Some Notes on the Revelation Principle

John Morgan
Haas School of Business and Department of Economics
University of California, Berkeley

1 Direct and Indirect Mechanisms

What is a contract?

On a general level, a contract is a game designed by the principal and played by the agent or agents. There may be various institutional restrictions placed on the set of feasible game forms the principal may choose from.

Thus, a contract can be thought of as a game Γ that designates the strategies available to the agents as well as the payoffs from the realizations of the strategies. We call such a contract an *indirect mechanism.* Suppose agent i's type is θ_i and her equilibrium strategy in game Γ is $\psi_i(\theta_i)$. Let $\psi(\theta)$ denote the vector of strategies chosen by all of the agents given their type.

So we have

$$\begin{array}{ccc} & & \theta \\ & & \searrow \\ v_i(\psi(\theta)) & \leftarrow & \psi(\theta) \end{array}$$

The mechanism is *indirect* since ultimately the vector of types θ determines the payoffs to each of the agents; however this information is transmitted indirectly through the strategy $\psi(\cdot)$.

In contrast, the principal might elect to use a *direct mechanism.* In a direct mechanism:

- Agents report their types to an neutral intermediary
- The intermediary then implements the strategy $\psi_i(\cdot)$ on behalf of the agent
- And the agent earns $v_i(\psi(\theta))$.

Thus, the direct mechanism looks like:

$$\begin{array}{cc} & \theta \\ \swarrow & \\ v(\psi(\theta)) & \psi(\theta) \end{array}$$

The Revelation Principle Suppose we use the solution concept of Bayes–Nash equilibrium then it must be the case that for agent i of type θ_i,

$$E_{\theta_{-i}} v_i \left(\psi_i \left(\theta_i \right), \psi_{-i} \left(\theta_{-i} \right) \right) \geq E_{\theta_{-i}} v_i \left(\psi_i', \psi_{-i} \left(\theta_{-i} \right) \right)$$

for all ψ_i' in the indirect mechanism. That is, no agent should have a profitable deviation.

We might also require that the agent have an incentive to participate in the game in the first place. That is

$$E_{\theta_{-i}} v_i \left(\psi \left(\theta \right) \right) \geq \bar{v}$$

for some outside option $\bar{v}$.

The point of the direct mechanism is that, through the revelation principle, it allows us to ignore analyzing all of the indirect mechanisms and focus only on the direct mechanism. Specifically,

Theorem 1 (Revelation Principle, Myerson, 1981) *Suppose that ψ was a Bayes-Nash equilibrium of the indirect mechanism Γ. Then there exists a direct mechanism that is payoff-equivalent and where truthful revelation is an equilibrium.*

Sketch Proof. By construction, if agents tell the truth in the direct mechanism, it is payoff equivalent to playing the ψ equilibrium strategies in the indirect mechanism Γ. To see that truth-telling is an equilibrium, notice that if player i with type θ_i deviates and reports his type as θ_i' then she earns $E_{\theta_{-i}} v_i \left(\psi_i \left(\theta_i' \right), \psi_{-i} \left(\theta_{-i} \right) \right) = E_{\theta_{-i}} v_i \left(\psi_i', \psi_{-i} \left(\theta_{-i} \right) \right)$ for some ψ_i' and, from above, we know

$$E_{\theta_{-i}} v_i \left(\psi_i \left(\theta_i \right), \psi_{-i} \left(\theta_{-i} \right) \right) \geq E_{\theta_{-i}} v_i \left(\psi_i', \psi_{-i} \left(\theta_{-i} \right) \right)$$

Therefore, this is not profitable. QED

2 Putting the Revelation Principle to Work

This set of lecture notes covers a general model of adverse selection as well as a leading example—that of a price discriminating monopolist—due to Maskin and Riley (1984). At the end of this unit, students should understand the application of the revelation principal to this class of models, the general solution techniques, and the key economic trade-off—efficiency versus information rent savings.

2.1 Two Type Model of Monopoly Price Discrimination

We begin with the problem of a price discriminating monopolist.

