Contract Theory*

Gabriel Carroll

August 22, 2021

1 Introduction

In many matching markets, the matching itself is not the end of the story; participants make decisions or take actions that in turn affect how valuable the match is. Workers in online labor markets decide how carefully to work or how fast to get the job done. Buyers matched with sellers of goods may be able to choose from among multiple versions of a product and multiple methods of delivery. These parties' behavior depends on the incentives they face, such as rewards to the worker for good performance, or the prices on different options offered to the buyer.

The traditional branch of economic theory dealing with design of incentives is known as *contract theory*, and its elements will be presented in this chapter. As usual, incentive issues are tightly linked with asymmetries of information. Essentially, if all parties were perfectly informed, there would be no incentive problems to solve. These asymmetries are often divided into two kinds: *hidden actions*, where one party does something that is not perfectly observed by the other, and *hidden information*, where one party knows something that affects preferences (as in most mechanism design problems). Both kinds of issues are studied through the *principal-agent framework*, a modeling approach that focuses on interactions between two participants, the *principal* (who designs the incentives) and the *agent* (who has the superior information and responds strategically).

A few words about modeling philosophy: Like most models in economics—and unlike most models in computer science and operations research—those in this chapter are "toy" models; that is, they are gross oversimplifications whose purpose is to develop concepts. Specialized algorithms to calculate optimal incentive schemes are typically of limited interest, because in practical situations one usually cannot describe the environment in enough detail to provide accurate input to the algorithm anyway. Given that the purpose is conceptual, we may

*This chapter will appear in *Online and Matching-Based Market Design*, Federico Echenique, Nicole Immorlica, and Vijay V. Vazirani, editors, © 2021, Cambridge University Press.

as well formulate the models to be simple enough so that the solutions can be found by hand, and then we can examine their salient properties.

2 Hidden-Action Models

In hidden-action models, as the name implies, the agent can choose among several actions. For such models to be interesting, it should be the case that the action the agent would choose in the absence of incentives is different from the one the principal would like chosen; such a situation is also referred to as one of *moral hazard*.

We will first offer a simple version of such a model that illustrates its essential elements. As we shall see, this model also has a simple solution—but arguably too simple for many purposes. We will then explore a couple of variants that address some shortcomings of the basic model and thereby highlight various considerations in incentive design.

It is common to introduce hidden-action models with the interpretation that the agent is a worker, and the principal is the boss, who has to motivate the agent to work. Although we follow this framing here, it is worth keeping in mind that these models have many other applications; Section 2.4 discusses some.

2.1 A simple benchmark model

The agent can take various possible *actions* that create revenue, or *output*, for the principal, but these actions may be costly for himself. We take as given a finite set $Y \subseteq \mathbb{R}_+$ of possible *output levels*. The output produced is random, but its distribution depends on the action taken. We are not concerned with the physical description of the actions the agent can take, but only with their consequences for payoffs, and thus model them as follows:

Definition 1. An *action* is a pair $a = (F, c)$, where $F \in \Delta(Y)$ and $c \in \mathbb{R}_+$.

The interpretation is that the action generates output drawn from distribution F, and it costs c to the agent. As a convention, we write $F(y)$ for the probability of drawing an output level y or less, and $f(y)$ for the probability of drawing exactly y.

Remark 2. Depending on the application, the cost c might be interpreted as money the agent has to spend, or simply as the monetary equivalent of the displeasure the agent experiences from exerting effort to perform the action.

Definition 3. A *technology* is a nonempty, finite set of actions.

An instance of the model is given by a triple $(Y, \mathcal{A}, \underline{u})$, where Y is the set of possible output levels, $\mathcal{A}$ is the technology available to the agent, and $\underline{u} \in \mathbb{R}$ is the agent's *outside option*, the payoff he would receive by declining to transact with the principal.

The principal can give incentives via a *contract*, which specifies a recommended action and a payment for each level of output that might be produced:

Definition 4. A *contract* (a, w) consists of $a \in \mathcal{A}$ and a function $w : Y \to \mathbb{R}$.

The interaction between the parties is envisioned to proceed as follows:

- The principal proposes a contract.
- The agent can either reject the contract and earn his outside option $\underline{u}$, or accept the contract.
- If the agent accepts the contract, he chooses an action (F, c) from his technology. The principal does not observe the action directly, but she does observe the level of output it produces, $y \sim F$.
- The agent is then paid as promised by the contract, $w(y)$, and the principal keeps the rest, $y - w(y)$.

Here, and throughout this chapter, we assume the parties have quasi-linear utility (see page **??**) and evaluate random outcomes by expected utility. So the agent's overall payoff from taking action (F, c), taking the cost into account, is $\mathbb{E}_{y\sim F}[w(y)] - c$, and the principal's corresponding payoff is $\mathbb{E}_{y\sim F}[y - w(y)]$.

Definition 5. Say the contract $((F, c), w)$ is *valid* if it satisfies the following two conditions:

- *(Incentive compatibility)* $\mathbb{E}_{y\sim F}[w(y)]-c \geq \mathbb{E}_{y\sim F'}[w(y)]-c'$ for all $(F', c') \in \mathcal{A}$.
- *(Individual rationality)* $\mathbb{E}_{y\sim F}[w(y)] - c \geq \underline{u}$.

The incentive compatibility condition is also sometimes referred to by saying that w *implements* action (F, c). Individual rationality expresses that the agent should be willing to accept the contract.

Remark 6. Why do we require contracts to include an action recommendation (and write incentive compatibility explicitly)? We could have instead just said that w itself is the contract, and the agent chooses whichever action he prefers. But this approach ensures there is no ambiguity about what the agent does if he has multiple optimal actions.

The principal's problem is to find a valid contract that maximizes her payoff. In this benchmark model, there is an easy solution. Define the *surplus* from an action (F, c) as $s(F, c) = \mathbb{E}_{y\sim F}[y] - c$. Let s^* be the maximum surplus among all actions in $\mathcal{A}$, and $(F^*, c^*) \in \mathcal{A}$ an action attaining it. Then:

Observation 7. *Offering payment $w(y) = y - s^* + \underline{u}$, together with recommending action (F^*, c^*), constitutes an optimal valid contract.*

Proof. To see that the contract is valid: for any action (F, c) the agent takes, his payoff is $s(F, c)$ plus the constant $-s^* + \underline{u}$, so taking the surplus-maximizing action (F^*, c^*) is incentive compatible. The agent's resulting payoff is $\underline{u}$, so individual rationality is satisfied too.

To see that the contract is optimal: notice that it gives the principal a payoff of $s^* - \underline{u}$. There is no way to do better, because for any contract, whatever action (F, c) the agent takes, the sum of the two parties' payoffs equals the surplus $s(F, c) \leq s^*$. Since the agent's payoff needs to be at least $\underline{u}$, the principal gets at most $s^* - \underline{u}$. □

Remark 8. The contract above is sometimes referred to as "selling the firm to the agent" for the price $s^* - \underline{u}$, i.e. it can be interpreted as asking the agent to pay $s^* - \underline{u}$ to the principal, then giving the agent ownership of the output.

In the above analysis, the principal does well by simply making the agent the beneficiary of the fruits of his labor, thus giving him incentives to choose his action efficiently.

Although this idea is intuitive, the resulting contract is often unrealistic: If the range of possible output levels is large (for example, the agent is responsible for a big project), this contract may end up specifying huge positive or negative payments depending on the realized y. There are various reasons why this might not be appropriate; here are two main ones:

- First, the agent might be *risk averse*: he might prefer a less-variable payment, even if its expected value is lower. This is modeled by assuming that the agent maximizes an expression such as $\mathbb{E}_{y\sim F}[u(w(y))] - c$, where $u : \mathbb{R} \to \mathbb{R}$ is some concave function, rather than simply maximizing $\mathbb{E}_{y\sim F}[w(y)] - c$. In this case, the principal can do better by offering the agent some insurance against the uncertainty in output.

