An Invitation to Mechanism Design

Justin Hadad[*]

Last edit: June 22, 2024

Before I studied it in graduate school, *mechanism design* was elusive to me. It (a) appeared to cover material that was essential "economics" while (b) not having immediately-accessible introductory content. Additionally, (c) people within the discipline colloquially looped in mechanism design with my study of *market design*, and it wasn't immediately obvious to me how to loop the two together. I'm writing this note to convey an introduction to the field. *Economic theory is cool*, I think, and mechanism design is at the heart of its modern wave.

At the center of economics are *prices* and *allocations*. For a moment, forget the idea that some seller just sets a price for a good and allocates it to whoever can pay it. Is there a better way? Specifically, I mean, is there a better *allocation rule* (a probability of allocation to a buyer with a certain value for the good) and *payment rule* than just price-setting?

The field of *mechanism design* gives precise results as to what allocations are possible, given agents' values are unknown to the seller. This note considers models with one agent and one seller, who would like to sell a good for which she has zero value.[1]

We differentiate agents by their values; intuitively, we say there are different *types* of agents. A little unintuitively, we say that an agent's *type* is just her value.

Without any loss of generality, we say that the agent has a type t valued on the interval $[0, 1]$; we write this $t \in [0, 1]$. Define an allocation rule $q(t) := [0, 1] \to [0, 1]$ as a mapping from the type space to the probability space, that a good is allocated given the agent has type t. This answers, *Given an agent's type, with what probability should she expect to be allocated the good?* Then define the payment rule as a vector $p(t) := [0, 1] \to \mathcal{R}$, equivalently answering, *what payment should an agent with type t expect to make?*

[*] University of Oxford, Department of Economics. The economic content here was taught to me by Ludvig Sinander; I additionally read and found useful Borgers et al (2015). The title is a nod to the article "An Invitation to Market Design" by two of my research supervisors.

[1] We generally say the seller has "zero marginal cost" for the good. The seller having some value $r > 0$ for the good has a direct relationship to the reserve price r that can be used when designing the optimal auction mechanism.

The expected utility of the seller is given by the payment $p(T)$, where T is the random variable associated with the unknown type of the buyer. Given she reports some type $r \in [0,1]$, the buyer receives utility $tq(r) - p(r)$; that is, she receives utility equal to the probability of the allocation times her true value, minus her payment from said report. In this way we assume that utility is directly transferable between both agents; utility loss $p(r)$ for the buyer corresponds exactly to the utility gain $p(r)$ for the seller.

We think of a *mechanism* as a pair of an allocation and a payment rule, $(q(t), p(t))$, or for short, (q, p). There's a famous theorem—commonly called the *revelation principle*—which allows us to focus on certain classes of mechanisms in our analysis. These mechanisms are the *incentive compatible* (IC) mechanisms, for which agents have the weakly dominant strategy to report their true type. This means, then, that buyer utility is always $tq(t) - p(t)$, which supremely simplifies our analysis.

We call a *direct revelation mechanism* (DRM) a setting where the agent is asked to make a report $r \in [0,1]$ of her type, and is then given the good with probability $q(r)$ and pays $p(r)$.

Theorem 1 (Revelation Principle)**.** *If the allocation and payment rule (q,p) are induced by some mechanism, then (q,p)—viewed as a DRM—is IC.*

Proof. Fix an allocation and payment rule (q,p). Because the allocation and payment rule are induced by some mechanism, each agent makes an optimal report according to their type. Now consider the DRM (q,p). The only deviations are to mimic the reports of other types $r \neq t$, which we just showed is weakly dominated. So the mechanism is IC. □

Define $V(t) := tq(t) - p(t)$ as the buyer's payoff from a truthful report, and define $f(r,t) := tq(r) - p(r)$, where $r \in [0,1]$ is the agent's report to the mechanism and t is her type. That a mechanism is IC implies that it additionally is *locally* IC: agents have no incentive to mimic *nearby* types. Hence $\frac{\partial}{\partial m} f(t+m,t)\big|_{m=0} := f_1(t,t) = 0$. Now consider the derivative of $V(t)$, which evaluates as

$$V'(t) = f_1(t,t) + f_2(t,t)$$

by the chain rule. Integrating upward and plugging in our local-IC condition gives

$$V(t) = V(0) + \int f_2(s,s)\mathrm{d}s.$$

After plugging in for our definition of $V(t)$, it's clear that any IC mechanism satisfies the following *envelope formula*:

$$tq(t) - p(t) = -p(0) + \int_0^t q$$

for every $t \in [0, 1]$. That every IC mechanism satisfies the envelope formula is the premise of the *envelope theorem.*