A Seller (P) can produce a quantity q of a good at a constant MC c and no fixed costs. He seeks to contract with a buyer (A) who has preferences

$$V \left(q, T, \theta_i \right) = \theta_i v \left(q \right) - t$$

where t is the transfer paid from the buyer to the seller and θ_i is the buyer's "type" which may be thought of as a parameter affecting the buyer's willingness to pay. The function $v(q)$ is the value function of the buyer depending on the quantity of the good purchased. We assume that $v' > 0$, $v'' < 0$ and that things are such that interior solutions always apply.

When the seller sells q units to the buyer and receives a transfer t, he earns:

$$\pi = t - cq.$$

Suppose that buyers come in two type: those with high willingness to pay and those with low willingness to pay. Thus, the set of states is $\{\theta_h, \theta_l\}$ where θ_l occurs with probability λ.

The extensive form of the game is as follows: P offers A a contract (or a menu of contracts) on a take it or leave it basis. If A rejects, both sides earn zero. If A accepts, the contract is executed.

First-best benchmark

To being with, suppose that P knew A's type. What contracts should he propose?

If the agent's type is θ_i then P's problem is:

Solve

$$\max_{t_i, q_i} t_i - cq_i$$

subject to

$$\theta_i v(q_i) - t_i \geq 0 \tag{1}$$

Here, the constraint on the problem, equation (1), reflects the fact the A must prefer the contract to the outside option. Notice that there's little point for the seller to leave any surplus to the buyer, hence equation (1) holds with equality. Thus, we know that

$$t_i = \theta_i v(q_i)$$

and the problem becomes a univariate problem:

$$\max_{q_i} \theta_i v(q_i) - cq_i$$

Which yields

$$\theta_i v'(q_i^*) = c \tag{2}$$

The economic interpretation of equation (2) is that P chooses a quantity where the marginal benefit to the buyer equals the seller's marginal cost of production. That is, the socially efficient quantity of the good is offered. This maximizes the size of the economic pie to be split between the two parties. Having done this, the seller then uses the transfer to capture all of this surplus.

Adverse selection

Interest in the problem then stems from the fact that the seller does not know the buyer's willingness to pay but must instead use contracting to figure it out. In

general, P selects a message space M and then specifies quantities and transfers as a function of the message selected by A. From the **revelation principle**, we know that it is without loss of generality to (i) restrict the message space to be the type space; and (ii) restrict attention to truth-telling equilibria.

Thus, P's problem is to specify a menu of contracts $\left(t\left(\hat{\theta}\right), q\left(\hat{\theta}\right)\right)$ as a function of A's report, $\hat{\theta}$, about his type. Since we can restrict attention to truth-telling, then such a menu must satisfy

$$\theta_i v\left(q\left(\hat{\theta}=\theta_i\right)\right)-t\left(\hat{\theta}=\theta_i\right)\geq 0 \tag{3}$$

for $i=h,l$. This is simply the analog of equation (1) above for the case of menus of contracts. In words, equation (3) says that an agent must prefer to report his type truthfully than to reject the contract entirely.

Next, for truth-telling to be an equilibrium, it must be the case that an agent of type θ_i prefers to reveal his type truthfully rather than to pretend to be some other type θ_j. Thus, truth-telling must be **incentive compatible.** That is

$$\theta_i v\left(q\left(\hat{\theta}=\theta_i\right)\right)-t\left(\hat{\theta}=\theta_i\right)\geq \theta_i v\left(q\left(\hat{\theta}=\theta_j\right)\right)-t\left(\hat{\theta}=\theta_j\right) \tag{4}$$

The monopolist's problem is to maximize its expected profits subject to equations (3) and (4) . That is

$$\max_{t(\theta),q(\theta)} \lambda\left(t\left(\theta_l\right)-cq\left(\theta_l\right)\right)+\left(1-\lambda\right)\left(t\left(\theta_h\right)-cq\left(\theta_h\right)\right)$$

subject to equations (3) and (4) .

