15780: GRADUATE AI (SPRING 2018)

Practice Final
(Solutions)

April 28, 2019

1 Social Choice: Strategyproofness

Consider the library allocation problem from slides (Social choice II. Slide 9.), where we pick the location to set up a library. For this problem, we will consider the real plane ($\mathbb{R}^2$) as opposed to the real line ($\mathbb{R}$). Recall that each player has a true preference for the location of the library, which we will refer to as a *peak*.

Assume that the utility function of a player whose peak is $x \in \mathbb{R}^2$ is $-d(x, y)$ for a facility located at y, where d denotes Euclidean distance. Given player peaks $x^1, ..., x^n$, consider the mechanism that locates the library at $(\text{med}\{x_1^i\}, \text{med}\{x_2^i\})$. Prove that this mechanism is strategyproof, i.e., player i cannot increase their utility by reporting a peak that is different from x^i, regardless of the reports of other players.

Note: For simplicity, you can assume that the number of voters n is odd.

Solution: Consider an arbitrary player j whose peak is at x^j. And let $x^1, \ldots, x^{j-1}, x^{j+1}, \ldots x^n$ be arbitrary peaks reported by other players. On truly reporting peak x^j, let the location chosen for the library be $y^\star = (\text{med}\{x_1^i\}, \text{med}\{x_2^i\})$. Note that, the reported value of x_1^j does not affect $y_2^\star$ and the reported value of x_2^j does not affect $y_1^\star$. WLOG, let us consider the effect of misreporting the first coordinate. We have two cases possible:

Case 1: $x_1^j \leq y_1^\star$
If j misreports x_1^j by reporting any value smaller than or equal to $y_1^\star$, then $y^\star$ remains the median. Hence, j's utility does not change. On the other hand, if j reports a value larger than $y_1^\star$, the median of the first coordinate will either remain the same or strictly increase; let the new median be $\hat{y}_1^\star$. The new utility for j is then $-d(x^j, (\hat{y}_1^\star, y_2^\star)) = -\sqrt{(\hat{y}_1^\star - x_1^j)^2 + (y_2^\star - x_2^j)^2} \leq -\sqrt{(y_1^\star - x_1^j)^2 + (y_2^\star - x_2^j)^2} = -d(x^j, y^\star)$, where the inequality holds because $x_1^j \leq y_1^\star \leq \hat{y}_1^\star$. Hence, player j cannot increase their utility in this case.

Case 2: $x_1^j > y_1^\star$
If j misreports x_1^j by reporting any value bigger than or equal $y_1^\star$, then $y^\star$ remains the median. Hence, j's utility does not change. On the other hand, if j reports a value smaller than $y_1^\star$, the median of the first coordinate will either remain the same or strictly decrease; let the new median be $\hat{y}_1^\star$. The new utility for j is then $-d(x^j, (\hat{y}_1^\star, y_2^\star)) = -\sqrt{(x_1^j - \hat{y}_1^\star)^2 + (x_2^j - y_2^\star)^2} \leq -\sqrt{(x_1^j - y_1^\star)^2 + (x_2^j - y_2^\star)^2} = -d(x^j, y^\star)$, where the inequality holds because $x_1^j > y_1^\star \geq \hat{y}_1^\star$. Hence, player j cannot increase their utility in this case.

Therefore, whatever is reported for x_2^j, player j's utility is maximized by truly reporting x_1^j. Similarly, we can show that whatever is reported for x_1^j, player j's utility is maximized by truly reporting x_2^j. Hence, any player j cannot increase their utility by reporting a peak that is different of x^j.

2 Probabilistic Modeling: MLE and MAP

(a) [4 points] Given a collection of observed (independent) data points $X = \{x^{(1)}, \ldots, x^{(m)}\}$ from a uniform distribution over $[-2\alpha, \alpha]$ (for $\alpha > 0$), derive the maximum likelihood estimator of α, which maximizes the probability of observing X.

Solution: If any of the data points is not in $[-2\alpha, \alpha]$ (i.e., $\exists x \in X$ s.t. $x \notin [-2\alpha, \alpha]$), then we observe X with 0 likelihood. If all the data points are within $[-2\alpha, \alpha]$, then the likelihood of observing X is $\left(\frac{1}{3\alpha}\right)^m$. To maximize this likelihood, we want to find the minimum α such that all the data points are in $[-2\alpha, \alpha]$. All the data points are in $[-2\alpha, \alpha]$ if and only if $\alpha \geq \frac{-\min(X)}{2}$ ($\Leftrightarrow -2\alpha \leq \min(X)$) and $\alpha \geq \max(X)$. Thus,

$$\mathcal{L}(\alpha) = \begin{cases} \left(\frac{1}{3\alpha}\right)^m & \text{if } \alpha \geq \max(\max(X), \frac{-\min(X)}{2}) \\ 0 & \text{otherwise.} \end{cases}$$

The maximum likelihood estimator of α is $\hat{\alpha} = \max(\max(X), \frac{-\min(X)}{2})$ because for $\alpha < \hat{\alpha}$, $\mathcal{L}(\alpha) = 0 < \mathcal{L}(\hat{\alpha})$. Furthermore, for $\alpha > \hat{\alpha}$, $\mathcal{L}(\alpha) = \left(\frac{1}{3\alpha}\right)^m < \mathcal{L}(\hat{\alpha}) = \left(\frac{1}{3\hat{\alpha}}\right)^m$.