- Second, there might be *limited liability*: large negative payments might just be impossible (there's no way to take away more money than the agent owns), or might be possible but illegal.

Risk aversion is widespread in many economic applications. However, we will turn our focus next to limited liability, as it allows us to stay within the quasi-linear utility framework.

Remark 9. We have made a rather stark assumption about how the parties interact: the principal gets to propose a contract, and the agent has to take it or leave it. In reality, we might imagine that there is some bargaining over the terms of the contract. Although we cannot predict the outcome without assuming more specifics on the bargaining process, it might be reasonable to assume they choose a contract w that is Pareto optimal (see page **??**). In this case, notice that if w gives the agent an expected payoff u, then w must be an optimal contract for the principal in the model when the agent's outside option $\underline{u}$ is replaced by u. In other words, for a given $(Y, \mathcal{A})$, we can find all the Pareto optimal contracts by varying $\underline{u}$ and solving for the principal's optimal contract for each $\underline{u}$. In the model above this doesn't do much because the contract will always be of the form $w(y) = y - (\text{constant})$, but the same observation can be applied to more complex models, such as the one in the next section.

2.2 A model with limited liability

We can change the model to incorporate limited liability by adding a constraint that contracts should never pay less than some minimum amount. It is without loss of generality to assume this minimum is 0 (otherwise, we can renormalize it to 0 by simply translating the payments in all contracts, and the agent's outside option, by a constant).

Thus, for this section, we define valid contracts as follows:

Definition 10. The contract $((F, c), w)$ is *valid* if it satisfies the incentive compatibility and individual rationality constraints from Definition 5 and further also satisfies

- *(Limited liability)* $w(y) \geq 0$ for all y.

Remark 11. If $\underline{u} \leq -(\min_{(F,c)\in\mathcal{A}} c)$, then the individual rationality condition is redundant. In the previous section, individual rationality was needed to make the principal's problem interesting (otherwise the principal could make arbitrarily large profits by setting all values $w(y)$ to large negative numbers). With limited liability, this is no longer the case, so sometimes we just assume that $\underline{u}$ is low enough that individual rationality can be ignored.

The interaction between the parties, and goal of maximizing the principal's payoff, remain as before.

An optimal contract for this model can be found as follows. For any fixed action $(F, c) \in \mathcal{A}$, we can consider the payment functions w for which $((F, c), w)$ is valid (if any such w exists). Note that optimizing the principal's payoff among all such w is equivalent to minimizing $\mathbb{E}_{y\sim F}[w(y)]$ over such w. Moreover, given the choice of (F, c), the constraints in the definition of a valid contract are linear inequalities on w. Thus we have the following algorithm:

Algorithm 12.

1. For each $(F, c) \in \mathcal{A}$, solve the LP to determine a payment function w that minimizes $\mathbb{E}_{y\sim F}[w(y)]$, subject to $((F, c), w)$ being valid. Record the payoff $v^*(F, c) = \mathbb{E}_{y\sim F}[y - w(y)]$ accordingly. (If the LP is infeasible, set $v^*(F, c) = -\infty$.)
2. Identify the action (F, c) for which $v^*(F, c)$ is maximized, and choose the corresponding optimal payment function w.

At this point, not much can be said about optimal contracts. For example, they may not be *monotone*: the agent may sometimes be paid less for higher levels of output. (See Exercise 1.) And in general, an optimal contract will no longer implement the surplus-maximizing action. To make progress in describing properties of optimal contracts, we need to first make more specific assumptions about the structure of the problem.

One common way of imposing such structure is to assume that the agent's actions are ordered; higher actions may be interpreted as higher levels of effort, and we might assume that higher effort makes higher levels of output relatively more likely. Together with a convexity assumption on output distributions (see Remark 15), this leads to the following definition.

Definition 13. Say the instance $(Y, \mathcal{A}, \underline{u})$ is *monotone convex* if the elements of Y can be labeled $\{y_1, \ldots, y_K\}$ with $y_1 < \cdots < y_K$, and the actions can be labeled $\{(F_1, c_1), \ldots, (F_J, c_J)\}$ with $c_1 < \cdots < c_J$, such that:

- *(Full support)* $f_j(y_k) > 0$ for all j and k;
- *(Monotone likelihood ratio property)*

$$\frac{f_j(y_k)}{f_j(y_{k-1})} > \frac{f_{j-1}(y_k)}{f_{j-1}(y_{k-1})} \text{ for all } 1 < j \leq J, 1 < k \leq K;$$

- *(Convexity)* for each $1 < j < J$, and each $1 \leq k < K$,

$$\frac{F_{j-1}(y_k) - F_j(y_k)}{c_j - c_{j-1}} \geq \frac{F_j(y_k) - F_{j+1}(y_k)}{c_{j+1} - c_j}.$$

Remark 14. In the simple case $K = 2$, we can think of the outcomes as "success" or "failure"; thus, higher effort makes success more likely.

Remark 15. The convexity assumption implies that the the probability of output above y_k (namely $1 - F_j(y_k)$) is a concave function of the effort c_j. It can be understood as saying that an intermediate effort level generates at least as good an output distribution as an equally costly randomization between high and low effort levels.

Proposition 16. *Suppose that $(Y, \mathcal{A}, \underline{u})$ is monotone convex, and assume $\underline{u}$ is low enough so that the individual rationality constraint is redundant.*

There exists an optimal contract $((F, c), w)$ such that $w(y) = 0$ for all $y \neq y_K$.

Moreover, let (F_{j^*}, c_{j^*}) be the surplus-maximizing action (if there is more than one maximizer, let j^* be the highest). Then, any optimal contract implements an action (F_j, c_j) with $j \leq j^*$.

Proof. First, a preliminary observation: we must have $f_{j-1}(y_K) < f_j(y_K)$ for each $j > 1$. This is so because iterated application of the monotone likelihood ratio property implies

$$\frac{f_{j-1}(y_k)}{f_j(y_k)} > \frac{f_{j-1}(y_K)}{f_j(y_K)} \tag{1}$$

for each $k < K$, and so if $f_{j-1}(y_K) \geq f_j(y_K)$ then $f_{j-1}(y_k) > f_j(y_k)$ for each other k, and then we could not have $\sum_k f_{j-1}(y_k) = 1 = \sum_k f_j(y_k)$.

Now on to the contracting problem. It is evident that action (F_1, c_1) can be implemented by paying 0 for every output level. We claim that, for each $j > 1$, action (F_j, c_j) can be implemented by the payment function

$$w(y_K) = \frac{c_j - c_{j-1}}{f_j(y_K) - f_{j-1}(y_K)}, \qquad w(y_k) = 0 \text{ for } k < K,$$

leading to the expected payment

$$\mathbb{E}_{y\sim F_j}[w(y)] = f_j(y_K)\frac{c_j - c_{j-1}}{f_j(y_K) - f_{j-1}(y_K)}, \tag{2}$$

and that there is no cheaper way to implement (F_j, c_j).