It helps to break out the four constraints separately:

$$\theta_h v\left(q_h\right)-t_h\geq \theta_h v\left(q_l\right)-t_l \tag{ICH}$$

$$\theta_l v\left(q_l\right)-t_l\geq \theta_l v\left(q_h\right)-t_h \tag{ICL}$$

$$\theta_h v\left(q_h\right)-t_h\geq 0 \tag{IRH}$$

$$\theta_l v\left(q_l\right)-t_l\geq 0 \tag{IRL}$$

We'll use some economic intuition to figure out which constraints are actually binding. In the general model, we'll use some math to figure out which constraints matter.

To gain intuition, suppose that P offers the first best contract. Clearly, this satisfies all the IR constraints. Notice, however, that a high willingness to pay buyer would rather switch to the contract intended for the low type. Why is this? Consuming the quantity intended for the low type, yields strictly more benefit for the high type than for the low type. That is

$$\theta_h v\left(q\left(\theta_l\right)\right)>\theta_l v\left(q\left(\theta_l\right)\right).$$

The transfer, however, is designed to capture the surplus of the low type

$$t(\theta_l) = \theta_l v(q(\theta_l)).$$

Hence, by pretending to be a low type, a high type obtains positive surplus. In contrast, telling the truth earns the high type zero surplus. Therefore, the incentive compatibility constraint must be binding for high types pretending to be low types.

Is the same thing true the other way around? Would low types want to pretend to be high types under the first best contract? The same reasoning shows that low types would not want to do this. They'll earn negative surplus by pretending to be high types as compared to zero surplus from telling the truth.

What about individual rationality? Recall that whatever surplus low types are left with, high types must obtain strictly more surplus since they can always pretend to be low types and earn more than what the low types earn. Thus, for each dollar of surplus we leave for low types, we have to give at least this much to high types too. Moreover, by simply scaling the transfers 1 for 1 between the two types, we do not affect incentive compatibility (since IC is only affect by the difference in the levels of the transfers). Hence, we'll leave the low types with zero surplus.

A different way of getting to this: Since low types cannot credibly pretend to be high types, they have no bargaining power vis a vis the seller. Hence, their in the same bargaining position as the first-best problem and the seller captures all the surplus from them.

Thus, we now know that

$$\theta_l v(q_l) - t_l = 0$$

and

$$\theta_h v(q_h) - t_h = \theta_h v(q_l) - t_l$$

Hence

$$t_l = \theta_l v(q_l)$$

and

$$t_h = \theta_h v(q_h) - (\theta_h - \theta_l) v(q_l)$$

Substituting, we now have the dead-simple problem

$$\max_{q_i} \lambda(\theta_l v(q_l) - cq_l) + (1-\lambda)(\theta_h v(q_h) - (\theta_h - \theta_l) v(q_l) - cq_h)$$

Differentiating wrt q_h yields

$$(1-\lambda)(\theta_h v'(q_h) - c) = 0 \tag{5}$$

and notice that this is our old friend the first-best solution.

Differentiating with respect to q_l yields

$$\lambda(\theta_l v'(q_l) - c) - (1-\lambda)(\theta_h - \theta_l) v'(q_l) = 0$$

which is nicer to rewrite as

$$\theta_l v'(q_l) = \frac{c}{1 - \left(\frac{1-\lambda}{\lambda}\frac{\theta_h - \theta_l}{\theta_l}\right)} \tag{6}$$

Notice that this implies that the low guys get less than the socially optimal quantity.

Equations (5) and (6) reveal the main insight of adverse selection: To cope with the problem of private information, P uses both margins available to him. He chooses a contract that reduces the overall size of the economic pie by distorting the output of low types and leaves surplus on the table for high types. Why does he do this? The key trade-off is the cost of the distortion versus the size of the information rents that must be paid to high types. To see that distortion makes sense, consider a contract that gives all types the socially efficient quantities. In such a contract, high types are paid considerable information rents. Now consider a small distortion downward in the quantity offered to low types. Since low types are already at the social optimum, the reduction in surplus from such a distortion is first-order equal to zero. However, when high types try to imitate low types, they are away from the social optimum, thus the effect on the welfare on high types who imitate low types is first-order negative. Hence, P can make a first-order reduction in the information rent paid to high types and this increases P's profits.

