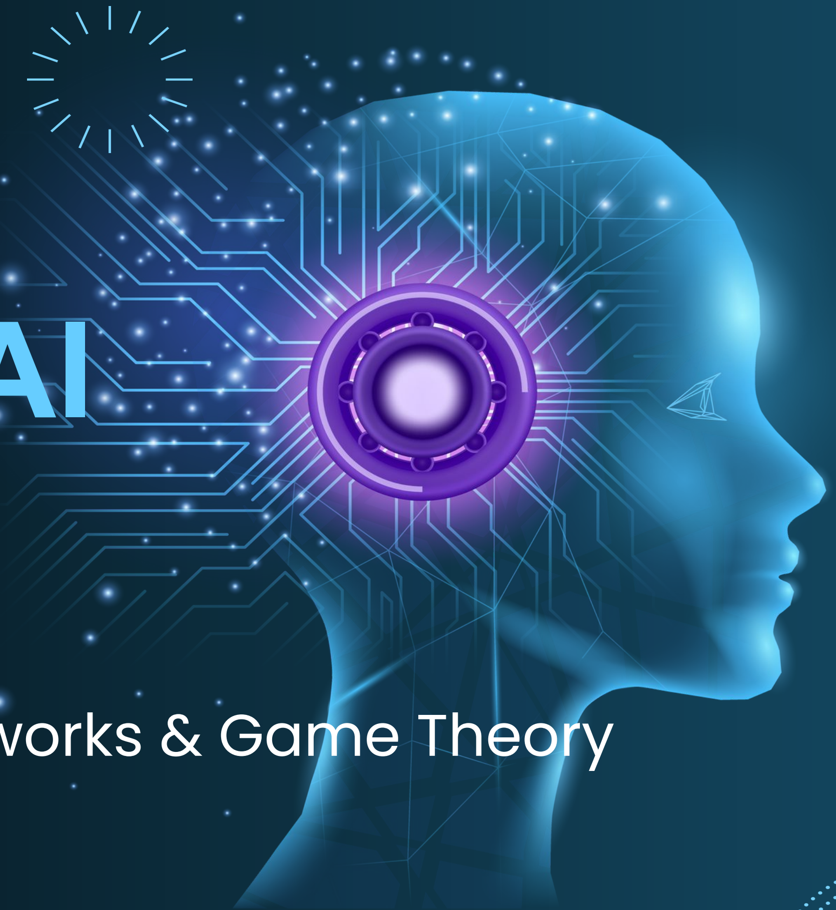
MONASH University

com/#pricingchtahttp
n/#pricingchat

# FIT5230
# MALICIOUS AI

S2 2025



## Week 7:

## Generative Adversarial Networks & Game Theory

# Overview

- GAN Quick Recap
- G/D Min-Max Game
- Game Theory
- Prisoner's Dilemma
- Nash Equilibrium
- Matching Coins Game
- GAN Motivation
- GAN for AI Security

# Generative Adversarial Network

## Quick Recap

# Generative Adversarial Networks

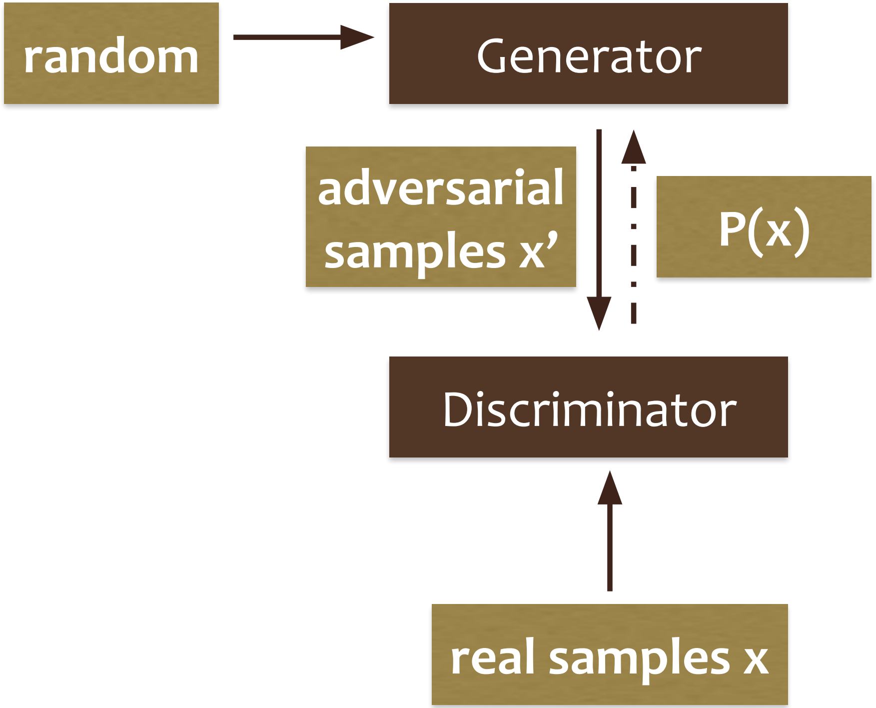
## Generative Adversarial Nets

**Ian J. Goodfellow, Jean Pouget-Abadie[*], Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair[†], Aaron Courville, Yoshua Bengio[‡]**
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

