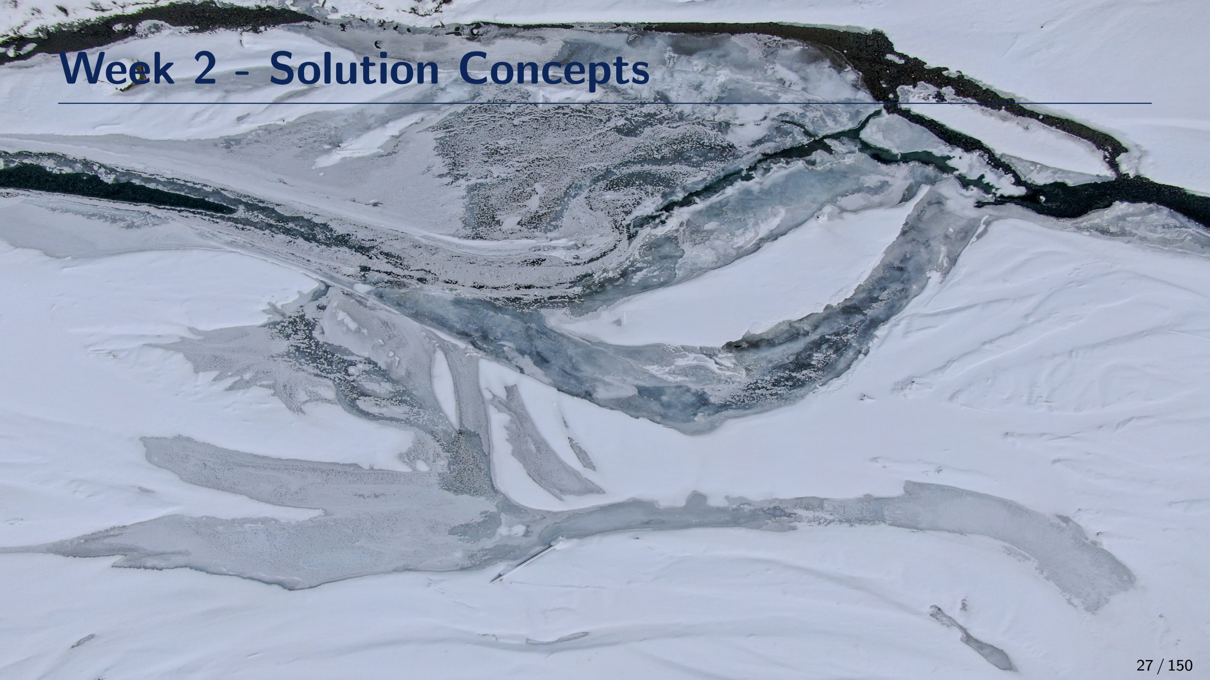
Week 2 - Solution Concepts



Normal Form Games

- Unlike RL, the return/reward is a function of all the agents in the game
- Fixing the the the policies of other players results in a single agent environment
- Minmax vs Nash solution concepts

Fixed Strategies

- We assume the agents follow a fixed strategy
  - Does not allow for opponent adaptation or online learning
- The analysis is for tabular, explicit representations
  - But works for implicit and online algorithms as long as they allow for tabularization [1]
- Training and evaluation paradigm - as used in all of the major games AI milestones

[1] Sustr M, Schmid M, Moravcik M, Burch N, Lanctot M, Bowling M. Sound search in imperfect information games.

Train/Eval Paradigm

(a) Train

(b) Eval

Search

- Online algorithm
- Common idea used in many perfect information games (chess, go, ...)
- Many appealing properties
- We still assume that it is consistent with a fixed, tabular strategy
- Makes only sense in sequential decision making - we will come back to this later

Solution Concepts

(a) Maximin

(b) Nash equilibrium

Maximin

- Maximizing the worst-case scenario
- Assumes everyone else is "out there to get you"

Definition: Maximin Policy

Maximin policy of a player i is:

$$\arg\max_{\pi_i \in \Pi_i} \min_{\pi_{-i} \in \Pi_{-i}} R_i(\pi_i, \pi_{-i}) = \arg\max_{\pi_i \in \Pi_i} BRV_i(\pi_i) \tag{1}$$

Nash equilibrium

- Everyone is happy

Definition: Nash Equilibrium

Strategy profile (π_i, π_{-i}) forms a Nash equilibrium if none of the players benefit by deviating from their policy.

$$\forall i \in N, \forall \pi_i' : R_i(\pi_i, \pi_{-i}) \geq R_i(\pi_i', \pi_{-i})$$

Maximin vs Nash

- One is defined for strategy, the other for strategy profile
- We will see some interesting differences
- But we will also see that they are sometimes the same!