2.2 General Model

Having now looked at a specific model and gotten some economic intuition for the key trade-off in adverse selection, we now turn to a reasonably general model to provide the mechanics for solving these types of problems.

Let $q \in \Re$ denote an economically relevant action (i.e. output level, quantity sold, etc.) and $t \in \Re$ denote a money transfer. Let Θ denote the space of types, which we shall assume to be univariate and continuous with positive density everywhere.

Suppose that P's objective function is $W(q,t)$. That is, the principal does not care about the agent's type directly. Suppose that A's objective function is $U(q,\theta,t) \equiv u(q,\theta) - t$. That is, the agent's preferences are quasi-linear in money. Finally, let $V\left(\hat{\theta},\theta\right) \equiv u\left(q\left(\hat{\theta}\right),\theta\right) - t\left(\hat{\theta}\right)$ denote a type θ agent's utility when it reports type $\hat{\theta}$.Throughout we'll assume everything is differentiable and nicely behaved.

From the revelation principle, we can restrict attention to direct mechanisms and truth-telling. Thus, for q,t to be incentive compatible, we require the first and second order conditions:

$$\begin{aligned} \frac{\partial V(\theta,\theta)}{\partial \hat{\theta}} &= 0 \\ \frac{\partial^2 V(\theta,\theta)}{\partial \hat{\theta}^2} &\leq 0 \end{aligned}$$

Substituting for V, we can rewrite the first order condition as

$$\frac{\partial u\left(q\left(\theta\right),\theta\right)}{\partial q}\frac{\partial q}{\partial \theta}-\frac{\partial t\left(\theta\right)}{\partial \theta}=0$$

or

$$\frac{\partial t\left(\theta\right)}{\partial \theta}=\frac{\partial u\left(q\left(\theta\right),\theta\right)}{\partial q}\frac{\partial q\left(\theta\right)}{\partial \theta} \tag{7}$$

Similarly, we can rewrite the second order condition as

$$\frac{\partial^{2}t\left(\theta\right)}{\partial \theta^{2}}\geq\frac{\partial^{2}u\left(q\left(\theta\right),\theta\right)}{\partial q^{2}}\left(\frac{\partial q\left(\theta\right)}{\partial \theta}\right)^{2}+\frac{\partial u\left(q\left(\theta\right),\theta\right)}{\partial q}\frac{\partial^{2}q\left(\theta\right)}{\partial \theta^{2}} \tag{8}$$

Now, differentiate (7) with respect to θ to obtain

$$\frac{\partial^{2}t\left(\theta\right)}{\partial \theta^{2}}=\frac{\partial^{2}u\left(q\left(\theta\right),\theta\right)}{\partial q^{2}}\left(\frac{\partial q\left(\theta\right)}{\partial \theta}\right)^{2}+\frac{\partial u\left(q\left(\theta\right),\theta\right)}{\partial q}\frac{\partial^{2}q\left(\theta\right)}{\partial \theta^{2}}+\frac{\partial^{2}u\left(q\left(\theta\right),\theta\right)}{\partial q\partial \theta}\frac{\partial q\left(\theta\right)}{\partial \theta}$$

which we can then substitute into the inequality in equation (8) to obtain

$$\frac{\partial^{2}u\left(q\left(\theta\right),\theta\right)}{\partial q\partial \theta}\frac{\partial q\left(\theta\right)}{\partial \theta}\geq0 \tag{9}$$

Thus, incentive compatibility requires

$$\begin{aligned}\frac{\partial t\left(\theta\right)}{\partial \theta} &= \frac{\partial u\left(q\left(\theta\right),\theta\right)}{\partial q}\frac{\partial q\left(\theta\right)}{\partial \theta}\\ \frac{\partial^{2}u\left(q\left(\theta\right),\theta\right)}{\partial q\partial \theta}\frac{\partial q\left(\theta\right)}{\partial \theta} &\geq 0\end{aligned}$$

What do we know about equation (9)? What does it even mean?

We certainly don't know that equation (9) holds, but most models assume the following

$$\frac{\partial^{2}u\left(q\left(\theta\right),\theta\right)}{\partial q\partial \theta}>0$$

which is called the **Spence-Mirrlees single-crossing property**. One can generalize this outside of the continuum framework and it says that the function u is **supermodular** in its arguments. Economically, it means that the marginal utility of a given unit of q is higher for higher types. Notice that the buyers in our monopoly example above satisfied this condition.