First, let us show that action j cannot be implemented more cheaply than claimed. Consider any payment function w. We have

$$\begin{aligned}
\mathbb{E}_{y\sim F_{j-1}}[w(y)] &= \sum_{k=1}^{K} f_{j-1}(y_k)w(y_k) \\
&\geq \sum_{k=1}^{K} \frac{f_{j-1}(y_K)}{f_j(y_K)} f_j(y_k)w(y_k) \\
&= \frac{f_{j-1}(y_K)}{f_j(y_K)}\mathbb{E}_{y\sim F_j}[w(y)].
\end{aligned}$$

(In the inequality step, we have used (1) together with $w(y_k) \geq 0$.) So if w implements (F_j, c_j), then

$$\mathbb{E}_{y\sim F_j}[w(y)] - c_j \geq \mathbb{E}_{y\sim F_{j-1}}[w(y)] - c_{j-1} \geq \frac{f_{j-1}(y_K)}{f_j(y_K)}\mathbb{E}_{y\sim F_j}[w(y)] - c_{j-1},$$

hence

$$\mathbb{E}_{y\sim F_j}[w(y)] \geq \frac{c_j - c_{j-1}}{1 - f_{j-1}(y_K)/f_j(y_K)}$$

which is the asserted lower bound. (The last step uses $f_{j-1}(y_K) < f_j(y_K)$ to ensure the denominator is positive.)

Now let us show that the claimed contract implements action j. That is, writing $r = (c_j - c_{j-1})/(f_j(y_K) - f_{j-1}(y_K))$, we need to show that

$$f_j(y_K)r - c_j \geq f_{j'}(y_K)r - c_{j'} \tag{3}$$

for each j'. Consider the convexity assumption for $k = K-1$, and any $1 < \widetilde{j} < J$. Since $f_{\widetilde{j}}(y_K) = 1 - F_{\widetilde{j}}(y_{K-1})$, we can rewrite the assumption as

$$\frac{f_{\widetilde{j}}(y_K) - f_{\widetilde{j}-1}(y_K)}{c_{\widetilde{j}} - c_{\widetilde{j}-1}} \geq \frac{f_{\widetilde{j}+1}(y_K) - f_{\widetilde{j}}(y_K)}{c_{\widetilde{j}+1} - c_{\widetilde{j}}}. \tag{4}$$

In particular, for $\widetilde{j} \leq j$, we have $f_{\widetilde{j}}(y_K) - f_{\widetilde{j}-1}(y_K) \geq \frac{1}{r}(c_{\widetilde{j}} - c_{\widetilde{j}-1})$, and so for any $j' < j$, by summing over $\widetilde{j} = j'+1, \ldots, j$ we conclude $f_j(y_K) - f_{j'}(y_K) \geq \frac{1}{r}(c_j - c_{j'})$. This implies (3) for $j' < j$. Likewise, (4) gives us $f_{\widetilde{j}}(y_K) - f_{\widetilde{j}-1}(y_K) \leq \frac{1}{r}(c_{\widetilde{j}} - c_{\widetilde{j}-1})$ for $\widetilde{j} > j$, and therefore, for any $j' > j$, summing over $\widetilde{j} = j+1, \ldots, j'$ gives $f_{j'}(y_K) - f_j(y_K) \leq \frac{1}{r}(c_{j'} - c_j)$. This implies (3) for $j' > j$. Thus, the contract implements action j.

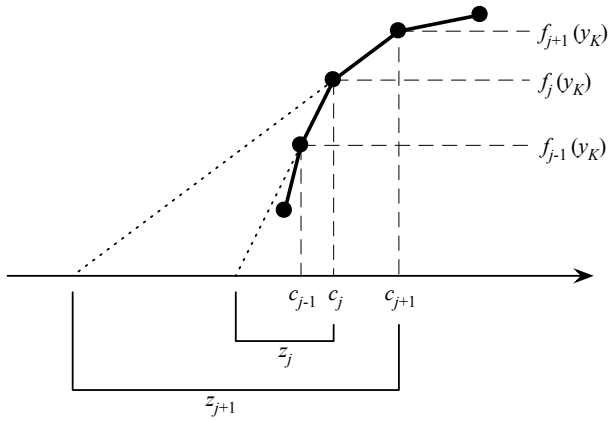

Figure 1: Costs to implement successive actions

We have now shown that, for every action j, the cheapest way to implement it involves paying 0 for every $y \neq y_K$, which proves the first assertion of the proposition.

It remains to prove that any optimal contract implements an action $j \leq j^*$. For $j > 1$, let z_j denote the expected payment given in (2), and put $z_1 = 0$. We claim that $z_j + (c_{j+1} - c_j) \leq z_{j+1}$ for each $j < J$. If $j > 1$, we have

$$\begin{aligned} z_j + (c_{j+1} - c_j) &= f_j(y_K)\frac{c_j - c_{j-1}}{f_j(y_K) - f_{j-1}(y_K)} + (c_{j+1} - c_j) \\ &\leq f_j(y_K)\frac{c_{j+1} - c_j}{f_{j+1}(y_K) - f_j(y_K)} + (c_{j+1} - c_j) \\ &= f_{j+1}(y_K)\frac{c_{j+1} - c_j}{f_{j+1}(y_K) - f_j(y_K)} \\ &= z_{j+1}, \end{aligned}$$

where we have used (4). And if $j = 1$, the claim is immediate from the definition of z_j, so it holds in this case too. Thus, $z_j - c_j$ is increasing in j, or equivalently, $c_j - z_j$ is decreasing. Figure 1 may be helpful in visualizing the situation: if we take the segment between two successive points on the $(c_j, f_j(y_K))$ curve and extend it leftwards, it meets the horizontal axis at point $c_j - z_j$, and the figure shows how concavity of the curve implies $c_j - z_j \geq c_{j+1} - z_{j+1}$.

The principal's payoff from implementing action (F_j, c_j) is $\mathbb{E}_{y\sim F_j}[y] - z_j$. We know that for any $j > j^*$, we have

$$\mathbb{E}_{y\sim F_j}[y] - c_j < \mathbb{E}_{y\sim F_{j^*}}[y] - c_{j^*}$$

by definition of j^*. Adding to this $c_j - z_j \leq c_{j^*} - z_{j^*}$, we have

$$\mathbb{E}_{y\sim F_j}[y] - z_j < \mathbb{E}_{y\sim F_{j^*}}[y] - z_{j^*},$$

so that it is not optimal for the principal to implement (F_j, c_j).

□

Remark 17. The optimal contract identified in Proposition 16 pays only for the highest output level and no others. An intuition is that, due to the monotone likelihood ratio assumption, this output level is the one whose probability is most reduced if the agent deviates to a lower action, and so loading all the payment on this outcome is the most efficient way to discourage such deviations.

Remark 18. One of the basic lessons from this model is that, in general, the agent's action is "distorted downward" relative to the socially optimal action. Due to limited liability, the principal cannot extract the full surplus of whatever action she induces, as she could in the previous section; some of the surplus is left to the agent, and more so for higher actions. Consequently, the principal's preference for inducing high actions is less strong than in the setting without limited liability. Note, however, that this lesson depends on the monotone convex structure we have imposed; see Exercise 3.

Remark 19. Notice that under the optimal contract specified, the agent gets the same payoff from the targeted action (F_j, c_j) or (F_{j-1}, c_{j-1}) (if $j > 1$). This confirms the importance, noted in Remark 6, of specifying how such indifference should be broken.

Although the analysis here leads to some useful insights—the idea of placing payment where it can efficiently discourage deviations, and the downward distortion—arguably the prediction remains fairly unrealistic. In particular, the optimal contract in Proposition 16 gives no incentives over output levels other than the highest one. This relies on a lot of faith in the monotone convex structure. If this structure is assumed when it is not actually correct (either because the principal oversimplifies to apply Proposition 16, or because her belief about the technology is just wrong), then things can go haywire. Suppose, say, that the principal has adopted the optimal contract from Proposition 16, targeting some action (F_j, c_j), but in fact, the agent also has the ability to spend a cost just slightly less than c_j to produce the highest output y_K with probability equal to $f_j(y_K)$ and otherwise produce no output. Then he would prefer to do this, potentially resulting in a severe drop in the principal's payoff.

Our next model incorporates this concern by assuming less knowledge of the technology.