(b) [8 points] Given a collection of observed (independent) data points $X = \{x^{(1)}, \ldots x^{(m)}\}$ from a uniform distribution over $[0, e^{\alpha}]$ where α follows a prior distribution

$$p(\alpha) \propto e^{-\alpha^2},$$

derive the estimator of α that maximizes the posteriori probability $p(\alpha|X)$. **(Hint: use** $p(\alpha|X) \propto p(X|\alpha)p(\alpha)$**).**

Solution: If any of the data points is not in $[0, e^{\alpha}]$ (i.e., $\exists x \in X$ s.t. $x \notin [0, e^{\alpha}]$), then $p(X|\alpha) = 0$, and from the hint, $p(\alpha|X) = 0$. If all the data points are within $[0, e^{\alpha}]$, then

$$p(\alpha|X) \propto p(X|\alpha)p(\alpha) \propto \left(\frac{1}{e^{\alpha}}\right)^m e^{-\alpha^2} = e^{-\alpha^2 - m\alpha}. \tag{1}$$

Since $\exp(x)$ is a monotonically increasing function, we want to find α maximizing $f(\alpha) = -\alpha^2 - m\alpha$, while keeping all the data points in $[0, e^{\alpha}]$. Note that $f(\alpha)$ is a quadratic function maximized at $\alpha = -m/2$. If $\max(X) \leq e^{-m/2}$, then $\alpha = -m/2$ maximizes $f(\alpha)$ while satisfying $\forall x \in X$, $x \in [0, e^{\alpha}]$. If $\max(X) > e^{-m/2}$, then it means that $\alpha = -m/2$ is too small and we need to find a larger α. Since for $\alpha > -m/2$, $f'(\alpha) = -2\alpha - m < 0$ (i.e., f is strictly decreasing), $\hat{\alpha} = \log(\max(X))$ maximizes $f(\alpha)$ while satisfying $\forall x \in X$, $x \in [0, e^{\alpha}]$. For any $\alpha > \hat{\alpha}$, we would have $f(\alpha) < f(\alpha')$ and hence a smaller posterior probability. Hence,

$$\hat{\alpha} = \begin{cases} -m/2, & \text{if } \max(X) \leq e^{-m/2} \\ \log(\max(X)), & \text{otherwise,} \end{cases}$$

is the estimator of α maximizing $p(\alpha|X)$.

3 Game Theory: IESDS

One method of simplifying the search for Nash equilibria is through the iterated elimination of strictly dominated strategies (IESDS). We say that a player's pure strategy s_i' is strictly dominated by another pure s_i if $\forall s_{-i} \in S_{-i},\ u_i(s_i', s_{-i}) < u_i(s_i, s_{-i})$. In other words, s_1 dominates s_2 if, no matter what the other players do, player i always does strictly better by playing s_1 rather than s_2.

IESDS proceeds by repeatedly eliminating one strictly dominated strategy per round, until there are no more dominated strategies to eliminate. For example, IESDS on the following game proceeds as follows.

	North	East	South	West
Top	2,3	1,-1	4,0	3,-3
Middle	7,2	-2,0	5,2	6,7
Bottom	8,2	0,1	6,-1	4,0

- Column eliminates East, as playing North is strictly better.
- Row eliminates Top, as playing either Middle or Bottom is strictly better now that Column has eliminated East.
- Column eliminates South, as playing West is strictly better now.
- No more strategies can be eliminated; this leaves Row: [Middle, Bottom] and Column: [North, West] as the surviving strategies.

Prove the following: If IESDS eliminates all but one of the strategies of each player, then there is a unique Nash equilibrium in the game.

Hints:

- Start by proving that IESDS will never remove an action s_i that appears (with nonzero probability) in any Nash equilibrium.
- Conclude by applying Nash's Theorem: In any (finite) game, there exists at least one (possibly mixed) Nash equilibrium.

Solution: First, we prove that iterative elimination of strictly dominated strategies never removes an action that is in the support of any Nash equilibrium. Assume for the sake of contradiction that an action s_i^* in a Nash equilibrium s^* is removed, and, furthermore, that it is the first action in any Nash equilibrium that is removed. This means that there exists some s_i such that $u_i(s_i, s'_{-i}) > u_i(s_i^*, s'_{-i}),\ \forall s'_{-i} \in S_{-i}$ at this time. We know that because s_i^* is the first action in any Nash equilibrium to be removed, the rest of s^* has not been removed, so $s^*_{-i} \in S_{-i}$. Then, $u_i(s_i, s^*_{-i}) > u_i(s_i^*, s^*_{-i})$, which contradicts our assumption that s^* is a Nash equilibrium.