Incentive-Feasibility

We are now in a position to write down something general about incentive-feasible (or implementable) contracts.

Lemma 1 *Suppose u satisfies the single crossing property. Then the contract (q,t) is incentive-feasible if and only if q is non-decreasing.*

Proof: Recall that incentive feasible contracts must satisfy equations (7) and (9). It's obvious that if (q,t) is incentive-feasible, then q is non-decreasing (otherwise equation (9) will fail somewhere). We proceed to show that *any* non-decreasing q is incentive-feasible.

First, differentiate $V\left(\hat{\theta},\theta\right)$ with respect to $\hat{\theta}$.

$$\frac{\partial V\left(\hat{\theta},\theta\right)}{\partial\hat{\theta}}=\frac{\partial u\left(q\left(\hat{\theta}\right),\theta\right)}{\partial q}\frac{\partial q\left(\hat{\theta}\right)}{\partial\theta}-\frac{\partial t\left(\hat{\theta}\right)}{\partial\theta}$$

Now substituting (7) for $\frac{\partial V\left(\hat{\theta},\hat{\theta}\right)}{\partial\hat{\theta}}$ yields

$$\begin{aligned}&\frac{\partial u\left(q\left(\hat{\theta}\right),\theta\right)}{\partial q}\frac{\partial q\left(\hat{\theta}\right)}{\partial\theta}-\frac{\partial u\left(q\left(\hat{\theta}\right),\hat{\theta}\right)}{\partial q}\frac{\partial q\left(\hat{\theta}\right)}{\partial\theta}\\ =\;&\frac{\partial q\left(\hat{\theta}\right)}{\partial\theta}\left(\frac{\partial u\left(q\left(\hat{\theta}\right),\theta\right)}{\partial q}-\frac{\partial u\left(q\left(\hat{\theta}\right),\hat{\theta}\right)}{\partial q}\right)\end{aligned}$$

By the mean value theorem, there exists θ^* between θ and $\hat{\theta}$ such that $\frac{\partial V\left(\hat{\theta},\theta\right)}{\partial\hat{\theta}}$ has the same sign as

$$\frac{\partial q\left(\hat{\theta}\right)}{\partial\theta}\frac{\partial^2 u\left(q\left(\hat{\theta}\right),\theta^*\right)}{\partial q\partial\theta}\left(\theta-\hat{\theta}\right)$$

Now if q is non-decreasing, this implies that $V\left(\hat{\theta},\theta\right)$ has a global maximum at $\hat{\theta}=\theta$ and hence is incentive-feasible.

Optimal Contract

We'll now solve for the optimal contract. We'll make life easier for ourselves by assuming that P's objective is also quasi-linear in money: $W(q,t)=C(q)-t$.

We'll make the single-crossing assumption as well as one additional assumption on u

Assumption:

$$\frac{\partial u(q,\theta)}{\partial\theta}>0$$

This means that higher types enjoy strictly higher utility from the same q. We're doing this to ensure we'll only have to worry about individual rationality for one type of agent. To see this, define

$$v(\theta)=u\left(q(\theta),\theta\right)-t(\theta)$$

under an incentive-feasible contract (q,t). From the envelope theorem, we know

$$\frac{\partial v(\theta)}{\partial\theta}=\frac{\partial u\left(q(\theta),\theta\right)}{\partial\theta}\tag{10}$$

which is positive by assumption. Thus, if we satisfy IR for the lowest type, we satisfy it for all other types since their equilibrium utility under an incentive-feasible contract is strictly higher than for the lowest type. We'll normalize the outside option for the lowest type at zero and hence

$$v(\theta) = \int_{\theta_0}^{\theta} \frac{\partial u(q(t), t)}{\partial \theta} dt$$

We also know that

$$t(\theta) = u(q(\theta), \theta) - v(\theta)$$

Hence

$$t(\theta) = u(q(\theta), \theta) - \int_{\theta_0}^{\theta} \frac{\partial u(q(t), t)}{\partial \theta} dt$$

and we can eliminate the $t(\theta)$ from the optimization.

