Information Acquisition

Motivation

Until now, we have represented the information available to the agents by their privately known type. Indeed, the whole paradigm of mechanism design starts with the dispersed information held by different agents as a primitive of the economic situation. The task of the mechanism designer is to construct a game form or a mechanism that simultaneously collects the information and uses it efficiently to arrive at the desired allocation.

It should be clear that in many economically interesting settings, information is not exogenous. To the contrary, information is costly to come by and the amount of information acquisition depends on the economic value that the agents expect to get from it. Hence it is not clear why information ought to be seen as a primitive and independent of the final allocation mechanism. The payoffs from private information depend on the mechanism used, and as a consequence, the choice of the mechanism should have a feed-back effect to the information acquisition decisions.

Value of information

Consider the single-agent decision problem, where the agent chooses the optimal action $a \in A$ to maximize her expected utility. The utility depends on the state of nature $\theta \in \Theta$, and for the remainder of this lecture, we assume that $\Theta = \{\theta_1, ..., \theta_N\}$ is a finite set with $N$ elements. Assume also that the (state dependent) utility function $u(a, \theta)$ is continuous in $a$ for all $\theta$ and $A$ is compact. Let $p(i)$ denote the prior probability of the event $\{\theta = \theta_i\}$, and $p = (p(i), ..., p(N))$. We assume w.l.o.g. that $p(i) > 0$ for all $i$. The decision maker’s problem is then to

$$\max_a \sum_{i=1}^{N} p(i)u(a, \theta_i).$$

Let $a(p)$ denote the maximizer in the above program, and let

$$V(p) = \sum_{i=1}^{N} p(i)u(a(p), \theta_i)$$

be the value function of the program.

Observation 1.

$V(p)$ is convex in $p$. 

...Think back to the properties of profit function... This observation follows immediately from the observation that $V(p)$ is the maximum over a family of linear functions.

**Corollary**

If there are two states, then $p \in [0, 1]$. Therefore, Rothschild and Stiglitz from first year Micro implies that Mean preserving spreads are good for the decision maker.

The best the decision maker could hope for is to know the true state. Then $p \in 0, 1$. Clearly the perfectly revealing statistical experiment is the best for the decision maker (since then she can take optimal actions state by state). Loosely speaking, an experiment is a random variable whose outcome is correlated with the true state of the world. After seeing the outcome in the experiment, the decision maker updates her beliefs on the state and then chooses the optimal action. The above reasoning told us that one experiment is better than a second experiment if the posterior belief resulting from the first is second order stochastically dominated by the belief resulting from the second. Or if the posterior from the first is a mean preserving spread of the posterior from the second.

In this lecture, we shall see how to generalize this comparison to the case where there are multiple states (and hence mean preserving spreads are not well defined. Blackwell’s theorem shows that one experiment is more valuable to all decision makers than a second experiment if and only if the first experiment is statistically sufficient for the second.

**Statistical experiments and sufficiency**

An experiment is a random variable correlated with the state of the world. Here we will consider two arbitrary experiments $X^1$ and $X^2$. Experiment $X^j$ takes values in $1, ..., K^j$. Denote probability of outcome $k^j$ in experiment $j$ conditional on the state $\theta_i$ by $\pi^j(k^j | i)$. Assume w.l.o.g. that for all $k^j$, there is an $i$ such that $\pi^j(k^j | i) > 0$. Then the marginals on all signals are strictly positive:

$$\pi^j(k^j) \equiv \sum_i \pi^j(k^j | i)p(i) > 0.$$  

As a result, we can write Bayes’ rule:

$$p^j(i | k^j) = \frac{\pi^j(k^j | i)p(i)}{\pi^j(k^j)}.$$ 

Let $\Pi^j(k^j | i)$ denote the matrix where the $i^{th}$ column is $\pi^j(k^j | i)$.