Now, by the hint and Nash's theorem, we are done: because there must exist at least one Nash equilibrium, and because IESDS never removes an action that is in the support of any Nash equilibrium, then this must be the unique Nash equilibrium in the game.

4 Adversarial Attacks

Assume we are given a set of m training points $S = \{(x^{(i)}, y^{(i)}) \in \mathbb{R}^D \times \{-1, +1\} \mid i = 1, \dots m\}$. Consider a monotonically decreasing classification loss $\mathcal{L} : \mathbb{R} \to \mathbb{R}$ and a hypothesis function $h_\theta(x) = \theta^T x$ mapping from $\mathbb{R}^D$ to $\mathbb{R}$ for $\theta \in \mathbb{R}^D$.

For this problem, assume that the training data is such that for every i, the first co-ordinate of $x^{(i)}$ equals its label and all other co-ordinates are zero i.e., $x_1^{(i)} = y^{(i)}$, and $x_j^{(i)} = 0$ for $j > 1$. Consider values θ^a and θ^b of the parameter, that perfectly classify the training data:

$$\theta^a = (1, \overbrace{0, 0, \dots, 0}^{D-1 \text{ zeros}})$$
$$\theta^b = (1, 1, 1, \dots, 1).$$

We can see that for all i, $h_{\theta^a}(x^{(i)}) \cdot y^{(i)} = h_{\theta^b}(x^{(i)}) \cdot y^{(i)} = 1$, leading to perfect classification.

(a) [8 points] **Robustness of θ^a to adversarial attacks.** Consider ϵ such that for every sample i, there exists an adversarial perturbation $\Delta^{(i)}$ satisfying $\|\Delta^{(i)}\|_\infty \leq \epsilon$ and $h_{\theta^a}(x^{(i)} + \Delta^{(i)}) \cdot y^{(i)} \leq 0$? Show that the smallest value ϵ can take is 1.

Solution:

First, we show that ϵ can take the value 1. For any i, consider $\Delta^{(i)}$ such that its first co-ordinate equals $-y^{(i)}$ and all other co-ordinates are zero. Clearly, $\|\Delta^{(i)}\|_\infty = 1$. Furthermore, $h_{\theta^a}(x^{(i)} + \Delta^{(i)}) \cdot y^{(i)} = (\theta^a \cdot x^{(i)}) \cdot y^{(i)} + (\theta^a \cdot \Delta^{(i)}) \cdot y^{(i)} = 1 + \Delta_1^{(i)} \cdot y^{(i)} = 1 - (y^{(i)})^2 = 0$.

Next, we show that this is the smallest value of ϵ possible. In particular, if $\epsilon < 1$, we have

$$\begin{aligned} h_{\theta^a}(x^{(i)} + \Delta^{(i)}) \cdot y^{(i)} &= (\theta^a \cdot x^{(i)}) \cdot y^{(i)} + (\theta^a \cdot \Delta^{(i)}) \cdot y^{(i)} \\ &= 1 + \Delta_1^{(i)} \cdot y^{(i)} \\ &\geq 1 - |\Delta_1^{(i)}| \\ &\geq 1 - \epsilon > 0. \end{aligned}$$

Hence, $\epsilon = 1$ is the smallest value possible.

(b) [8 points] **Robustness of θ^b to adversarial attacks.** Consider ϵ such that for every sample i, there exists an adversarial perturbation $\Delta^{(i)}$ satisfying $\|\Delta^{(i)}\|_\infty \leq \epsilon$ and $h_{\theta^b}(x + \Delta^{(i)}) \cdot y^{(i)} \leq 0$. Show that the smallest value ϵ can take is $1/D$.

Solution:

First, we show that ϵ can take the value $1/D$. For any i, consider $\Delta^{(i)}$ such that all its co-ordinates equal $-\frac{y^{(i)}}{D}$. Clearly, $\|\Delta^{(i)}\|_\infty = \frac{1}{D}$. Furthermore, $h_{\theta^b}(x^{(i)} + \Delta^{(i)}) \cdot y^{(i)} = (\theta^b \cdot x^{(i)}) \cdot y^{(i)} + (\theta^b \cdot \Delta^{(i)}) \cdot y^{(i)} = 1 - \sum_{j=1}^{D} \frac{1}{D}(y^{(i)})^2 = 0$.

Next, we show that this is the smallest value of ϵ possible. In particular, if $\epsilon < \frac{1}{D}$, we have that

$$\begin{aligned} h_{\theta^b}(x^{(i)} + \Delta^{(i)}) \cdot y^{(i)} &= (\theta^b \cdot x^{(i)}) \cdot y^{(i)} + (\theta^b \cdot \Delta^{(i)}) \cdot y^{(i)} \\ &= 1 + y^{(i)} \sum_{j=1}^{D} \Delta_j^{(i)} \theta_j^b \\ &\geq 1 - \max_j \left|\Delta_j^{(i)}\right| \sum_{j=1}^{D} \left|\theta_j^b\right| \\ &\geq 1 - D\epsilon > 0. \end{aligned}$$

Hence, $\epsilon = \frac{1}{D}$ is the smallest value possible.