CIS 700 Differential Privacy in Game Theory and Mechanism Design January 31, 2014

Lecture 3

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Privacy as a Tool for Mechanism Design (for arbitrary objective functions) Without Money

1 Introduction

Today we return to the idea of differential privacy as a tool to be wielded in mechanism design. We will prove a simple, but remarkable theorem that deviates from our normal intuition as mechanism designers. Typically, we think of social welfare as a special objective function – we can always use the VCG mechanism to optimize social welfare, but in general, there do not exist truthful mechanisms to optimize arbitrary objective functions. Even with the welfare objective, we require mechanisms which can extract payments.

In this lecture we give a strictly dominant strategy truthful mechanism which can optimize arbitrary objective functions, and without payments. This is not in contradiction with known impossibility results, because we pay for this remarkable result – we do not optimize the objective exactly, but instead only approximately, with some additive loss. Nevertheless, this loss will generally become negligible in large economies.

Let us recall two features of the exponential mechanism from Lecture 1, one of which we have already remarked on, and one of which we have not:

1. Although we used the exponential mechanism to optimize for *revenue* in the first lecture and *social welfare* in the second lecture, it can equally well optimize for *arbitrary* low sensitivity objective functions. (Although it was important that it optimized for welfare in Lecture 2 to pair it with VCG payments).

2. The generic *approximate truthfulness* guarantee that it inherits by virtue of being differentially private *does not require payments.* Note that in the last two lectures, we used payments in different ways. In Lecture 1, we used payments to a) collect revenue, and b) guarantee that not *every* report was an approximate dominant strategy. In Lecture 2, we used payments to guarantee exact truthfulness. But there is more than one way to skin a cat[1]...

This combination makes it tempting to ask: can the exponential mechanism be used as a tool to design (exactly) truthful mechanisms for approximately optimizing arbitrary objective functions, without the use of money? Such general tools are rare in mechanism design: the Gibbard-Satterthwaite theorem tells us that in general settings, the only non-trivial deterministic truthful mechanisms are constant functions and dictator functions (in which the outcome is chosen as a function of only a single agent's report). In the lucky case when our objective is social welfare, we saw the VCG mechanism is a general tool, but it requires payments! In general, this is necessary.

In this lecture, however, we will give a general technique for truthfully optimizing arbitrary objective functions without payments. The tradeoff, as always, will be that we do not *exactly* optimize these functions, but only approximately, with some additive loss. However, the additive loss will generally become a diminishing fraction of the optimal objective value as the size n of the society grows large.

2 Making the Exponential Mechanism Exactly Truthful Without Money

We will work in the following setting:

[1] Non-native speakers: This is an odd English language idiom. Cat torture will not be essential to the content of this lecture.

Definition 1 (The Environment) *An environment is determined by:*

1. *A set N of n players.*
2. *A set of types T_i for each player $i \in N$.*
3. *A finite set S of social alternatives*
4. *A set of reactions R_i for each player $i \in N$.*
5. *A utility function $u_i : T_i \times S \times R_i \to [0,1]$ for each agent i.*

We write T_{-i} for $\prod_{j \neq i} T_i$ and $t_{-i} \in T_{-i}$. Write $r_i(t, s, \hat{R}_i) \in \arg\max_{r \in \hat{R}_i} u_i(t, s, r)$ to denote i's optimal reaction to type t and alternative s among choices $\hat{R}_i \subseteq R_i$.

A direct revelation mechanism $\mathcal{M}$ defines a game which is played as follows:

1. Each player i reports a type $t'_i \in T_i$.
2. The mechanism chooses an alternative $s \in S$ and a subset of reactions for each player $\hat{R}_i \subseteq R_i$.
3. Each player chooses a reaction $r_i \in \hat{R}_i$ and experiences utility $u_i(t_i, s, r_i)$.

Note that this setting is slightly unusual in the following sense: agents interact with the mechanism *twice*: first, they report their type. Then, the mechanism chooses not just a set of outcomes, but also a feasible set of *reactions*, and then the players choose a reaction from among that set. A player's utility depends on just on the outcome, but on what reaction he takes in response to it. Importantly, here, the mechanism has the power to *limit* the set of reactions that the player may choose from.