2.3 A robust model

Let us keep the limited liability assumption, but now assume that the principal does not fully know the agent's technology. Instead, she only knows some actions that are available to the agent, but envisions that there may be other actions available. Thus, an instance of the model is now given by $(Y, \mathcal{A}_0, \underline{u})$, where $\mathcal{A}_0$ is a technology representing the actions known to the principal.

For this section, let us make four assumptions that will simplify the analysis:

- Assume that $\underline{u} < -\min_{(F,c)\in\mathcal{A}_0} c$, so that individual rationality will not be a concern.
- Assume that $\min(Y) = 0$. (This is an innocuous normalization; it can be achieved by adding a constant to every element of Y without changing the principal's optimization problem.)
- Also assume that for every $(F, c) \in \mathcal{A}_0$, $c > 0$. (This gets rid of some messy edge cases.)
- Assume there exists $(F, c) \in \mathcal{A}_0$ whose surplus $s(F, c)$ is positive.

We redefine contracts for this section as follows:

Definition 20. A *contract* is a function $w : Y \to \mathbb{R}_+$.

Thus, limited liability is incorporated, but a contract no longer prescribes what action the agent should take, since this cannot be specified without knowing what actions are available.

Definition 21. Contract w *guarantees* a payoff level v to the principal if, for every technology $\mathcal{A}$ such that $\mathcal{A}_0 \subseteq \mathcal{A}$, there exists an action $(F, c) \in \mathcal{A}$ such that

- $\mathbb{E}_{y\sim F}[w(y)] - c \geq \mathbb{E}_{y\sim F'}[w(y)] - c'$ for every $(F', c') \in \mathcal{A}$, and
- $\mathbb{E}_{y\sim F}[y - w(y)] \geq v$.

Evidently, if a contract guarantees v, it also guarantees any $v' \leq v$. We may refer to the supremum of all v guaranteed by a given contract as *the guarantee* of the contract.

Thus, for any contract, we are interested in understanding what payoff it guarantees to the principal in spite of her uncertainty about the true technology $\mathcal{A}$.

To illustrate how one can show that a contract guarantees a certain payoff, we now consider a particularly simple class of contracts:

Definition 22. Contract w is *linear* if there exists a fraction $\alpha \in (0, 1)$ such that $w(y) = \alpha y$ for all $y \in Y$.

Linear contracts are widely seen in practice: for example, think of sales agents who are paid a fixed percentage of each sale they make.

Observation 23. *Suppose that the principal uses a linear contract $w(y) = \alpha y$. Write $u_0(\alpha) = \max_{(F,c)\in\mathcal{A}_0}(\alpha\mathbb{E}_{y\sim F}[y] - c)$. Then, the contract guarantees payoff $\frac{1-\alpha}{\alpha}u_0(\alpha)$.*

(In particular, note that the guarantee is positive if α is close to 1, because we assumed a positive-surplus action exists.)

Proof. Note that for any technology $\mathcal{A}$ containing $\mathcal{A}_0$, and any optimal action $(F, c) \in \mathcal{A}$ for the agent, we have

$$\alpha \mathbb{E}_{y\sim F}[y] \geq \alpha \mathbb{E}_{y\sim F}[y] - c \geq u_0(\alpha),$$

since any action in $\mathcal{A}_0$ is also in $\mathcal{A}$. Therefore

$$\mathbb{E}_{y\sim F}[y - w(y)] = (1-\alpha)\mathbb{E}_{y\sim F}[y] \geq \frac{1-\alpha}{\alpha} u_0(\alpha).$$

□

After seeing this argument, one might next ask: is there any theoretical reason to focus on linear contracts here? The next result shows that there is: given the goal of maximizing the guarantee, without loss of generality, we can restrict attention to linear contracts.

Theorem 24. *If any contract w guarantees a payoff level $v > 0$, then there exists a linear contract that also guarantees v.*

The proof proceeds roughly as follows. First, we write down a linear program that identifies the "worst-case" action the agent might choose if offered w. We then use a geometric separation argument to find a linear contract that matches w for this worst-case action, and show that because the linear contract better aligns the agent's interests with the principal's, its guarantee can only be better.

Proof. Denote $u_0(w) = \max_{(F,c)\in\mathcal{A}_0} (\mathbb{E}_{y\sim F}[w(y)] - c)$. Note that $u_0(w) < \max_y w(y)$, by our assumption that $c > 0$ for all $(F, c) \in \mathcal{A}_0$.

Consider the problem of minimizing $\mathbb{E}_{y\sim F}[y - w(y)]$ over all $F \in \Delta(Y)$ such that $\mathbb{E}_{y\sim F}[w(y)] \geq u_0(w)$. This is a linear program for F, and we claim that this problem has a solution where the constraint is satisfied with equality. If not, then the constraint can be dropped, which implies that the function $y - w(y)$ attains its minimum over Y at a point $\underline{y}$ with $w(\underline{y}) > u_0(w)$. Then consider the distribution F' that simply places probability 1 on this $\underline{y}$; if the technology is $\mathcal{A} = \mathcal{A}_0 \cup \{(F', 0)\}$, then the unique optimal action for the agent is $(F', 0)$. Since the principal then receives $\underline{y} - w(\underline{y}) \leq 0 - w(0) \leq 0$ (since $0 \in Y$), the contract cannot guarantee any payoff level above 0, a contradiction.

Now let $\widetilde{F}$ be a solution to the minimization problem above, and let $\widetilde{v}$ be the resulting objective value. We claim that $\widetilde{v}$ is exactly the guarantee of w, so that $\widetilde{v} \geq v$. Indeed, for any possible technology $\mathcal{A}$, an optimal action $(F, c) \in \mathcal{A}$ necessarily satisfies $\mathbb{E}_{y\sim F}[w(y)] \geq \mathbb{E}_{y\sim F}[w(y)] - c \geq u_0(w)$, i.e. F is feasible in the minimization problem, so $\mathbb{E}_{y\sim F}[y - w(y)] \geq \widetilde{v}$, and w guarantees at least $\widetilde{v}$. Conversely, for any $\varepsilon > 0$, by perturbing $\widetilde{F}$ we can find $F' \in \Delta(Y)$ such that $\mathbb{E}_{y\sim F'}[y - w(y)] < \widetilde{v} + \varepsilon$ and $\mathbb{E}_{y\sim F'}[w(y)] > u_0(w)$ (here we have used the fact that $u_0(w) < \max(w)$), so if we consider the technology $\mathcal{A} = \mathcal{A}_0 \cup \{(F', 0)\}$, the agent's unique optimal action is $(F', 0)$, showing that w does not guarantee $\widetilde{v} + \varepsilon$.

Now we proceed to the separation argument. Define two convex subsets S and T of $\mathbb{R}^2$ as follows: S is the convex hull of all points $(y, w(y))$ for $y \in Y$,

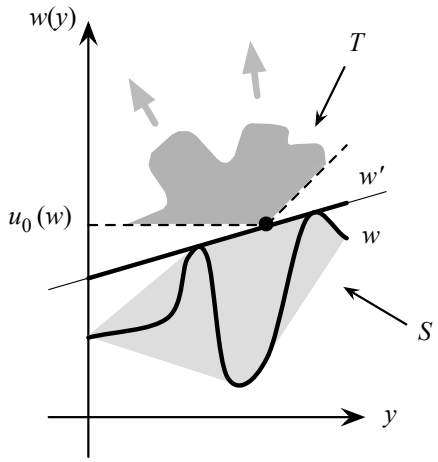

Figure 2: Separation argument and improvement to a linear contract

while T consists of all points (p, q) such that $q > u_0(w)$ and $p - q < \widetilde{v}$. These sets are disjoint: any point in their intersection would correspond to a feasible point in the minimization problem having objective value $p - q < \widetilde{v}$, which cannot exist. Therefore, there exists some line in $\mathbb{R}^2$ separating S and T. Expressing this statement algebraically: there exist values $\kappa, \lambda, \mu \in \mathbb{R}$, with κ, λ not both zero, such that

$$\kappa p + \lambda q \leq \mu \qquad \text{for all } (p, q) \in S; \tag{5}$$

$$\kappa p + \lambda q \geq \mu \qquad \text{for all } (p, q) \in T. \tag{6}$$

The argument is illustrated in Figure 2. Sets S and T are the two shaded regions. (Although we have assumed Y is finite, w is shown as a curve for visual clarity.)

