UPenn NETS 412: Algorithmic Game Theory March 27, 2018

Lecture 14

Lecturer: Bo Waggoner *Scribe: Bo Waggoner*

Information Elicitation Without Verification

In the previous lecture, we looked at proper scoring rules – basic tools for eliiting predictions from strategic agents. Now, we will use these as tools to elicit information from agents in situations where we cannot verify the reports.

Imagine there is a group of agents who have each observed some information. We would like to collect that information truthfully, but we cannot directly verify whether they are being truthful or not.

For instance, perhaps these agents have all gone to the same restaurant and observed some information about the restaurant's quality, such as good/bad or on a 1-to-5 scale. We want to elicit their true experiences, but we cannot go to all the restuarants ourselves to verify what we are told (and if we could, then we wouldn't need the agents in the first place).

As a second example, suppose we are asking a group of agents to label images for us (a common crowdsourcing task on the web). For each image, each agent observes some information about what is in the image, then makes a report. We do not have the resources to double-check all of the images, but we want to elicit truthful reports.

In all these cases, the idea is to have multiple agents completing the same task or reporting their own observations and score them against each other. If the others are truthful, then these agents will be truthful as well.

Modeling the game. In this setting, which we will call *information elicitation without verification*, we have:

- n agents $1, \dots, n$.
- For each agent i, a set of possible signals S_i and a utility function $u_i : S_1 \times \cdots \times S_n \to \mathbb{R}$.
- A common prior distribution $\mathcal{P}$ over all realizations of all of the signals. For instance $\mathcal{P}(s_1, \dots, s_n)$ is the probability that agent 1 observes s_1, ..., and agent n observes s_n.

The timing of the game is as follows:

1. The signals $s_1, \dots, s_n$ are drawn jointly from the distribution $\mathcal{P}$.
2. Each player i observes her signal s_i.
3. Each player i makes a report $\hat{s}_i$ possibly different from s_i.
4. Each player i gets a utility $u_i(\hat{s}_1, \dots, \hat{s}_n)$.

The question is: how do we design the utility functions $u_1, \dots, u_n$ so that players all want to report truthfully?

Solution concept. This game is a bit different from others we have seen because it involves the signals, which are randomly drawn. Each player is not completely certain what information the others hold. We need a new definition of equilibrium to capture this setting. This definition can be extended to more general games of *incomplete information*, but we will focus on information elicitation without verification.

In this game, the actions A_i of player i consist of all functions $a_i : S_i \to S_i$. In other words, player i decides on a function that takes in her observed signal s_i and outputs her reported signal $\hat{s}_i$. One function is the "truthful" function, which always reports the signal truthfully. Note that, given a strategy profile $a_1, \dots, a_n$, player i's expected utility can be written $\mathbb{E}\, u_i(a_1(s_1), \dots, a_n(s_n))$ where the expectation is taken over $s_1, \dots, s_n \sim \mathcal{P}$.

Definition 1 *A strategy profile $a_1, \ldots, a_n$ is a **Bayes-Nash equilibrium** if, for all players i and all $a_i' \neq a_i$,*

$$\mathop{\mathbb{E}}_{s_1 \ldots s_n \sim \mathcal{P}} u_i(a_1(s_1), \ldots, a_n(s_n)) \geq \mathop{\mathbb{E}}_{s_1 \ldots s_n \sim \mathcal{P}} u_i(a_{-i}(s_{-i}), a_i'(s_i)).$$

*It is a **strict** equilibrium if the inequality is always strict for $a_i' \neq a_i$.*

*It is a **truthful** equilibrium if every a_i is truthful, i.e. the identity function.*

Example and output agreement. A simple example is *output agreement*. Suppose there are two players who each observe the same signal, for instance, they are both asked to provide a label for an image. They are both rewarded with 1 if they report the same signal, and they both get 0 otherwise.

This might sound like a reasonable approach. For example, consider the restaurant setting. Let's say there are two agents, Alice and Bob. And say there are two signals, Like and Dislike. Alice and Bob each go to the restaurant and observe a signal, then report it. Under output agreement, they each get 1 if they both report the same thing, otherwise 0. In particular, for both $i = \text{Alice}$ and $i = \text{Bob}$ we have

$$u_i(\text{Like}, \text{Like}) = u_i(\text{Dislike}, \text{Dislike}) = 1$$
$$u_i(\text{Dislike}, \text{Like}) = u_i(\text{Like}, \text{Dislike}) = 0$$

But this is not always truthful! Consider the following prior distribution on signals:

- $\mathcal{P}(\text{Dislike}, \text{Dislike}) = 0.02$
- $\mathcal{P}(\text{Like}, \text{Dislike}) = \mathcal{P}(\text{Dislike}, \text{Like}) = 0.24$
- $\mathcal{P}(\text{Like}, \text{Like}) = 0.5$.