This isn't unreasonable though, nor as different from the standard setting as it seems at first. In fact, standard auction settings can be phrased in this way. For example, in the first lecture, we gave a differentially private mechanism that picks a sales price p as a function of player reports. Then, as is usually the case in auctions, we obligated any player who reported a value greater than this price to actually follow through with the purchase of the item at price p. If you like, the price p was the outcome chosen by the mechanism. Given the outcome, each agent had two possible reactions: buy the good, or not buy it – and their utility was a function both of the outcome (the purchase price), and whether or not they decided to buy (their chosen reaction). Finally, by obligating agents to buy the item if the price was below their reported valuation, the mechanism used its ability to restrict the set of allowable reactions. Restricted reaction sets are relevant in other settings as well. For example (related to an application we will see), suppose a local government is deciding where to build a collection of schools, as a function of reported citizen addresses. After the schools are open, the city can restrict each citizen to be able to use only the school that is closest to his *reported* address, and not a school that may instead be closer to his *actual* address.

Note that since there is no further interaction after the 3rd step, rational agents will always pick the reaction that maximizes their utility at the 2nd step:

$$r_i = r_i(t_i, s, \hat{R}_i),$$

and so we can ignore this as a strategic step.

Let $\mathcal{R}_i = 2^{R_i}$ and let $\mathcal{R} = \prod_{i=1}^{n} \mathcal{R}_i$. Then a mechanism is a randomized mapping $\mathcal{M} : T \to S \times \mathcal{R}$. We denote agents expected utilities for reporting a type t'_i when all other agents report type t'_{-i} as:

$$u(t_i, \mathcal{M}(t'_i, t'_{-i})) = \mathrm{E}_{s, \hat{R}_i \sim \mathcal{M}(t'_i, t'_{-i})}[u(t_i, s, r_i(t_i, s, \hat{R}_i))]$$

We want to design mechanisms that incentivize truthful reporting, but don't require payments...

We will say that:

Definition 2 *A mechanism $\mathcal{M}$ is* η-strictly dominant strategy truthful *if for all $i \in N$, $t_i \in T_i$ and $t'_{-i} \in T_{-i}$:*

$$u(t_i, \mathcal{M}(t_i, t'_{-i})) \geq u(t_i, \mathcal{M}(t'_i, t'_{-i})) + \eta$$

Finally, we will be interested in maximizing arbitrary objective functions $F : T \times S \times R \to \mathbf{R}$. We will normalize these objective functions to take values in $[0, 1]$ For example, our old favorite – social welfare – is:

$$F(t, s, r) = \frac{1}{n}\sum_{i=1}^{n} u_i(t_i, s, r_i)$$

Definition 3 *A mechanism $\mathcal{M}$ α-approximates an objective F if for all t:*

$$\mathrm{E}_{s,\hat{R}\sim\mathcal{M}}[F(t, s, r(t, s, \hat{R}))] \geq \max_{t,s,r} F(t, s, r) - \alpha$$

Ok! Now we can set about designing mechanisms. First lets consider *unrestricted* mechanisms that always output $\hat{R}_i = R_i$. We've already got a terrific one – the exponential mechanism.

Recalling the theorem:

Theorem 4 *$\mathcal{M}_\epsilon$ is ϵ-approximately truthful and α-approximates any objective F for:*

$$\alpha = O\left(GS(F) \cdot \frac{\log |S|}{\epsilon}\right)$$

We want to make the exponential mechanism truthful – but recall we can't use payments. The idea will be simple. Using the exponential mechanism, we get a good approximation to our objective – and agents have at most a *small* incentive to deviate. Suppose we had some other mechanism that perhaps did not get a good approximation to our objective, but gave agents a *strict* incentive to truth-tell. Then, there exists some lottery between the two mechanisms such that their resulting combination is exactly dominant strategy truthful! If the lottery still puts substantial weight on the exponential mechanism, then we will inherit much of its objective guarantee.