**Definition 1.** Experiment $X^1$ is sufficient for $X^2$ ($X^1 \geq X^2$) if there exists a positive $K^2 \times K^1$ matrix $B$ with a typical element $b_{k^2 k^1}$ such that

$$\Pi^2(k^2 | i) = B\Pi^2(k^2 | i,).$$
and
\[ \sum_{k^2} b_{k^2 k^1} = 1 \text{ for all } k^1. \]

The interpretation of sufficiency is rather immediate. \( X^1 \) is sufficient for \( X^2 \) if you can get the conditional probability of the signals in experiment \( X^2 \) through a two-stage procedure. First observe \( X^1 \) and the resulting \( k^1 \). Perform an additional randomization (that is independent of \( \theta \) and sends the original signals onto \( 1, \ldots, K^2 \). It is clear that observing only the signals after the second stage cannot improve information on \( \theta \) since the second randomization was by assumption independent of \( \theta \). Hence there should be at least as much statistical information about \( \theta \) in \( X^1 \) as in \( X^2 \).

Valuable information

We could think that experiment \( X^1 \) is more valuable than experiment \( X^2 \) if for all utility functions \( u(a, \theta) \), the optimization after observing signals \( k^1 \) result in a higher expected payoff than optimization after observing \( k^2 \). More formally, we define:

**Definition 2.** Experiment \( X^1 \) is more informative than experiment \( X^2 \) (\( X^1 \succeq X^2 \)) if:

\[ \sum_{k^1} \pi^1(k^1) u(a(p^1 \mid k^1), \theta) \geq \sum_{k^2} \pi^2(k^2) u(a(p^2 \mid k^2), \theta). \]

for all \( A \) and all \( u(a, \theta) \).

Blackwell’s theorem

Blackwell’s Theorem connects the above two definitions:

**Theorem 1.** \( X^1 \) is more informative than \( X^2 \) if and only in \( X^1 \) is sufficient for \( X^2 \).

To make the proof simple, interpret all matrices as long vectors. For example, \( \Pi^j(k^j \mid i) \) should be understood as a \( K^j \times n \) vector where the columns \( \pi(k^j \mid i) \) are stacked on top of one another. (I.e. first the conditional on signals given state \( \theta_1 \), then conditional on \( \theta_2 \), and so on until conditional distribution on signals given \( \theta_N \).) We denote this vector by \( \pi(\cdot \mid \cdot) \). Similarly \( b_{k^2 k^1} \) is the \( K^2 \times K^1 \) positive vector where the columns of \( B \) are stacked on top of each other. Let \( B \) denote the set of all such vectors, where \( \sum_{k^2} b_{k^2 k^1} = 1 \). Finally, let \( D \) denote the set of \( d \in \mathbb{R}^{K^2 \times N} \) (again to be read as the columns of the matrix formed by \( d \) stacked on top of each other) such that

\[ \sum_{k^1} b_{k^2 k^1} \pi^1(k^1 \mid i) = d_{k^2 i} \text{ for some } b_{k^2 k^1} \in B. \]

With this notation, we begin the proof.
Proof. The following are equivalent:

1. $X^1$ is sufficient for $X^2$.

2. $\pi^2 \in \mathcal{D}$.

3. For all $q \in \mathbb{R}^{K^2 \times N}$, there exists $b_{k2k1} \in \mathcal{B}$ such that

\[
\sum_{k^2, i} q_{k^2 i} \pi^2(k^2 \mid i) \leq \sum_{k^2, i} q_{k^2 i} \sum_{k^1} b_{k2k1} \pi^1(k^1 \mid i)
\]

\[= \sum_{k^1} \sum_{k^2} b_{k2k1} \sum_{i} q_{k^2 i} \pi^1(k^1 \mid i).\]

4. For all $q \in \mathbb{R}^{K^2 \times N}$,

\[
\sum_{k^2, i} q_{k^2 i} \pi^2(k^2 \mid i) \leq \max_{k^2} \sum_{i} q_{k^2 i} \pi^1(k^1 \mid i) \]

5. For all $\phi \in \mathbb{R}^{K^2 \times N}$,

\[
\sum_{k^2} \pi^2(k^2) \sum_{i} \phi_{k2i} p^2(i \mid k^2) \leq \sum_{k^1} \pi^1(k^1) \max_{k^2} \sum_{i} \phi_{k2i} p^1(i \mid k^1).
\]

1 $\iff$ 2 by definition of sufficiency. 2 $\iff$ 3 by separating hyperplane theorem. 3 $\iff$ 4 follows since $\sum_{k^1} \max_{k^2} \sum_{i} q_{k^2 i} \pi^1(k^1 \mid i)$ maximizes $\sum_{k^1} \sum_{k^2} b_{k2k1} \sum_{i} q_{k^2 i} \pi^1(k^1 \mid i)$ over $b_{k2k1} \in \mathcal{B}$. 4 $\iff$ 5 by change of variables: $q_{k2i} = p(i)\phi_{k2i}$.

