Harvard CS 182: Artificial Intelligence October 20, 2022

Section 7: Stackelberg Security Games and Social Choice

Lecturer: Ariel Procaccia *Authors: Lauren Cooke and Sanjana Singh*

Stackelberg Security Games

Stackelberg Games: Definitions

A *Stackelberg Game* is a two-player game where the leader will go first and the follower will go second. The game will play out according to the following steps:

1. The leader will commit to some strategy
2. The follower will observe the leader's selected strategy (not necessarily knowing the move the leader makes if he plays with a mixed strategy)
3. Given the observations they make in step 2, the follower will commit to a strategy

It's important to remember that, unlike other games, a stackelberg game does not lock the leader into an individually dominant strategy. All that a Stackelberg game will guarantee is that the follower will choose a best response strategy to the strategy that the leader is using.

Example 1 *Consider the following stackelberg game where the row player is the leader:*

$(1,1)$	$(3,0)$
$(0,0)$	$(2,1)$

Table 1: Example Game

We can find the maximum utility for the leader when they commit to the strategy $(0.5, 0.5)$*. Given that we break ties in favor of the leader and our follower has an equal probability of reaching their maximum utility from choosing either column, our follower will commit to the right column. Given this commitment, we can find the total expected utility for the leader by summing up our utility options multiplied by the probability that each option occurs:*

$$(.5)(3) + (.5)(2) = 1.5 + 1 = 2.5$$

Which you will notice is greater than the utility for the leader at our nash equilibrium point of $(1, 1)$*.*

Let the leader of our Stackelberg game play with a mixed strategy called x_1. We can define the set of *best response strategies*, or the set of pure strategies that player 2 can play that maximize the utility of the follower, that the follower can use as $B_2(x_1)$.

$$B_2(x_1) = \text{argmax}_{S_2 \in S} u_2(x_1, s_2)$$

Where s_2 represents the selected follower strategy, S represents the set of strategies, S_2 represents the set of pure strategies that player 2 can play, and $u_2(x_1, s_2)$ represents the utility for the follower given that the leader uses strategy x_1 and the follower uses the strategy s_2.

In a *strong Stackelberg equilibrium*, the leader plays a mixed strategy to maximize their outcome (and to favor the leader in a tie-breaking situation) given that the follower will choose the best response strategy, adhering to the following:

$$\text{argmax}_{x_1 \in \Delta(S)} \text{max}_{s_2 \in B_2(x_1)} u_1(x_1, s_2)$$

Where $\Delta(S)$ is the set of mixed strategies and $u_1(x_1, s_2)$ represents the utility for the leader given that the leader uses strategy x_1 and the follower uses the best response strategy s_2.

Problem 1 *Consider the following game:*

$(3, 1)$	$(5, 0)$
$(2, 0)$	$(4, 2)$

Table 2: Problem 1 Game

Find the nash equilibrium for this game and find the strategy that the leader can play to maximize their utility in a stackelberg game where the row player is the leader.

Stackelberg Games: Computation

In two-player normal form games, we can compute a Strong Stackelberg Equilibrium in polynomial time using a linear program where the variables are the $x(s_1)$s and our goal is to find a probability assignment for the mixed strategy for the leader:

$$\max \sum_{s_1 \in S} x(s_1) u_1(s_1, s_2^*)$$

$$\text{s.t } \forall s_2 \in S, \sum_{s_1 \in S} x(s_1) u_2(s_1, s_2^*) \geq \sum_{s_1 \in S} x(s_1) u_2(s_1, s_2)$$

$$\sum_{s_1 \in S} x(s_1) = 1$$

We are trying to maximize the expected utility for the leader given that the leader plays some mixed strategy $x(s_1)$ that represents the probability that each possible pure leader strategy $s_1 \in S$ is played with and our follower chooses a best response strategy s_2^*.

We want to solve this linear program for every possible follower strategy $s_2 \in S$, so that we can choose the leader strategy that gives us the highest utility among these different follower options. We maximize our leader utility subject to the following constraints:

1. *ensure that the selected follower strategy is optimal*: for all follower strategies $s_2 \in S$, we enforce that the expected utility for the follower when playing strategy s_2^*, $\sum_{s_1 \in S} x(s_1) u_2(s_1, s_2^*)$, is greater than or equal to the expected utility for the follower when using any other strategy s_2.
2. *ensure that our mixed strategy is valid*: the sum of all probabilities that each leader strategy is played as defined in $x(s_1)$ is 1.

Stackelberg Security Games

A security game has the following components:

1. a set of *targets* $T = \{1, \ldots, n\}$, or places that can be attacked
2. a set of m *security resources* Ω, or defense mechanisms that can protect targets
3. a set of *schedules* $\Sigma 2^T$, which are subsets of targets that can be defended using some resource $\omega \in \Omega$. These possible assignments are captured in the set $A(\omega)$.

Note that given how resources are allocated to schedule, we get a set of *coverage probabilities* for all targets that we can call $c = (c_1, \ldots, c_n)$. Given this setup, the attacker then gets to chose one target to attack.