Here is one such strictly truthful mechanism which is very simple, but not necessarily the best for a given problem:

Definition 5 *The commitment mechanism $M^P(t')$ selects $s \in S$ uniformly at random and sets $\hat{R}_i = \{r_i(t'_i, s, R_i)\}$. i.e. it picks a random outcome, and then forces everyone to react as if their reported type is their true type.*

Define the *gap* of an environment as:

$$\gamma = \min_{i, t_i \neq t'_i, t_{-i}} \max_{s \in S} (u_i(t_i, s, r_i(t_i, s, R_i)) - u_i(t_i, s, r_i(t'_i, s, R_i)))$$

i.e. γ is a lower bound over players and types of the worst-case cost (over s) of mis-reporting. Note that for each player, this worst-case is realized with probability at least $1/|S|$. Therefore we have the following simple observation:

Lemma 6 *For all i, t_i, t'_i, t_{-i}:*

$$u(t_i, \mathcal{M}^P(t_i, t_{-i})) \geq u(t_i, \mathcal{M}^P(t'_i, t_{-i})) + \frac{\gamma}{|S|}$$

Proof For any player i, let $s^* = \arg\min_{t'_i \neq t_i, t_{-i}} \max_{s \in S} (u_i(t_i, s, r_i(t_i, s, R_i)) - u_i(t_i, s, r_i(t'_i, s, R_i)))$. Then for any deviation t'_i:

$$\begin{aligned} u(t_i, \mathcal{M}^P(t_i, t_{-i})) &= \sum_{s \in S} \frac{1}{|S|} \cdot u(t_i, s, r_i(t_i, s, R_i)) \\ &= \frac{1}{|S|} \left(\sum_{s \neq s^*} (u(t_i, s, r_i(t_i, s, R_i))) + u(t_i, s, r_i(t_i, s^*, R_i)) \right) \\ &\geq \frac{1}{|S|} \left(\sum_{s \neq s^*} (u(t_i, s, r_i(t'_i, s, R_i))) + u(t_i, s, r_i(t'_i, s^*, R_i)) + \gamma \right) \\ &= u(t_i, \mathcal{M}^P(t'_i, t_{-i})) + \frac{\gamma}{|S|} \end{aligned}$$

■

In other words, the commitment mechanism is strictly truthful: every individual has at least a $\frac{\gamma}{|S|}$ incentive not to lie.

This suggests a way to achieve an exactly truthful mechanism that also gets good objective guarantees:

Definition 7 *The punishing exponential mechanism $\mathcal{M}^P_\epsilon(t)$ defined with parameter $0 \leq q \leq 1$ is:*

1. *With probability $(1-q)$ return $\mathcal{M}_\epsilon(t)$*
2. *With probability q return $\mathcal{M}^P(t)$.*

We can calculate for which values of q $\mathcal{M}^P_\epsilon(t)$ is truthful:

Theorem 8 *If $\epsilon \leq \frac{q\gamma}{(1-q)|S|}$ then $\mathcal{M}^P_\epsilon$ is strictly truthful.*

Proof Observe that by linearity of expectation, we have for all t_i, t'_i, t_{-i}:

$$\begin{aligned} u_i(t_i, \mathcal{M}^P_\epsilon(t_i, t_{-i})) &= (1-q) \cdot u_i(t_i, \mathcal{M}_\epsilon(t_i, t_{-i})) + q \cdot u_i(t_i, \mathcal{M}^P(t_i, t_{-i})) \\ &\geq (1-q)\left(u_i(t_i, \mathcal{M}_\epsilon(t'_i, t_{-i})) - \epsilon\right) + q\left(u_i(t_i, \mathcal{M}^P(t'_i, t_{-i})) + \frac{\gamma}{|S|}\right) \\ &= u_i(t_i, \mathcal{M}^P_\epsilon(t'_i, t_{-i})) - (1-q)\epsilon + q\frac{\gamma}{|S|} \end{aligned}$$

Setting $q\frac{\gamma}{|S|} > (1-q)\epsilon$ and solving for ϵ gives the theorem. ■

Remark This condition is satisfied whenever we set $q = \frac{|S|\epsilon}{\gamma + |S|\epsilon}$.

Note that we also have utility guarantees for this mechanism. Setting the parameter q so that we have a truthful mechanism:

$$\begin{aligned} \mathrm{E}_{s,\hat{R} \sim \mathcal{M}^P_\epsilon}[F(t, s, r(t, s, \hat{R}))] &\geq (1-q) \cdot \mathrm{E}_{s,\hat{R} \sim \mathcal{M}_\epsilon}[F(t, s, r(t, s, \hat{R}))] \\ &= \left(1 - \frac{|S|\epsilon}{\gamma + |S|\epsilon}\right) \cdot \mathrm{E}_{s,\hat{R} \sim \mathcal{M}_\epsilon}[F(t, s, r(t, s, \hat{R}))] \\ &= \left(1 - \frac{|S|\epsilon}{\gamma + |S|\epsilon}\right) \cdot \left(\max_{t,s,r} F(t, s, r) - O\left(GS(F) \cdot \frac{\log |S|}{\epsilon}\right)\right) \\ &\geq \max_{t,s,r} F(t, s, r) - \frac{|S|\epsilon}{\gamma + |S|\epsilon} - O\left(GS(F) \cdot \frac{\log |S|}{\epsilon}\right) \end{aligned}$$