Theorem 2 (Mirrlees Envelope Theorem)**.** *Every IC mechanism satisfies the envelope formula.*

We can go further: An IC mechanism implies the agent receives weakly negative profit when misreporting her type. So it must be that $V(t) - f(r, t) \geq 0$. Some manipulation allows us to characterize what is necessary for this condition to hold:

$$\begin{aligned} V(t) - f(r, t) &= V(t) - V(r) + [rq(r) - p(r)] - [tq(r) - p(r)] \\ &= \int_r^t q(s)\mathrm{d}s - (t - r)q(r) \\ &= \int_r^t [q(s) - q(r)]\mathrm{d}s. \end{aligned}$$

Thus if a mechanism is IC, then it additionally satisfies the envelope formula and has q increasing in the type (else there would be finite intervals along (r, t) for which deviations could be profitable). If a mechanism has an allocation increasing in type and satisfies the envelope formula, the mechanism is IC. The reverse also holds.

Theorem 3 (Spence-Mirrlees Lemma)**.** *A mechanism is IC iff it satisfies the envelope formula and q is increasing.*

Without loss we can focus on mechanisms that induce participation for all types; if type t were not participating, then we could invite her to participate and award her the outcome $(q(t), p(t)) = (0, 0)$. We say a mechanism is *individually rational* (IR) if every type gets weakly positive payoff.

Theorem 4 (Spence-Mirrlees Lemma Corollary)**.** *A mechanism is IC and IR iff it satisfies the envelope formula, q is increasing, and $p(0) \leq 0$.*

A proof is straightforward; clearly it is optimal to set $p(0) = 0$.

Now that we have characterized what an IC mechanism demands (an increasing q that satisfies the envelope formula), we can formalize the seller problem as choosing a mechanism to maximize her revenue $R(q)$, defined as follows:

$$R(q) := \mathrm{E}[p(T)] = \mathrm{E}\left[Tq(T) - \int_0^T q\right].$$

Call Q the set of increasing allocations $q : [0,1] \to [0,1]$ that satisfy the envelope formula. Note Q is convex (consider that the convex combination of an increasing function is an increasing function) and compact (it is closed and bounded), and R is linear (therefore convex) and continuous. Then we use the following fact to pin down what mechanism the seller should choose to maximize her profit.

Proposition 1. *Any convex and suitably continuous function defined on a convex and suitably compact space achieves a maximum at an extreme point.*

Proof. Say $q \in Q$ maximizes $R(q)$. Then by convexity we can write $q = \int_{\text{ext } Q} q' \mu \mathrm{d}q'$ where ext Q are the extreme points of Q. Then we can write $\phi(q) \le \int_{\text{ext } Q} \phi(q')\mu \mathrm{d}q'$ by Jensen's inequality, where ϕ is a convex and suitably continuous function. But then $\phi(q) \le \phi(q')$, so $q' \in \text{ext } Q$ must also maximize $R(q)$. □

Thus $R(q)$ is maximized at an extreme point of Q. But what are the extreme allocations? They are the following:

$$q(t) = \begin{cases} 0 & \text{for } t < t^\star \\ 1 & \text{for } t > t^\star \end{cases} \quad \text{and} \quad q\left(t^\star\right) \in \{0,1\} \quad \text{for some } t^\star \in [0,1].$$

A proof that functions of the above form (Ludvig Sinander in his notes calls these *impulses*) are extreme points is straightforward: any allocation $q \in \{0,1\}$ cannot be written as a convex combination of two distinct elements within Q. By inspection, note the extreme allocations are merely posting a price, and the price can be pinned down via the envelope formula, which holds because the mechanism is IC!

This means that price-posting—literally setting a price, and letting the buyer purchase at that price or walk away—is optimal for the seller in the setting of one buyer and one seller. What a simple and fascinating result.

Mechanism design continues to investigate, among other things (·) if the result holds up with multiple goods to sell (not necessarily; what are the extreme points?); (·) if the results hold up with multiple buyers, and how to construct the optimal auction; (·) dynamic allocation; (·) commitment. Some more modern literature additionally changes the objective function for the seller—what if she chose to minimize regret (payment less the buyer's type, say) rather than maximize payoff?