Maximin

Maximin in Pure Strategies

	Cooperate	**Defect**
Cooperate	(-6, -6)	(0, -10)
Defect	(-10, 0)	(-1, -1)

Table: Prisoner's dillema

	Stop	**Go**
Stop	(0, 0)	(0, 1)
Go	(1, 0)	(-10, -10)

Table: Chicken's game

Maximin in Pure Strategies

	Rock	**Paper**	**Scissors**
Rock	0	1	-1
Paper	-1	0	1
Scissors	1	-1	0

Table: Rock paper scissors

Maximin

- Let's consider pure strategies
- When we mix, we can do better!
- Opponent does not care about their reward at all!
- How can we find the best mixed strategy?

Optimizing Against Best Response

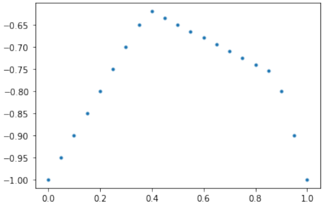
- For two-player zero sum games, we have

$$\arg\max_{\pi_i \in \Pi_i} \min_{\pi_{-i} \in \Pi_{-i}} R_i(\pi_i, \pi_{-i}) = \arg\max_{\pi_i \in \Pi_i} R_i(\pi_i, br_{-i}(\pi_i))$$

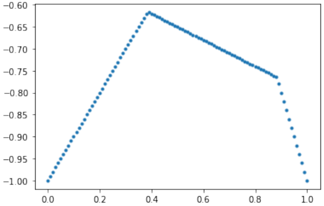
- We are thus optimizing against a best-responding player
- Let's visualise the best-response value function $f(\pi_i) = R_i(\pi_i, br_{-i}(\pi_i))$

Best Response Value Function

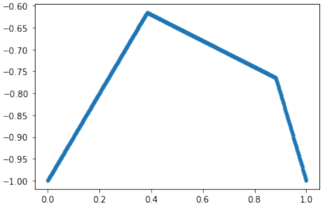
	A	B	C
X	-1	0	-0.8
1-X	1	-1	-0.5



(a) Step size 0.05



(b) Step size 0.01



(c) Step size 0.001

Nash

Nash equilibrium

Nash equlibrium

Strategy profile (π_i, π_{-i}) forms a Nash equilibrium if none of the players benefit by deviating from their policy.

$$\forall i \in N, \forall \pi_i' : R_i(\pi_i, \pi_{-i}) \geq R_i(\pi_i', \pi_{-i})$$

- Easy to verify - all strategies are best response
- Brute-force - enumerate all possible pairs and then verify

Nash in Pure Strategies

	Cooperate	**Defect**
Cooperate	(-6, -6)	(0, -10)
Defect	(-10, 0)	(-1, -1)

Table: Prisoner's dillema

	Stop	**Go**
Stop	(0, 0)	(0, 1)
Go	(1, 0)	(-10, -10)

Table: Chicken's game

Nash in Pure Strategies

	Rock	**Paper**	**Scissors**
Rock	0	1	-1
Paper	-1	0	1
Scissors	1	-1	0

Table: Rock paper scissors

Nash equilibrium

- Everyone is happy
- Pure Nash - enumerate, but might not exist!

Mixed Nash Equilibrium

$$\forall i \in N, \forall \pi_i' : R_i(\pi_i, \pi_{-i}) \geq R_i(\pi_i', \pi_{-i})$$

- Recall that the opponents are best-responding
- We also know that for best-response strategy, all the actions in the support have the same value
- For sparse x, y consider the corresponding elements of $xA_1, A_2 y^T$

Enumerating Support

- For sparse x, y consider the corresponding elements of xA_1, A_2y^T
- All the elements must correspond to the best-response value from the perspective of the **other** player
- In other words, player needs to mix actions in their support so that the action-values in the opponent's support are best-responding (and thus all the same)

Enumerating Support

	A	B	C
X	(0, 0)	(0, 1)	(-10, -10)
Y	(1, 0)	(-10, -10)	(-10, -10)
Z	(-10, -10)	(-10, -10)	(-10, -10)

Column player needs to mix their actions in the support (A, B) so that the values of the row player for actions in their support (X, Z) are all the same

$$0p(A) + 0p(B) = v_1 \tag{2}$$

$$-10p(A) - 10p(B) = v_1 \tag{3}$$

Same reasoning for the other side

$$0p(X) - 10p(Z) = v_2 \tag{4}$$

$$0p(X) - 10p(Z) = v_2 \tag{5}$$

No solution $\Rightarrow$ no Nash for this support

Enumerating Support

	A	B	C
X	(0, 0)	(0, 1)	(-10, -10)
Y	(1, 0)	(-10, -10)	(-10, -10)
Z	(-10, -10)	(-10, -10)	(-10, -10)