Picking ϵ to minimize this expression, we find:

$$\mathrm{E}_{s,\hat{R}\sim\mathcal{M}_\epsilon^P}[F(t,s,r(t,s,\hat{R}))] \geq \max_{t,s,r} F(t,s,r) - O\left(\sqrt{GS(F)\cdot\frac{|S|\log|S|}{\gamma}}\right)$$

Therefore, we have shown:

Theorem 9 *There is a strictly dominant strategy truthful mechansim that does not use payments, and for any objective function F, α-approximates F for:*

$$\alpha = O\left(GS(F)\cdot\sqrt{\frac{|S|\log|S|}{\gamma}}\right)$$

e.g. if $GS(F) = 1/n$ (as it is for average social welfare:

$$\alpha = O\left(\sqrt{\frac{|S|\log|S|}{\gamma n}}\right)$$

which tends to zero as $n \to \infty$

We did of course need that $\gamma > 0$.

3 An application: Facility Location

Lets now consider an application of this framework: the problem of school location that we alluded to earlier. Suppose that a city wants to build k schools to minimize the average distance between each citizen and their closest school. To simplify matters, we make the mild assumption that the city is built on a discretization of the unit line. Formally, for all i let:

$$L(m) = \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$$

denote the discrete unit line with step-size $1/m$. $|L(m)| = m+1$. Let $T_i = R_i = L(m)$ for all i and let $|S| = L(m)^k$. Define the utility of agent i to be:

$$u_i(t_i, s, r_i) = \begin{cases} -|t_i - r_i|, & \text{If } r_i \in s; \\ -1, & \text{otherwise.} \end{cases}$$

Note that $r_i(t_i, s)$ is here the closest facility $r_i \in s$.

We can instantiate Theorem 9. Note that in our case, we have: $|S| = (m+1)^k$, and we can compute the gap:

$$\gamma \geq 1/m$$

This is because for any true type t_i, and reported type t_i', if s is the outcome that places a single facility at t_i and all remaining facilities at t_i'.

We can now use our tools to optimize over any objective function we want! For the social welfare objective (or any other $1/n$ sensitive objective):

$$(u_i(t_i, s, r_i(t_i, s, R_i)) - u_i(t_i, s, r_i(t_i', s, R_i))) = 0 + |t_i - t_i'| \geq \frac{1}{m}$$

Theorem 10 *M_ϵ^P instantiated for the facility location game is strictly truthful and α-accurate for:*

$$\alpha = O\left(\sqrt{\frac{km(m+1)^k\log m}{n}}\right)$$

In fact, we do not need this exponential dependence in k which we inherit from the general theorem. Note that when we argued that $\gamma \geq 1/m$, we only needed a single kind of outcome s: For every pair t_i, t'_i, we needed an instance in which there are only schools on t_i and t'_i. Therefore, our punishment mechanism does not need to randomize over all $(m+1)^k$ possible outcomes, but only over these m^2 relevant ones. Therefore, we get a punishment mechanism which is $1/m^3$-strictly truthful, and a correspondingly the stronger theorem:

Theorem 11 M_ϵ^P *instantiated for the facility location game is strictly truthful and α-accurate for:*

$$\alpha = O\left(\sqrt{\frac{k \cdot m^3 \log m}{n}}\right)$$

An even more careful analysis can remove another factor of $\sqrt{m}$.

In either case, the approximation quickly becomes exact as the population size n grows – and we have strict dominant strategy truthfulness always.

Bibliographic Information The contents of this lecture are taken entirely from Nissim, Smorodinsky, and Tennenholtz, "Approximately Optimal Mechanism Design via Differential Privacy" [NST12].

References

[NST12] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. Approximately optimal mechanism design via differential privacy. In *ITCS*, pages 203–213, 2012.