that for this to be an equilibrium

$$\begin{aligned} 1 &= k\theta_1 + \sum_{i=0}^{k-1} 4bi \\ 1 &= k\theta_1 + 4b\frac{k}{2}(k-1) \\ \theta_1 &= \frac{1}{k} - 2b(k-1) \end{aligned}$$

So the largest $k$ will be the largest integer that lies strictly below the positive root of

$$0 = 1/k - 2b(k-1).$$

Which means that the largest possible $k$ as a function of the bias is

$$\left\lfloor \frac{1+\sqrt{\frac{2}{b}+1}}{2} \right\rfloor$$

where the floors make sure that $k$ is actually an integer.

A lot of stuff sort of matches our intuition here. Communication is impossible with large biases, and communication is easier with more aligned preferences (smaller $b$), in the sense that more messages can be sent in the most informative equilibrium. The intervals are getting longer for higher messages, which we need to discourage lying up. Communication is always coarse. Arbitrarily small conflicts of interest ($b$ close to 0) still really limit how much information can be transmitted. We did this here for uniform types and quadratic utility, but there's nothing particularly special here. All we really needed was single crossing to get this partition structure, this simply allowed for a nice closed form solution.

There's a massive literature on cheap talk, and perhaps an even larger literature on cheap talk where the sender can commit to a strategy. I don't really want to talk about these, so I won't.

## 6.4 Disclosure

Suppose we have a seller selling a single item to multiple buyers who compete ala Bertrand for the good.

$$u_s(x,\theta) = x$$

$$u_r(x,\theta) = \theta - x$$

The sender always gets paid the receiver's expectation about $\theta$. Now the sender can disclose "hard evidence",

$$m \in \{\theta, \emptyset\}$$

For simplicity, let's assume a finite type space $\theta \in \{\theta_1, \theta_2, \ldots \theta_N\}$, $\theta_1 < \theta_2 < \ldots \theta_N$.

What does an equilibrium in this game look like? You might expect that maybe the high types reveal their type, and the low types pool on saying nothing. And I suppose that's technically true.

**Theorem 12.** *All information is disclosed in equilibrium.*

*Proof.* Suppose some type reports $\emptyset$ with positive probability. We'll show that all information is disclosed in this equilibrium. We proceed by induction. Clearly $\theta_N$ reveals their type.[31] Now suppose that all types above $\theta_n$ reveal their types, $n > 1$. If $\theta_n$ reports

[31] If they were the only type to not reveal, all other types would have incentive to deviate. If other

$\emptyset$, then they receive $E(\theta|\emptyset) \leq \theta_{n-1} Pr(\theta < \theta_n|\emptyset) + \theta_n Pr(\theta = \theta_n|emptyset)$. If they report truthfully they receive $\theta_n$. So disclosure is optimal. Therefore, all types other than $\theta_1$ find it optimal to reveal. Since only $\theta_1$ doesn't disclose, they also reveal effectively reveal their type. So, all types $\theta_n$ $n > 1$ discloses, and $\theta_1$ either discloses or doesn't (or randomizes between the two) are the set of possible equilibria. It's easy to verify that these are all equilibria. □

This is a very stark result. Hard evidence, and the knowledge that the sender has this evidence, unravels all private information. Even in games where no types want to reveal their private information, one can imagine how this logic would lead to equilibrium where all information is revealed, for instance to a monopolist who can then perfectly price discriminate. At the same time, being able to perfectly prove your type perhaps seems a bit too strong. What if we changed the evidence structure so that there was some uncertainty that the sender has access to evidence?

Now suppose that with probability $q$, the sender cannot prove their type, and let $Pr(\theta_i) = p_i$. Now the sender with evidence reveals if and only if

$$\theta_i > E(\theta|\emptyset).$$

Clearly, the equilibrium is going to have a cutoff structure. Can we support no revelation for some types?

Let $k$ be the cutoff type, this type and all types above it reveal if they have evidence and all lower types don't reveal. We know at least the highest type must reveal. Now, in equilibrium,

$$E(\theta|\emptyset) = \frac{\sum_{i=1}^{k-1} p_i\theta_i + \sum_{j=k}^{N} qp_i\theta_j}{\sum_{i=1}^{k-1} p_i + \sum_{j=k}^{N} qp_i}.$$

This is a convex combination of the types, with strictly positive weight on every type. So, the lowest type never reveals and the highest type always does. We have an interior cutoff, and the equilibrium seems to match our intuition.

Relative to the case with no asymmetric information, we can see that high type sellers are doing "worse" in expectation, while low type sellers are better off. This model leads to the natural question, what if the seller could choose whether or not to acquire evidence.

For simplicity, suppose that there are two types $\theta_1$ and $\theta_2$, which are unknown to both the buyer and the seller ex-ante. The seller can pay a cost $c(1-\varepsilon)$ to learn their type and simultaneously produce hard evidence they can choose to reveal to buyers, where $\varepsilon > 0$ is also the probability that info acquisition fails (take $\varepsilon$ small), so that if

types don't reveal, then $E(\theta|\emptyset) < \theta_N$.

both acquire information for sure the off-path beliefs are still pinned down.[32]

Let $P$ be the equilibrium price following no revelation. We know that $\theta_1$ is never going to reveal, and $\theta_2$ will always reveal after acquiring information. Then we know that information is acquired when

$$p_2\theta_2 + (1-p_2)P - c \geq P.$$

Using Bayes rule, if the sender acquires information with probability $\gamma \in [0, 1-\varepsilon]$ (omitting $\varepsilon$s by collapsing them into $\gamma$) then

$$P = \frac{(1-\gamma)p_2\theta_2 + (1-p_2)\theta_1}{(1-\gamma)p_2 + (1-p_2)}$$

As this is decreasing in $\gamma$, there is always going to be at most one equilibrium. Solving for where this intersects the equilibrium condition, after some algebra we get

$$1 - \gamma = \frac{(1-p_2)(p_2(\theta_2 - \theta_1) - c)}{cp_2}$$

(or 0 or 1 if this is either less than $\varepsilon$ or larger than 1). So the larger the gap between types, relative to costs, the less likely they are to acquire information.

For a final exercise, how would this equilibrium change if the government mandated disclosure of any information acquired. Then the equilibrium condition becomes

$$p_2\theta_2 + (1-p_2)\theta_1 - c \geq P,$$

and $P$ must be $p_2\theta_2 + (1-p_2)\theta_1$, since the agent must be uninformed if they don't reveal anything. Then there is no incentive to acquire information, so this disclosure policy is counterproductive. If we'd like this information to be revealed, this disclosure mandate should be accompanied with a mandate to acquire information. A requirement to disclose any problems you find vs. a requirement to run these specific tests + disclose the results have very different impacts.

[32] Otherwise there's an additional PBE where both types acquire and reveal because no revelation is interpreted as proof that the sender is the lowest type. The analysis is identical modulo this. Sequential equilibrium isn't strong enough to rule this out, but the trembles that generate the appropriate beliefs require the high type to tremble to a weakly dominated strategy relatively frequently, which is maybe a bit strange.