Note that for large M, the points $(-M, u_0(w)+1/M)$ and $(\widetilde{v}+M, M+1/M)$ are both in T. Hence, we must have $\kappa \leq 0$ and $\kappa+\lambda \geq 0$, since otherwise one or the other of these points would violate (6) for large enough M. Together with $(\kappa, \lambda) \neq (0, 0)$, these imply $\lambda > 0$. Also, note that the point $(u_0(w)+\widetilde{v}, u_0(w)) = (\mathbb{E}_{y\sim\widetilde{F}}[y], \mathbb{E}_{y\sim\widetilde{F}}[w(y)])$ lies in S, and it also lies in the closure of T, so it must lie on the line:

$$\kappa(u_0(w) + \widetilde{v}) + \lambda u_0(w) = \mu. \tag{7}$$

The latter implies in turn that $\kappa < 0$, since if $\kappa = 0$, then (5) would imply $w(y) \leq \mu/\lambda = u_0(w)$ for all y, contradicting $u_0(w) < \max(w)$.

Now we can define our new contract. Condition (5) implies that $w(y) \leq \frac{\mu-\kappa y}{\lambda}$ for all $y \in Y$. Accordingly, define w' by $w'(y) = \frac{\mu-\kappa y}{\lambda}$. Then $w'(y) \geq w(y) \geq 0$ for all y, so w' is indeed a contract. (It is shown by the thick straight line in Figure 2.)

We claim that w' again guarantees at least $\widetilde{v}$. The calculation is a variant of the one in Observation 23. Write $\alpha = -\kappa/\lambda > 0$ and $\beta = \mu/\lambda$, so $w'(y) = \alpha y + \beta$. Equation (7) can be rewritten as $(1-\alpha)u_0(w) - \alpha\widetilde{v} = \beta$. Now consider any technology $\mathcal{A}$ containing $\mathcal{A}_0$, and any $(F', c') \in \mathcal{A}$ optimal for the agent:

$$\mathbb{E}_{y\sim F'}[w'(y)] - c' \geq \max_{(F,c)\in\mathcal{A}_0} (\mathbb{E}_{y\sim F}[w'(y)] - c) \geq \max_{(F,c)\in\mathcal{A}_0} (\mathbb{E}_{y\sim F}[w(y)] - c) = u_0(w),$$

where the second inequality follows from $w'(y) \geq w(y)$. Hence,

$$\alpha\mathbb{E}_{y\sim F'}[y] + \beta = \mathbb{E}_{y\sim F'}[w'(y)] \geq u_0(w),$$

and so

$$\mathbb{E}_{y\sim F'}[y - w'(y)] = (1-\alpha)\mathbb{E}_{y\sim F'}[y] - \beta \geq (1-\alpha)\left(\frac{u_0(w)-\beta}{\alpha}\right) - \beta = \widetilde{v}.$$

This shows that w' guarantees $\widetilde{v}$ as claimed.

At this point, we have shown that there exists an "affine" contract, i.e. one of the form $w'(y) = \alpha y + \beta$ with α, β constants, that guarantees $\widetilde{v} \geq v$. We have noted $\alpha > 0$, and $\beta = w'(0) \geq 0$. Note also that $\alpha < 1$, because otherwise $w'(y) \geq y$ for all y, contradicting the fact that w' has a positive guarantee. Finally, if $\beta > 0$ strictly, then replacing $w'(y)$ with just αy can only improve the guarantee (since, for any technology $\mathcal{A}$, the agent's optimal action is the same as before, and now the principal's payoff increases by β). After this change, our definition of a linear contract is met.

□

This theorem shows that, if contracts are evaluated by their worst-case guarantee, then a linear contract is optimal. (To be precise, we have not yet shown that the optimum is attained; for this, see Exercise 4.)

2.4 Applications of hidden-action models

While we have focused on the employment application for hidden-action models, it is worth emphasizing that such models have many other applications: The concepts are relevant any time one party designs incentives that influence the action taken by another, by promising material rewards (which may be money, or something else, e.g. social status) contingent on a noisy signal of the action chosen. Here are just a few more examples to illustrate the breadth of applications:

- Financial contracting: an investor writes a contract with a startup founder, specifying how profits are shared between the parties; the investor needs to be offered an adequate return, while the founder needs to be given incentives to invest the money productively.
- Health insurance: a classic application featuring moral hazard and risk aversion. The company wants to design the insurance contract to protect the consumer against risks without incentivizing the consumer to spend too much on unnecessary procedures.

- Reputation systems: the design of the system by which ratings (for sellers on eBay, drivers on Lyft, etc.) translate into future business affects their incentives for performance.
- Political accountability: voters can reelect politicians or not, depending how well they (appear to) have governed; some theorists have studied the extent to which these reelection incentives can induce good performance.

3 Hidden-Information Models

In hidden-information models, the agent holds private information about his preferences, and an allocation is chosen as a function of this information. Like hidden-action models, there are a wide variety of applications. Here, for concreteness, we will envision the principal as a seller of a product and the agent as a buyer. The product can be offered in different quality levels (think of a hotel offering various room types, or a cell phone provider offering multiple service plans), from which the buyer will choose depending on his preferences. Section 3.2 will list some other applications.

We will present a classic version of such a model here. We will adopt a formulation with continuous types, as this makes the mathematics particularly clean (and, if the reader has seen the theory of optimal auctions, much of the analysis will look familiar); but discrete formulations are also possible.

Models of this sort are also often called *screening* models: the seller "screens" the different types of buyer by offering multiple options that are chosen by different types.

Remark 25. Some authors also call these *adverse selection* models. There is some confusion in the literature about this term. It comes from the world of insurance, where different buyers may choose different products, and thereby sort themselves in a way that makes the products more costly for the seller to provide. For example, a health insurance contract that is sold at a high price but covers a large percentage of costs is especially likely to attract very sick people, and these are precisely the people that the insurance company would prefer *not* to attract. In the model presented below, there is nothing "adverse": the cost of providing a given quality q does not depend on who buys it. Nonetheless, the term is sometimes applied to such a model.

3.1 A price-discrimination model

The agent can be given a product of *quality* q, which can be chosen from some interval $[\underline{q}, \overline{q}] \subseteq \mathbb{R}$. The agent's preference over qualities is determined by his *type* θ, drawn from a given interval $[\underline{\theta}, \overline{\theta}] \subseteq \mathbb{R}$. It is important that both the possible qualities and the types are "ordered," with higher types both valuing the product more overall and having stronger preference for high quality; we shall formalize this shortly.

If the agent of type θ purchases a good of quality q and pays a transfer t for it, his overall payoff is $u(q, \theta) - t$. We normalize the agent's outside option to 0. Here u is a function satisfying the following conditions:

- u is twice continuously differentiable. (We will denote its derivatives with respect to particular arguments via subscripts, thus writing u_θ, $u_{q\theta}$, etc.)
- u is weakly increasing in θ, for each $q \in [\underline{q}, \overline{q}]$. (It is common to interpret the minimum quality $\underline{q}$ as not receiving a product, in which case one often sets $\underline{q} = 0$ and assumes $u(0, \theta) = 0$ for all θ.)
- u is *strictly supermodular*: for qualities $q' > q$, $u(q', \theta) - u(q, \theta)$ is strictly increasing in θ. (Given the differentiability assumption, this is equivalent to $u_{q\theta} \geq 0$, with strict inequality on a dense set.)