Column player needs to mix their actions in the support (A, B) so that the values of the row player for actions in their support (X, Y) are all the same

$$0p(A) + 0p(B) = v_1 \tag{6}$$

$$1p(A) - 10p(B) = v_1 \tag{7}$$

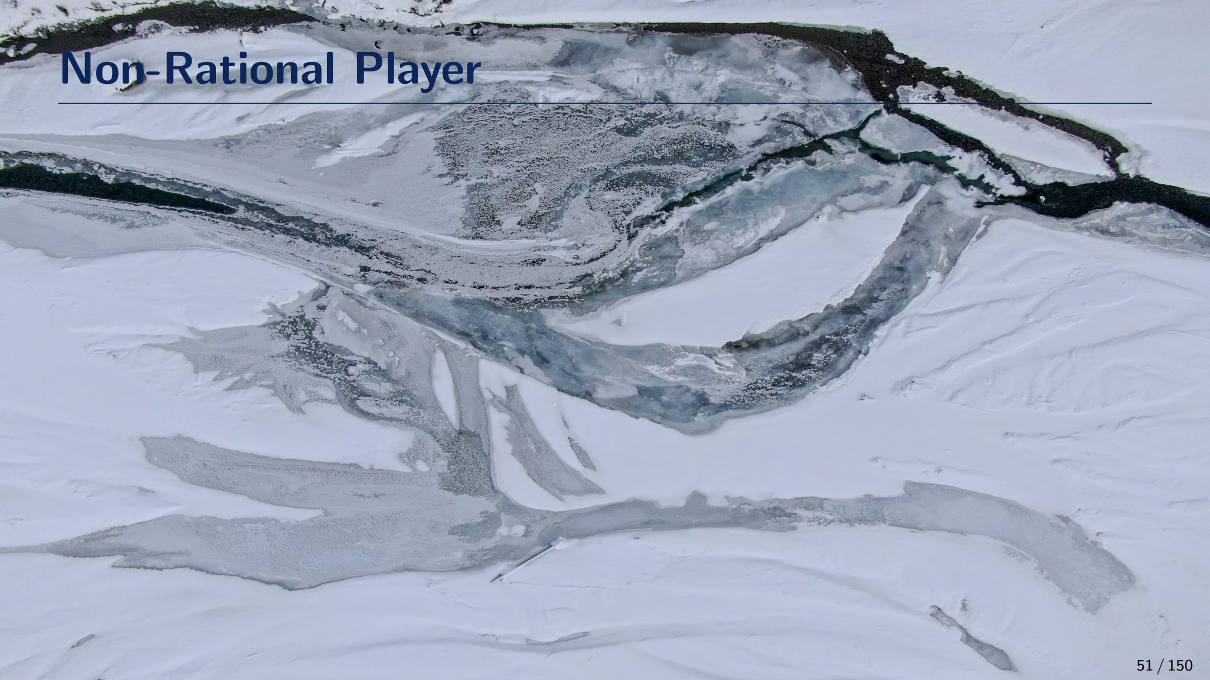
Same reasoning for the other side

$$0p(X) + 0p(Y) = v_2 \tag{8}$$

$$1p(X) - 10p(Y) = v_2 \tag{9}$$

Solution: $p(A) = 0.909091, p(B) = 0.090909, p(X) = 0.909091, P(Y) = 0.090909$

Non-Rational Player



Non-Rational Players - Deviating from Nash

Let's elaborate on the properties of this solution concept!

- What are the implications for the players?
- What are the situations we would/wouldn't play Nash?

Non-Rational Players

Suppose we play versus a stupid opponent

- Non-rational player does not maximize his utility, he can play arbitrarily
- Given Nash equilibrium $\pi = (\pi_0, \pi_1)$, we decided to play π_0, what do we know?
- Even though π_1 maximizes the utility for the opponent, he can make mistakes and select different (non-equilibristic) strategy π_1'
- Choosing different strategy than π_1' is no better for the opponent
- But it can be much worse for us! It can be the case that $u_0(\pi_0, \pi_1) \gg u_0(\pi_0, \pi_1')$

Moral of the story: opponent mistakes can hurt us!

Deviating from Nash

	Cooperate	**Defect**
Cooperate	(-6, -6)	(0, -10)
Defect	(-10, 0)	(-1, -1)

Table: Prisoner's dillema

	Stop	**Go**
Stop	(0, 0)	(0, 1)
Go	(1, 0)	(-10, -10)

Table: Chicken's game

Rational Players, Multiple Equilibria (I)

Suppose there are two optimal strategy profiles in the game (π_0, π_1^a) and (π_0, π_1^b)

- The opponent is indifferent between his two strategies, he does not care which strategy he chooses (given our strategy π_0), since $u_1(\pi_0, \pi_1^a) = u_1(\pi_0, \pi_1^b)$
- We care!
- $u_i(\pi_0, \pi_1^a) \neq u_0(\pi_0, \pi_1^b)$

Moral of the story: even thought both players play optimally, different optimal strategies can lead to different utilities!

Week 2 Homework

1. Draw best-response value function for a $(2 \times N)$ matrix game
2. For a two-player matrix game (does not have to be a zero-sum)
   2.1 Given a support for both players, construct a set of linear equations and see whether there exists a Nash equilibrium for that support. Use SciPy to solve the constructed LP `https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linprog.html`
   2.2 Enumerate all possible supports and try to find a Nash Equilibrium for each support