It's worth thinking about what this prior encodes. It says that we all expect that the restaurant will be quite good. In fact, with probability 0.5, we expect both agents to like the restaurant. Each player alone has a total probability of 0.74 of liking the restaurant and only 0.26 of disliking it.

Now we can compute the best responses for each player as follows. If Alice observes her signal is Like, then her posterior belief is, using Bayes' rule,

$$\begin{aligned}\Pr\left[s_{\text{Bob}} = \text{Like} \mid s_{\text{Alice}} = \text{Like}\right] &= \frac{0.5}{0.5 + 0.24} \\ &= \frac{25}{37}. \\ \Pr\left[s_{\text{Bob}} = \text{Dislike} \mid s_{\text{Alice}} = \text{Like}\right] &= \frac{0.24}{0.5 + 0.24} \\ &= \frac{12}{37}.\end{aligned}$$

Suppose Bob is reporting truthfully, i.e. Bob always reports the signal he observes. Is it a strict best response for Alice to report truthfully? When she observes the signal Like, she believes there is a $\frac{25}{37}$ chance Bob is reporting Like and only a $\frac{12}{37}$ chance Bob is reporting Dislike. So she gets expected payoff $\frac{25}{37}$ for reporting Like truthfully, and only $\frac{12}{37}$ expected payoff for misreporting Dislike instead. So when Alice observes Like, she best-responds by being truthful.

But what about when observing Dislike? Then her posterior belief is

$$\begin{aligned}\Pr\left[s_{\text{Bob}} = \text{Like} \mid s_{\text{Alice}} = \text{Dislike}\right] &= \frac{0.24}{0.24 + 0.02} \\ &= \frac{12}{13}. \\ \Pr\left[s_{\text{Bob}} = \text{Dislike} \mid s_{\text{Alice}} = \text{Dislike}\right] &= \frac{0.24}{0.24 + 0.02} \\ &= \frac{1}{13}.\end{aligned}$$

So if Bob is reporting truthfully, when Alice observes Dislike, she believes Bob has a $\frac{12}{13}$ chance of reporting Like. So Alice is better off misreporting Like!

We have shown by counterexample:

Fact 2 *Output agreement sometimes has no truthful equilibria.*

Peer prediction. We can design a truthful mechanism using our knowledge of proper scoring rules. The **peer prediction** mechanism, proposed by Miller, Resnick, and Zeckhauser in 2005, works as follows for two agents.

1. Alice reports a signal $\hat{s}_{\text{Alice}}$.
2. The mechanism calculates the posterior distribution p_{Alice} on Bob's signal, conditioned on $s_{\text{Alice}} = \hat{s}_{\text{Alice}}$.
3. Meanwhile, Bob reports a signal $\hat{s}_{\text{Bob}}$.
4. Alice's utility is
$$u_{\text{Alice}}(\hat{s}_{\text{Alice}}, \hat{s}_{\text{Bob}}) = S(p_{\text{Alice}}, \hat{s}_{\text{Bob}})$$
where S is a strictly proper scoring rule.
5. Symmetrically, the mechanism calculates Bob's posterior distribution p_{Bob} on Alice's signal conditioned on his report, and Bob's utility is
$$u_{\text{Bob}}(\hat{s}_{\text{Alice}}, \hat{s}_{\text{Bob}}) = S(p_{\text{Bob}}, \hat{s}_{\text{Alice}}).$$

We make a small assumption on the prior distribution of signals: Each of Alice's signals produces a different posterior belief on Bob's signals; and symmetrically for Bob.

Theorem 3 *Under the above assumption, the peer prediction mechanism has a strict, truthful Bayes-Nash equilibrium.*

Proof Fix a player, say Alice without loss of generality, and suppose the other player is playing truthfully. We must show that it is a best response for Alice to play truthfully, and any other strategy gives strictly less utility.

If Bob is reporting truthfully, then Alice's utility will be

$$u_i(\hat{s}_{\text{Alice}}, s_{\text{Bob}}) = S(p_{\text{Alice}}, s_{\text{Bob}}).$$

Because S is a strictly proper scoring rule, Alice maximizes this utility when p_{Alice} is equal to her true belief, i.e. her posterior distribution on s_{Bob} conditioned on her true signal s_{Alice}. Furthermore, plugging in any other value for p_{Alice} would give her strictly worse expected score.

But Alice can achieve this maximum utility only by reporting $\hat{s}_{\text{Alice}} = s_{\text{Alice}}$, i.e. reporting truthfully. In this case, her true belief will be plugged in to the scoring rule, but for any other report, a different p will be plugged in, resulting in strictly worse expected score. ■

Notice that this mechanism only worked for two agents. This is however quite easy to fix: we can split the agents up into pairs if there are more than two, or even do more fancy things such as choosing, for each Alice, a random "reference report" Bob and scoring her prediction on that report.

A more serious drawback with this method is that it needs to know the common prior distribution that the signals are drawn from. This might not always be realistic. There has been a significant amount of research on weakening this assumption either by requesting more information from the agents, by considering several different tasks simultaneously, or so on.