Producing (or acquiring) a product of quality q costs $c(q)$ to the principal. Thus, if the agent receives q and pays t, the principal's payoff is $t - c(q)$. Assume that c is continuous.

From the principal's point of view, the agent's type is unknown; it is drawn from a distribution F, assumed to have a density f on $[\underline{\theta}, \overline{\theta}]$, which is continuous and strictly positive throughout the interval. As before, we write $F(\theta)$ for the probability of drawing a type $\leq \theta$.

An instance of the model is then given by the tuple $(\underline{q}, \overline{q}, \underline{\theta}, \overline{\theta}, u, c, f)$.

We can envision the principal offering a price for each quality q (or perhaps a subset of the possible qualities), and letting the agent choose his favorite from the offered (quality, price) pairs. However, by the revelation principle (see Section **??**), the outcome of such an interaction can equivalently be described by a *(direct) mechanism* that specifies the quality and price chosen by each type:

Definition 26. A *mechanism* (q, t) consists of two measurable functions $q : [\underline{\theta}, \overline{\theta}] \to [\underline{q}, \overline{q}]$ and $t : [\underline{\theta}, \overline{\theta}] \to \mathbb{R}$. Sometimes q is called the *allocation function* (or *allocation rule*) and t is the *payment function*.

Although we use the same notation q both for a typical quality level and for an allocation function (and t likewise), the meaning should be clear from context.

Definition 27. The mechanism (q, t) is *valid* if it satisfies the following conditions:

- *(Incentive compatibility)* $u(q(\theta), \theta) - t(\theta) \geq u(q(\widehat{\theta}), \theta) - t(\widehat{\theta})$ for all $\theta, \widehat{\theta} \in [\underline{\theta}, \overline{\theta}]$.
- *(Individual rationality)* $u(q(\theta), \theta) - t(\theta) \geq 0$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$.

The principal's expected payoff from mechanism (q, t) is $\mathbb{E}_{\theta \sim F}[t(\theta) - c(q(\theta))]$. The problem is to find a valid mechanism that maximizes this payoff.

The allocation function q is *implementable* if there is some payment function t such that (q, t) is a valid mechanism (and we say that t *implements* q). A first question is what allocation functions are implementable—and what payments implement them.

Proposition 28. *If (q,t) is a valid mechanism, then the function q is weakly increasing.*

Conversely, any weakly increasing q is implementable, and it is implemented by t if and only if there is some constant $C \geq 0$ such that

$$t(\theta) = u(q(\theta),\theta) - \int_{\underline{\theta}}^{\theta} u_\theta(q(\widetilde{\theta}),\widetilde{\theta})\, d\widetilde{\theta} - C \tag{8}$$

for all θ.

Proof. First, suppose (q,t) is valid, and suppose for contradiction that $\theta < \theta'$ are two distinct types with $q(\theta) > q(\theta')$. By incentive compatibility,

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

from which

$$u(q(\theta),\theta) - u(q(\theta'),\theta) \geq t(\theta) - t(\theta').$$

Similarly, $u(q(\theta'),\theta') - u(q(\theta),\theta') \geq t(\theta') - t(\theta)$. Combining,

$$u(q(\theta),\theta) - u(q(\theta'),\theta) \geq t(\theta) - t(\theta') \geq u(q(\theta),\theta') - u(q(\theta'),\theta').$$

This contradicts the strict supermodularity assumption, which tells us that $u(q(\theta),\cdot) - u(q(\theta'),\cdot)$ is strictly increasing. This shows that q must be weakly increasing.

Henceforth, fix any q that is weakly increasing. We first show every function t implementing q has the form in (8). For such a t, let

$$U(\theta) = u(q(\theta),\theta) - t(\theta) \tag{9}$$

be the payoff earned by an agent of type θ in the mechanism. We claim that U is Lipschitz continuous. To see this, let λ be an upper bound for $|u_\theta|$ (which exists since u_θ is continuous); then for any θ, θ', incentive compatibility implies

$$U(\theta') \geq u(q(\theta),\theta') - t(\theta) \geq (u(q(\theta),\theta) - \lambda|\theta' - \theta|) - t(\theta) = U(\theta) - \lambda|\theta' - \theta|.$$

Writing the corresponding inequality with θ and θ' reversed, and combining, we obtain $|U(\theta') - U(\theta)| \leq \lambda|\theta' - \theta|$, so U is Lipschitz continuous as claimed. This in turn implies U is absolutely continuous, i.e. it is differentiable almost everywhere and is given by the integral of its derivative.

However, if we fix any $\theta \in (\underline{\theta}, \overline{\theta})$ at which U is differentiable, incentive compatibility says that $U(\theta') \geq u(q(\theta),\theta') - t(\theta)$ for all θ', with equality when $\theta' = \theta$. Both sides of this inequality are differentiable in θ' at the equality point $\theta' = \theta$, so their derivatives there must coincide, i.e.

$$\frac{dU}{d\theta} = u_\theta(q(\theta),\theta).$$

Thus we have, for each θ,

$$U(\theta) - U(\underline{\theta}) = \int_{\underline{\theta}}^{\theta} u_\theta(q(\widetilde{\theta}),\widetilde{\theta})\, d\widetilde{\theta}.$$

Substituting into (9) and solving for $t(\theta)$ shows that $t(\theta)$ has the form given in (8), with $C = U(\underline{\theta})$. Finally, $C \geq 0$ follows from individual rationality.

Let us conversely show that any payment function as in (8) implements q. First, we verify incentive compatibility. Consider any θ and $\widehat{\theta}$. Define U as in (9), and note that $U(\theta) = \int_{\underline{\theta}}^{\theta} u_\theta(q(\widetilde{\theta}), \widetilde{\theta})\, d\widetilde{\theta} + C$.

Now,

$$u(q(\widehat{\theta}), \theta) - t(\widehat{\theta}) = U(\widehat{\theta}) + (u(q(\widehat{\theta}), \theta) - u(q(\widehat{\theta}), \widehat{\theta})).$$

Checking incentive compatibility is thus equivalent to checking

$$U(\theta) - U(\widehat{\theta}) \geq u(q(\widehat{\theta}), \theta) - u(q(\widehat{\theta}), \widehat{\theta}),$$

or equivalently

$$\int_{\widehat{\theta}}^{\theta} u_\theta(q(\widetilde{\theta}), \widetilde{\theta})\, d\widetilde{\theta} \geq \int_{\widehat{\theta}}^{\theta} u_\theta(q(\widehat{\theta}), \widetilde{\theta})\, d\widetilde{\theta}. \tag{10}$$

For $\widehat{\theta} < \underline{\theta}$, this is true because monotonicity of q and supermodularity of u imply $u_\theta(q(\widetilde{\theta}), \widetilde{\theta}) \geq u_\theta(q(\widehat{\theta}), \widetilde{\theta})$ for each $\widetilde{\theta} > \widehat{\theta}$. For $\widehat{\theta} > \theta$, (10) has the integration ranges "reversed," with the lower endpoint above the upper endpoint, so it may be easier to understand in the equivalent form

$$\int_{\theta}^{\widehat{\theta}} u_\theta(q(\widetilde{\theta}), \widetilde{\theta})\, d\widetilde{\theta} \leq \int_{\theta}^{\widehat{\theta}} u_\theta(q(\widehat{\theta}), \widetilde{\theta})\, d\widetilde{\theta}.$$

Again, this is true because monotonicity and supermodularity ensure $u_\theta(q(\widetilde{\theta}), \widetilde{\theta}) \leq u_\theta(q(\widehat{\theta}), \widetilde{\theta})$ on the relevant range of $\widetilde{\theta}$.