With these preliminary results, we can complete the proof.

more informative implies sufficiency

By contrapositive. If $X^1$ is not sufficient for $X^2$, then item 5 above does not hold, and there is a $\phi \in \mathbb{R}^{K^2 \times N}$ such that

\[
\sum_{k^1} \pi^1(k^1) \max_{k^2} \sum_{i} \phi_{k2i} p^1(i \mid k^1)
\]

\[< \sum_{k^2} \pi^2(k^2) \sum_{i} \phi_{k2i} p^2(i \mid k^2)
\]

\[\leq \sum_{k^2} \pi^2(k^2) \max_{\lambda^2} \sum_{i} \phi_{k2i} p^2(i \mid k^2)
\]
Then for $A = K^2$ and $u(k^2, \theta_i) = \phi_{k^2 i}$, we have

$$\sum_{k^1} \pi^1(k^1) u(a(p^1 | k^1), \theta) < \sum_{k^2} \pi^2(k^2) u(a(p^2 | k^2), \theta).$$

**sufficiency implies more informative**

Let $\phi_{k^2 i} = u(a(p^2(\cdot | k^2)), \theta_i)$ and conclude from item 5 above that the value in $X^1$ must be at least as high as in $X^2$.

\[
\square
\]

**Comments**

Blackwell’s theorem is a fantastic result conceptually. At the same time, its message is somewhat disappointing. The resulting order on experiments is very incomplete. In other words, for most experiments $X, Y$ neither $X$ is more informative than $Y$ nor is $Y$ more informative than $X$. This follows from the fact that statistical sufficiency is not easily satisfied. Another problem with the Theorem is that in practice, it is of little help for decision theory. It is notoriously difficult to check whether one experiment is sufficient for another. In fact, the theorem is often useful in showing that sufficiency fails by constructing an appropriate decision problem. For economics, this is of course not very helpful.

Consider an example to see this. Suppose that $X = \theta + \epsilon_X$ and $Y = \theta + \epsilon_Y$, where $\theta$ is uniformly distributed on $[-M, M]$, $\epsilon_X$ is uniformly distributed on $[-a, a]$, and $\epsilon_Y$ is uniformly distributed on $[-b, b]$, and $a < b$. Intuitively, one would think that $X$ is more informative than $Y$, nonetheless neither of these experiments is more informative than the other. I leave it as a (very good and somewhat challenging) exercise to verify this. On the other hand, it should be noted that adding another copy of an independent experiment leads to a more informative experiment.

In the next section, we will search for tighter orders by imposing additional structure on the decision problems. The approach is thus similar to that taken in expected utility theory when looking for various notions of stochastic orders. (Recall there that FOSD allowed for all increasing utility functions whereas SOSD was defined when utility functions were restricted to increasing and concave functions).
Lehmann’s order

Lehmann’s condition

Consider two experiments $X, Y$ on a state space $\Theta$. Denote the conditional distribution of signals (i.e. outcomes of the experiment) for $X$ by $F(x | \theta)$ and the distribution of signals in $Y$ by $G(y | \theta)$. Select two states $\theta_1, \theta_2$ with $\theta_1 < \theta_2$. Plot the following parametric curves in the unit square:

$$(F(x, \theta_1), F(x, \theta_2)) \text{ and } (G(y, \theta_1), G(y, \theta_2)).$$

Clearly both of these curves start at the lower left corner of the square and end at the higher right corner. If we assume MLRP, then the curves start with a slope less than 1 and the slopes increase as the parameter $x$ or $y$ grows. This is seen easily by differentiating with respect to the parameter.

Lehmann’s condition for $X$ to be more accurate than $Y$ is that the graph of $(F(x, \theta_2), F(x, \theta_1))$ lie below $(G(y, \theta_2), G(y, \theta_1))$ for all choices of $\theta_1 < \theta_2$. An equivalent statement of this is that for all $\theta_1 < \theta_2$,

$$F(F^{-1}(p; \theta_1); \theta_2) \leq G(G^{-1}(p; \theta_1); \theta_2).$$

Another equivalent way of stating Lehmann’s condition is that for all $x, y$ in the support of $F, G$ respectively,

$$F(x; \theta) - G(y; \theta)$$

has at most one change of sign from negative to positive as $\theta$ increases.