Thus, P's problem is to choose a non-decreasing function q to maximize

$$\begin{aligned} W(q,t) &= \int_{\theta_0}^{\theta_1} (t(\theta) - C(q(\theta))) f(\theta) d\theta \\ &= \int_{\theta_0}^{\theta_1} \left(u(q(\theta), \theta) - \int_{\theta_0}^{\theta} \frac{\partial u(q(t), t)}{\partial \theta} dt - C(q(\theta)) \right) f(\theta) d\theta \end{aligned}$$

where f is the density of the type space which is assumed everywhere positive.

An aside: What to do about the term $\int_{\theta_0}^{\theta_1} \left(\int_{\theta_0}^{\theta} \frac{\partial u(q(t),t)}{\partial \theta} dt \right) f(\theta) d\theta$? From Fubini's theorem (or just draw a picture) we can reorder the integrals as

$$\begin{aligned} & \int_{\theta_0}^{\theta} \left(\int_{\theta_0}^{\theta_1} f(\theta) d\theta \right) \frac{\partial u(q(t), t)}{\partial \theta} dt \\ = & \int_{\theta_0}^{\theta_1} \left(\int_{\theta}^{\theta_1} f(t) dt \right) \frac{\partial u(q(\theta), \theta)}{\partial \theta} d\theta \\ = & \int_{\theta_0}^{\theta_1} (1 - F(\theta)) \frac{\partial u(q(\theta), \theta)}{\partial \theta} d\theta \end{aligned}$$

Hence, we have

$$W(q,t) = \int_{\theta_0}^{\theta_1} \left\{ [u(q(\theta), \theta) - C(q(\theta))] f(\theta) - \frac{\partial u(q(\theta), \theta)}{\partial \theta} (1 - F(\theta)) \right\} d\theta$$

Now multiply and divide by f to obtain

$$W(q,t) = \int_{\theta_0}^{\theta_1} \left\{ u(q(\theta), \theta) - C(q(\theta)) - \frac{\partial u(q(\theta), \theta)}{\partial \theta} \frac{1}{h(\theta)} \right\} f(\theta) d\theta \qquad (11)$$

where $H(\theta)$ is the **hazard rate** of the type distribution.

What the heck does the expression of P's problem given in equation (11) mean? Notice that the first pair of terms $u(q(\theta),\theta) - C(q(\theta))$ is simply the social surplus from the allocation. The last term $\frac{\partial u(q(\theta),\theta)}{\partial \theta}\frac{1}{h(\theta)}$ represents the distortion associated with the trade-off of social surplus for information rent savings. The whole expression given in $\{\}$ is often termed the **virtual surplus.**

In general, we could use methods of **optimal control** to solve this problem, but since we're lucky, we can instead maximize it pointwise and see how things turn out. Notice however that the pointwise optimization ignores the constraint that q be non-decreasing, so we'll have to go back and check on this later. Maximizing pointwise yields the first-order condition:

$$\frac{\partial u(q,\theta)}{\partial q} - C'(q) - \frac{\partial^2 u(q,\theta)}{\partial q \partial \theta}\frac{1}{h(\theta)} = 0$$

or

$$\frac{\partial u(q,\theta)}{\partial q} - \frac{\partial^2 u(q,\theta)}{\partial q \partial \theta}\frac{1}{h(\theta)} = C'(q)$$

which should have a familiar ring. The term $\frac{\partial u(q,\theta)}{\partial q}$ is the marginal benefit of providing a unit of q. The term $C'(q)$ is the marginal cost. Finally, the term $\frac{\partial^2 u(q,\theta)}{\partial q \partial \theta}\frac{1}{h(\theta)}$ is the distortion associated with adverse selection. Notice that this term is always negative; hence, we see the same distortion as in the monopoly case: too little q is provided relative to the social optimum. Finally, Notice that, at θ_1 the hazard rate is infinite; hence the weight on the distortion term goes to zero—the output distortion on high types is small. To solve things beyond this point requires further structure on the problem.