Finally, $U(\underline{\theta}) = C \geq 0$, so individual rationality is satisfied for type $\underline{\theta}$. For any other type θ, individual rationality then follows from incentive compatibility and monotonicity of u in the type, as

$$u(q(\theta), \theta) - t(\theta) \geq u(q(\underline{\theta}), \theta) - t(\underline{\theta}) \geq u(q(\underline{\theta}), \underline{\theta}) - t(\underline{\theta}) \geq 0.$$

□

Given Proposition 28, we can rewrite the principal's problem as follows: choose a weakly increasing function $q : [\underline{\theta}, \overline{\theta}] \to [\underline{q}, \overline{q}]$ and a constant $C \geq 0$ to maximize

$$\int_{\underline{\theta}}^{\overline{\theta}} \left(u(q(\theta), \theta) - \int_{\underline{\theta}}^{\theta} u_\theta(q(\widetilde{\theta}), \widetilde{\theta})\, d\widetilde{\theta} - C - c(q(\theta)) \right) f(\theta)\, d\theta.$$

Clearly it is optimal to set $C = 0$. For the inner integral, we can swap the order of integration between θ and $\widetilde{\theta}$, then relabel $\widetilde{\theta}$ as θ; with these steps, we can rewrite the principal's objective as

$$\int_{\underline{\theta}}^{\overline{\theta}} \left(u(q(\theta), \theta) - \frac{1 - F(\theta)}{f(\theta)} u_\theta(q(\theta), \theta) - c(q(\theta)) \right) f(\theta)\, d\theta. \tag{11}$$

The nice thing about this expression is that, for each value of θ, the integrand depends on $q(\theta)$ but not on the values of q at any other points. Thus, we can hope to optimize $q(\theta)$ for each θ separately.

With this in mind, we define the following objects:

Definition 29. $\overline{u}(q,\theta) = u(q,\theta) - \frac{1-F(\theta)}{f(\theta)} u_\theta(q,\theta)$ is the *virtual value* of type θ for quality q. $\overline{s}(q,\theta) = \overline{u}(q,\theta) - c(q)$ is the *virtual surplus* with type θ and quality q.

For each θ, let $q^\circ(\theta)$ be the quality that maximizes the virtual surplus $\overline{s}(q,\theta)$ (or the highest maximizer, if there is more than one). With a few further assumptions, we can characterize an optimal solution to the principal's problem.

Theorem 30. *Suppose that the utility function u is three times continuously differentiable, with $u_{q\theta\theta} \leq 0$ everywhere, and that the function $h(\theta) = f(\theta)/(1 - F(\theta))$ is differentiable and increasing in θ.*

Then, taking $q(\theta) = q^\circ(\theta)$, and taking payments $t(\theta)$ given by (8) with $C = 0$, gives a valid mechanism that maximizes the principal's payoff.

Proof. It is immediate that the suggested allocation function maximizes (11) as long as it is indeed weakly increasing. To check this, we show that the virtual surplus $\overline{s}(q,\theta)$ is weakly supermodular (i.e. for $q' > q$, $\overline{s}(q',\theta) - \overline{s}(q,\theta)$ is weakly increasing in θ). This will imply the result, since if $\theta < \theta'$ but $q^\circ(\theta) > q^\circ(\theta')$, we would have $\overline{s}(q^\circ(\theta),\theta') - \overline{s}(q^\circ(\theta'),\theta') < 0 \leq \overline{s}(q^\circ(\theta),\theta) - \overline{s}(q^\circ(\theta'),\theta)$, contradicting supermodularity.

Writing $\overline{s}(q,\theta) = u(q,\theta) - u_\theta(q,\theta)/h(\theta) - c(q)$, we take the cross-partial derivative and find

$$\overline{s}_{q\theta} = u_{q\theta} - \frac{u_{q\theta\theta}}{h} + \frac{u_{q\theta} h_\theta}{h^2}.$$

Every term is nonnegative, so $\overline{s}_{q\theta} \geq 0$, and $\overline{s}$ is supermodular. □

Remark 31. The technical condition that h be increasing, or equivalently that $1 - F(\theta)$ be log-concave, is satisfied by many standard probability distributions, such as uniform, truncated normal, or truncated exponential.

It is common to compare the qualities in the principal's optimal mechanism with the socially optimal qualities. For each type θ, let $q^*(\theta)$ be the quality that maximizes the surplus $s(q,\theta) = u(q,\theta) - c(q)$ (or the highest such quality, if there is more than one). Note that q^* is weakly increasing in θ, by an argument similar to the proof of Theorem 30, because the surplus is supermodular.

Observation 32. *For each θ, we have $q^\circ(\theta) \leq q^*(\theta)$, with equality at $\theta = \overline{\theta}$.*

Proof. If $q^\circ(\theta) > q^*(\theta)$, we would have $s(q^\circ(\theta),\theta) < s(q^*(\theta),\theta)$ and $\overline{s}(q^*(\theta),\theta) \leq \overline{s}(q^\circ(\theta),\theta)$, implying $s(q^\circ(\theta),\theta) - \overline{s}(q^\circ(\theta),\theta) < s(q^*(\theta),\theta) - \overline{s}(q^*(\theta),\theta)$. But $s(q,\theta) - \overline{s}(q,\theta) = u_\theta(q,\theta)/h(\theta)$ which is weakly increasing in q, so we get a contradiction.

Also, to see the equality statement, just note that $\overline{s}(q,\theta) = s(q,\theta)$ when $\theta = \overline{\theta}$. □

Observation 32 is often summarized by saying there is "no distortion at the top, downward distortion elsewhere." (Compare to Remark 18 from the hidden-action model.)

An intuition is that giving a less-than-efficient quality level to type θ reduces the total surplus, but it increases the amount of money the principal can extract from types above θ, because they are willing to pay an especially large amount to avoid getting stuck with a low quality. The optimal choice of quality trades off these two effects. For the highest type $\overline{\theta}$, there are no higher types, so the second effect disappears, which is why no distortion arises.

Remark 33. The idea behind the distortion is sometimes credited to the nineteenth-century civil engineer Jules Dupuit, who wrote the following about third-class train cars: "It is not because of the few thousand francs which would have to be spent to put a roof over the third-class carriages or to upholster the third-class seats that some company or other has open carriages with wooden benches... What the company is trying to do is prevent the passengers who can pay the second-class fare from traveling third class; it hits the poor, not because it wants to hurt them, but to frighten the rich."

However, Dupuit went on to assert that first-class carriages were made excessively luxurious for the same reason. In the model here, this would mean that high types receive above-efficient quality levels. But the model does not predict this. One might think that giving a high type θ an excessive quality allows the seller to extract more payments from lower types, but this is incorrect: prices are limited by the high types' incentive to imitate lower ones, not vice versa. (This depends on our assumption about outside options; see Exercise 6.)

Remark 34. As already mentioned, if the reader is familiar with the theory of optimal auctions, many of the steps here are similar. This is more than coincidence: Indeed, the one-buyer optimal "auction" is just a special case of this model, where q is interpreted as a probability of receiving the object rather than a quality, $u(q, \theta) = q\theta$ and $c(q) = 0$.