Finally, it should be noted that Lehmann’s condition is equivalent to Blackwell’s condition when applied to dichotomies (i.e. two-point sets of states).

Restriction on preferences and Persico’s theorem

Definition 3. A utility function $u(a, \theta)$ satisfies the single crossing property if for each pair of decisions $a, a' \in A$ with $a < a'$, and $\theta < \theta'$,

$$u(a', \theta) - u(a, \theta) \geq 0 \Rightarrow u(a', \theta) - u(a, \theta) \geq 0.$$  

The single crossing property is said to be strong if the above equation holds with strict inequalities.
It should be clear that supermodular preferences satisfy single crossing. This class of preferences is also important in the literature on monotone comparative statics. Milgrom and Shannon, Econometrica, 1994, explore conditions that guarantee monotone comparative statics results for decision problems and for games. Athey, QJE, 2001, extends this analysis to decisions under uncertainty, and papers Athey, Reny, McAdams (all in Econometrica, early 2000’s) and others have used these notions in the study of existence of equilibria in monotone pure strategies in Bayesian games.

For our purposes the following Theorem (due to Nicole Persico) is of most importance.

**Theorem 2.** Consider two experiments \( X, Y \) that satisfy the MLRP and that have conditional distributions \( F(x; \theta), G(x; \theta) \). Then the following two statements are equivalent.

1. \( X \) is more accurate than \( Y \) in the sense of Lehmann’s accuracy.

2. For all prior probabilities \( \mu \) on \( \Theta \) and for all utility functions \( u(a, \theta) \) satisfying the single crossing property,

\[
\int_{\theta \in \Theta} \int u(a(x), \theta) dF(x; \theta)d\mu(\theta) \geq \int_{\theta \in \Theta} \int u(a(y), \theta) dG(y; \theta)d\mu(\theta).
\]

Hence a Bayesian decision maker with single crossing preferences values information that is accurate in the sense of Lehmann. We end our discussion on this topic with a decision theoretic example (due to Athey and Levin) of information acquisition under this class of preferences.

**Example 3.** This example examines the demand of information for a monopolist producer.

Let \( P(q) \) be the inverse demand curve. Suppose the cost of producing \( q \) units is \( C(q, \theta) \), where (letting subscripts denote partial derivatives) \( C_q \) is nonincreasing in \( \theta \).

The monopolists payoff is \( u_M(q, \theta) = qP(q) - C(q, \theta) \), while the social planners payoff is

\[
u_S(q, \theta) = \int_0^q P(t)dt - C(q, \theta).
\]

Both payoff functions are supermodular, so consider a family of experiments about the true cost \( \theta \) whose signal distributions are indexed by a continuous parameter \( \sigma \), i.e., \( \{F^\sigma(x)\}_{\sigma \in \Sigma} \). Assume that for \( \sigma' > \sigma \), \( \{F^\sigma'(x)\} \) is more accurate than \( \{F^\sigma(x)\} \) in the sense of Lehmann (therefore satisfying MLRP). Assume also that the distribution functions are differentiable w.r.t. the parameter \( \sigma \).
Persico’s theorem implies that both the monopolist and the planner prefer more accurate signals to less accurate. To see who has a higher marginal value for information, we must consider the derivatives of the value functions $\frac{dV^M(\tilde{\sigma})}{d\sigma}$ and $\frac{dV^S(\tilde{\sigma})}{d\sigma}$ at a particular parameter value $\tilde{\sigma}$.

We want to show that the difference on the planner’s and monopolist’s objective function satisfies single crossing. To this effect, consider

$$\frac{d}{d\sigma}u^S(q^S(x), \theta) - \frac{d}{d\sigma}u^M(q^M(x), \theta).$$

Eliminating terms that do not depend on $\theta$, we express this difference as

$$-C_q(q^S(x), \theta)q^S_x(x) + C_q(q^M(x), \theta)q^M_x(x).$$

To check that this is non-decreasing in $\theta$ is an exercise in using the Implicit Function Theorem to find first expressions for $q^M(x)$ and $q^S(x)$ and then checking that the derivative of the above difference w.r.t. $\theta$ has the right sign. (The conditions in the end will unfortunately be nasty).