Per target $t \in T$, we have four measurable values of utility:

1. $u_d^+(t)$ the defender's utility if defense protects the target t and t is attacked
2. $u_d^-(t)$ the defender's utility if defense does not protect the target t and t is attacked
3. $u_a^+(t)$ the attacker's utility if defense protects the target t and t is attacked
4. $u_a^-(t)$ the attacker's utility if defense does not protect the target t and t is attacked

With these terms, we can then define the expected utility to the defender $u_d(t, c)$ and the expected utility to the attacker $u_a(t, c)$ if target t is attacked under coverage probability setup c:

$$u_d(t, c) = (c_t)(u_d^+(t)) + (1 - c_t)(u_d^-(t))$$
$$u_a(t, c) = (c_t)(u_a^+(t)) + (1 - c_t)(u_a^-(t))$$

Social Choice

Social choice theory is the theoretical framework to analyze the combination of opinions, preferences, interests, or welfares of individual agents to reach a collective decision. One of the most prominent uses of social choice theory is to study voting procedures. The voting model includes

- set of agents (voters) $N = \{1, ..., n\}$
- voters make decisions w.r.t. set of possible alternatives $A = \{a1, a2, ..., am\}$
- voter preferences over the alternatives $\sigma_i L$: $x \succ_i y$ means voter i prefers alternative x to y
- preference profile $\sigma \in L_n$ of all voter preferences
- voting rule $f : L^n \to A$

Our challenge in social choice theory is to combine these preferences using a social welfare function to come up with a social preference order, ranking the alternatives from most preferred to least preferred.

Voting Rules

A voting rule is a function that maps a preference profile to a preferred alternative.

Plurality The score vector for the plurality rule is $(1, 0, ..., 0)$. The score of an alternative is the number of voters who ranked it first. Plurality with runoff, used in the presidential election in France, is a variant of plurality with two rounds in which the top two alternatives advance to the second round, and then applying the plurality rule again.

Problem 2 *Americans typically revile third-party presidential candidates like Ralph Nader as "spoilers," candidates that distort an election, but many countries have (reasonably) well-functioning governments with multiple parties. Concept check: Why does America's "winner-take-all" majority voting rule function poorly for three, but well for two, parties?*

Borda An alternative gets k points from voter i if i prefers a to k other alternatives. The score vector is $(m1, m2, ..., 0)$. The alternative with the highest score is the social choice. Borda is used for elections in Slovenia and Nauru. Plurality and Borda are both *positional scoring* rules in which each alternative a receives s_i points for every voter that puts a in the ith position.

Single transferable vote (STV) STV successively eliminates alternatives that are ranked first by the smallest number of voters. Voter preferences are updated so that the second choice can take the place of the first choice for voters who selected that alternate first, and repeated for $m1$ rounds until a single alternative (which is our winner) is left. STV is common called "ranked-choice voting" with instant runoff and is used in presidential elections in Ireland, state and federal elections in Australia, and more recently in Maine.

A *pairwise election* evaluates two alternatives x and y. Alternative x beats y if the majority of voters prefer x to y, that is, $|i \in N : x \succ_i y| > n/2$.

An alternative is a *Condorcet winner* if it wins pairwise elections against all other alternatives. A voting rule is *Condorcet consistent* if the Condorcet winner, if exists, must be elected by the rule. A Condorcet winner does not always exist, and many voting rules are not Condorcet consistent.

Example 2 (Condorcet's Paradox) *Suppose we have three alternatives $A = x, y, z$ and three voters $N = 1, 2, 3$ with preferences:*

1. $x \succ_1 y \succ_1 z$
2. $z \succ_2 x \succ_2 y$
3. $y \succ_2 z \succ_2 x$

The paradox is that 2/3 of voters prefer z over x, 2/3 of voters prefer x over y, and 2/3 of voters prefer y over z.

Dodgson's rule Defining the distance between profiles as the number of swaps between adjacent alternatives, the Dodgson score of alternative x is the minimum distance from a profile where x is a Condorcet winner. Dodgson's rule selects the alternative that minimizes Dodgson score.

In other words: we select the alternative that can be made a Condorcet winner by interchanging as few adjacent alternatives in the individual rankings as possible.

Let preference profile σ' be obtained from by pushing $x \in A$ upwards for some individual preference. A voting rule is *monotonic* if whenever $f(\sigma) = x$ and σ' is obtained from σ by pushing x upwards, then $f(\sigma') = x$.

Problem 3 *The year is 2019, and you and your housemates are having some friends over for a party. You're tasked with planning food. To ensure that your guests are as happy as can be, you send out a survey asking them to rank their favorite cheeses: gouda, brie, or edam. Their results come in, and the responses are:*

- *4 votes: 1. edam 2. gouda 3. brie*
- *3 votes: 1. brie 2. gouda 3. edam*
- *2 votes: 1. gouda 2. brie 3. edam*

Unfortunately, there is no clear winner and you can only select one cheese. To be the best possible host, how should you decide?

Fortunately, you've just had a lecture on computational social choice, so you decide to put some voting rules into practice. What would your decision be if you used each of the following voting rules?

Plurality, Borda, STV, Plurality with runoff.

How do the pairwise elections between the following alternatives turn out?

- *Edam vs. Brie*
- *Edam vs. Gouda*
- *Gouda vs. Brie*

Looking at these pairwise results, are any alternatives a Condorcet winner?

What is the Dodgson score of each alternative?

Problem 4 *Consider the STV voting rule. Suppose that Arrow Academy, a high school with very democratic-minded students, is holding an election for class president. Suppose the students running comprise the set of alternatives is $A = \{$Frances Allen, Daniel Bernoulli, Mary Cartwright$\}$ and we have the preference profile:*

- *27 votes: 1. Allen 2. Bernoulli 3. Cartwright*
- *42 votes: 1. Cartwright 2. Allen 3. Bernoulli*
- *24 votes: 1. Bernoulli 2. Cartwright 3. Allen*

a) Cartwright puts forth a very rousing campaign to extend the lunch period by 10 minutes, and four votes switch from [Allen $\succ$ Bernoulli $\succ$ Cartwright] to [Cartwright $\succ$ Allen $\succ$ Bernoulli]. How does this change the result?

b) What would happen if, in a parallel world, Cartwright instead enacted a voter suppression campaign, causing four voters with preference [Allen Bernoulli Cartwright] to not vote?