Remark 35. What happens if the extra assumptions of Theorem 30 do not hold? In this case, the allocation function q obtained by maximizing virtual surplus may not be increasing, and therefore may not be implementable. Instead, a process called "ironing" must be applied to account for the monotonicity constraint on q, resulting in intervals of types that all receive same quality. Regardless, it remains true that an optimal mechanism only distorts downwards: if we consider any weakly increasing $q(\theta)$ that sometimes distorts upwards, the alternative allocation function $\min\{q(\theta), q^*(\theta)\}$ is also weakly increasing and yields higher profit than q, by logic similar to Observation 32.

Remark 36. We could also imagine giving the seller more power by allowing randomized mechanisms, that specify a probability distribution over qualities and payments for each type.

Randomizing the payment is not relevant, since our agent has quasi-linear utility and so paying a random amount is equivalent to just paying the expected

value. However, randomizing the allocation function can be useful in better screening the different types of agents. One can repeat the analysis above, allowing for randomized mechanisms, and show that the expected profit still equals the expected virtual surplus. For each type θ, the virtual surplus can be maximized over all randomized qualities by just putting probability 1 on $q^\circ(\theta)$. So if $q^\circ(\theta)$ is increasing, as it is under the assumptions of Theorem 30, then it remains optimal even among randomized mechanisms. However, when the assumptions do not hold, one can give instances in which randomized mechanisms do strictly better than deterministic ones.

3.2 Applications of hidden-information models

As with hidden-action models, hidden-information models have many applications. For example, they can apply to sales of a product or service, not only when the allocation variable is quality, but also when it is some other dimension, such as quantity or speed of service, that some buyer types value more than others. Here are some other domains of application, with suitable reinterpretations of the variables:

- Seller-side information: these models can also apply to trades where the seller, rather than the buyer, holds private information. For example, perhaps a firm is buying some input from a supplier and negotiating over the terms of sale, and the firm has uncertainty about how costly it is for the supplier to produce a given quality or quantity of the input.
- Taxation: models based on the framework here are widely applied to study tax systems. Here, individuals have private information about their income-earning ability, and the tax schedule shapes their preferences for choosing one job (or level of work intensity) over another.

Many of the applications from Section 2 can also be approached through hidden-information models instead; one or the other approach may be more relevant depending on the situation. For example, suppose an employer hires a worker to produce output, and the worker's effort maps *deterministically* to output; however, the worker has private information about his ability, which determines how hard it is for him to produce any given level of output. This is naturally described by a hidden-information model, where θ is the worker's ability, and q is the level of output he produces; the employer offers wages for each possible output level, and the different types of worker self-select. This is sometimes called a "false moral hazard" model.

4 Exercises

1. Consider the hidden-action model with limited liability as in Section 2.2. Suppose that there are just two actions. Assume that $\underline{u}$ is low enough so that individual rationality is not relevant.

(a) Show that there is an optimal contract that pays a positive amount for at most one output level. Identify this contract explicitly.

(b) Give an example of an instance where any optimal contract fails to be monotone: that is, there exist output levels $y < y'$ such that $w(y) > w(y')$.

2. Consider the hidden-action model with limited liability as in Section 2.2. Assume that $\underline{u} > 0$. Let (F^*, c^*) be the surplus-maximizing action, with surplus s^*, and suppose there exists some output level y that receives positive probability under F^*, but zero probability under F for any other $(F, c) \in \mathcal{A}$. Prove that the principal can achieve a payoff of $s^* - \underline{u}$.

3. Consider the hidden-action model with limited liability as in Section 2.2. Give an instance where there is a unique surplus-maximizing action, and where an optimal contract induces an action that has *higher* cost than the surplus-maximizing one.

4. Consider the robust contracting model with limited liability as in Section 2.3. Consider a linear contract $w(y) = \alpha y$ with $\alpha \in (0, 1)$.

   (a) Suppose that $u_0(\alpha) > 0$. Show that the guarantee of this contract is exactly $\frac{1-\alpha}{\alpha} u_0(\alpha)$.

   (b) In terms of $\mathcal{A}_0$, characterize explicitly the α for which this guarantee is maximized.

5. Consider the robust contracting model, but now without limited liability. Thus, a contract is defined as a function $w : Y \to \mathbb{R}$, such that $\max_{(F,c)\in\mathcal{A}_0} (\mathbb{E}_{y\sim F}[w(y)] - c) \geq \underline{u}$ (to ensure individual rationality). A guaranteed payoff (for the principal) is defined as before. Identify the highest possible guaranteed payoff, and a contract that achieves it.

6. Consider the hidden-information model in Section 3.1, but now suppose that higher types of agent have much higher outside options: the individual rationality constraint is replaced by requiring $u(q(\theta), \theta) - t(\theta) \geq \underline{u}(\theta)$ for each θ, where $\underline{u}$ is a function such that, for each q, $\underline{u}(\theta) - u(q, \theta)$ is increasing in θ. Repeat the analysis. What changes? (Note: to give a characterization of the optimal mechanism analogous to Theorem 30, you may need to modify the extra regularity assumptions made in that theorem.)

7. Consider the hidden-information model, but now drop the differentiability assumptions on u: instead, assume only that u is continuous, weakly increasing in θ, and strictly supermodular. Show by example that there can be two different payment functions $t, t' : [\underline{\theta}, \overline{\theta}] \to \mathbb{R}$ that both implement the same allocation function q, such that $t' - t$ is not constant.

8. Consider the hidden-information model, but now suppose that the agent learns his type *after* agreeing to transact with the principal. That is:

- first the principal offers a mechanism, specifying qualities and prices (q, t);
- the agent decides whether to accept the mechanism or reject it and receive his outside option payoff of 0;
- if the agent has accepted the mechanism, his type is drawn, $\theta \sim F$; and then
- he selects a (quality, price) pair from among those offered by the principal (he cannot exit the mechanism at this point, even if his best option gives him negative payoff).

Show how to formulate the incentive compatibility and individual rationality constraints in this model. Identify an optimal mechanism. What is the principal's payoff?

5 Notes

A careful treatment of a hidden-action model with risk aversion (which we have not covered here) can be found in Grossman and Hart (1983). (The conditions in the "monotone convex" formulation here also appear in that paper.) A canonical reference for a formulation without risk aversion, and with limited liability, is Innes (1990). The robust contracting model of Section 2.3 draws heavily on Carroll (2015).

Classic references for the basic hidden-information model are Mussa and Rosen (1978) and Maskin and Riley (1984). The remark on randomized mechanisms draws on Strausz (2006). For the theory of optimal auctions, which uses much of the same machinery, see Myerson (1981). The passage from Jules Dupuit is as quoted in Ekelund (1970).

6 Acknowledgments

This writing was supported by an NSF CAREER grant. Ellen Muir provided valuable research assistance.

References

Carroll, Gabriel. 2015. Robustness and Linear Contracts. *American Economic Review*, **105**(2), 536–63.

Ekelund, Jr, Robert B. 1970. Price Discrimination and Product Differentiation in Economic Theory: An Early Analysis. *The Quarterly Journal of Economics*, 268–278.

Grossman, Sanford J, and Hart, Oliver D. 1983. An Analysis of the Principal-Agent Problem. *Econometrica*, **51**(1), 7–46.

Innes, Robert D. 1990. Limited Liability and Incentive Contracting with Ex-Ante Action Choices. *Journal of Economic Theory*, **52**(1), 45–67.

Maskin, Eric, and Riley, John. 1984. Monopoly with Incomplete Information. *The RAND Journal of Economics*, **15**(2), 171–196.

Mussa, Michael, and Rosen, Sherwin. 1978. Monopoly and Product Quality. *Journal of Economic Theory*, **18**(2), 301–317.

Myerson, Roger B. 1981. Optimal Auction Design. *Mathematics of Operations Research*, **6**(1), 58–73.

Strausz, Roland. 2006. Deterministic versus Stochastic Mechanisms in Principal–Agent Models. *Journal of Economic Theory*, **128**(1), 306–314.