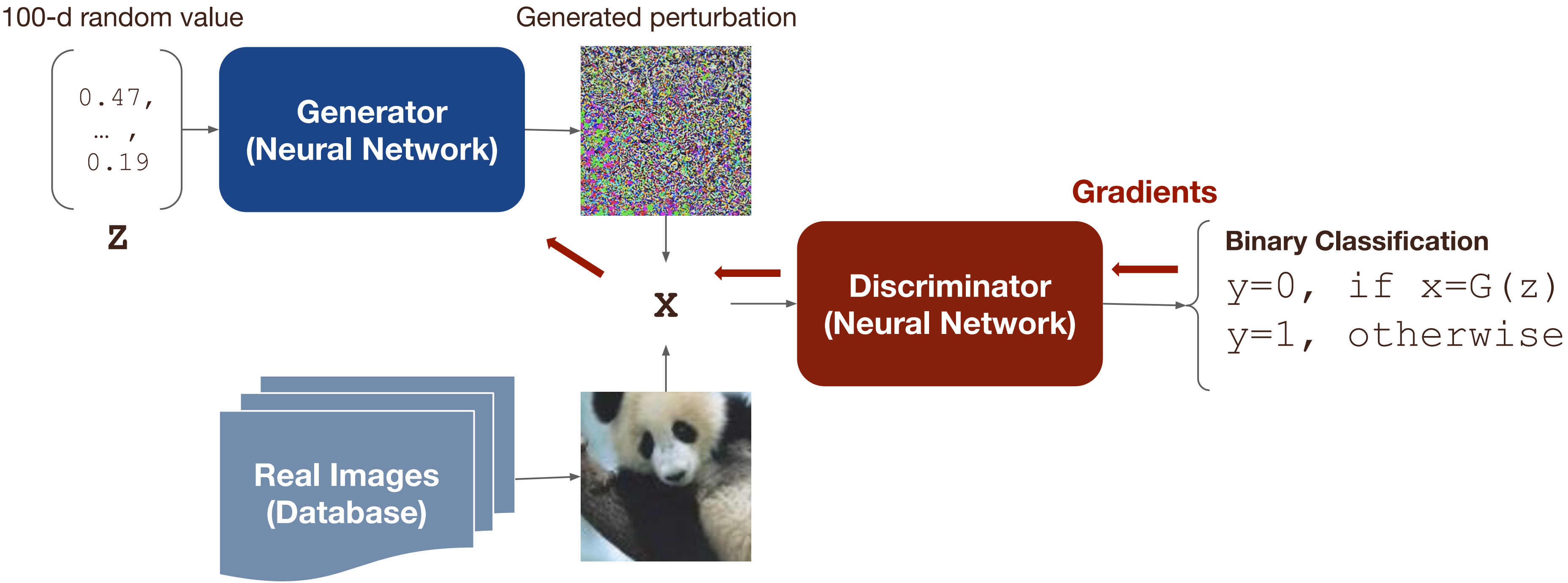
### Abstract

We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model $G$ that captures the data distribution, and a discriminative model $D$ that estimates the probability that a sample came from the training data rather than $G$. The training procedure for $G$ is to maximize the probability of $D$ making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions $G$ and $D$, a unique solution exists, with $G$ recovering the training data

# Generative Adversarial Networks



# G/D Game


100-d random value
Generated perturbation
0.47,
… ,
0.19
z
Generator
(Neural Network)
x
Gradients
Discriminator
(Neural Network)
Binary Classification
y=0, if x=G(z)
y=1, otherwise
Real Images
(Database)

# G/D Min-Max Game

The adversarial nature of GANs creates a Min-Max Game

- leading to the final loss function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

- The **discriminator** tries to maximize this function by correctly classifying real and fake data.
- The **generator** tries to minimize it by generating realistic samples that fool the discriminator.

# GANs Summary

- First presented by Ian Goodfellow and associates in 2014
- Composed of two neural networks that cooperate through a competitive process
  - **Generator**: To generate human imperceptible adversarial examples
  - **Discriminator**: To determine real vs fake data

The outcome of this adversarial process is extremely advanced models that can recognize and produce synthetic media that is remarkably realistic.

MONASH University

# Game Theory

# Prisoner’s Dilemma

# GAN & Game Theory

- Two players G and D in a game
- Shown that Nash equilibrium exists when function space corresponds to
$$p_{real} = p_{model}$$
- i.e. distribution of real = distribution of G's outputs

# Theory of Games

Game:

- players indexed by $i \in \{1, . . . , \ell\}$: $P_i$

Strategies $S_i$ = set of available actions for player i

- $s_i \in S_i$ = action by player i

Payoffs $u_i(S)$ = payoff/utility function of player i
where S is list of actions of player i

Q: payoff depends on whose strategies?

# Theory of Games: the Matrix

Matrix game / strategic form game

- 2-player game with identical rounds
- in each round: each player simultaneously makes a move
- outcome: tie or win/lose (loser pays winner)

| | Rock | Paper | Scissors |
|---|---|---|---|
| Rock | 0, 0 | -1, 1 | 1, -1 |
| Paper | 1, -1 | 0, 0 | -1, 1 |
| Scissors | -1, 1 | 1, -1 | 0, 0 |

# Theory of Games: the Matrix

Payoff matrix

- aij is payoff in a round
- P1 makes move i, P2 makes move j
- **+ve if P1 wins**, 0 if tie, -ve if P1 loses

Zero-sum game

- sum of winnings of all players = 0

| | Rock | Paper | Scissors |
|---|---|---|---|
| Rock | 0, 0 | -1, 1 | 1, -1 |
| Paper | 1, -1 | 0, 0 | -1, 1 |
| Scissors | -1, 1 | 1, -1 | 0, 0 |

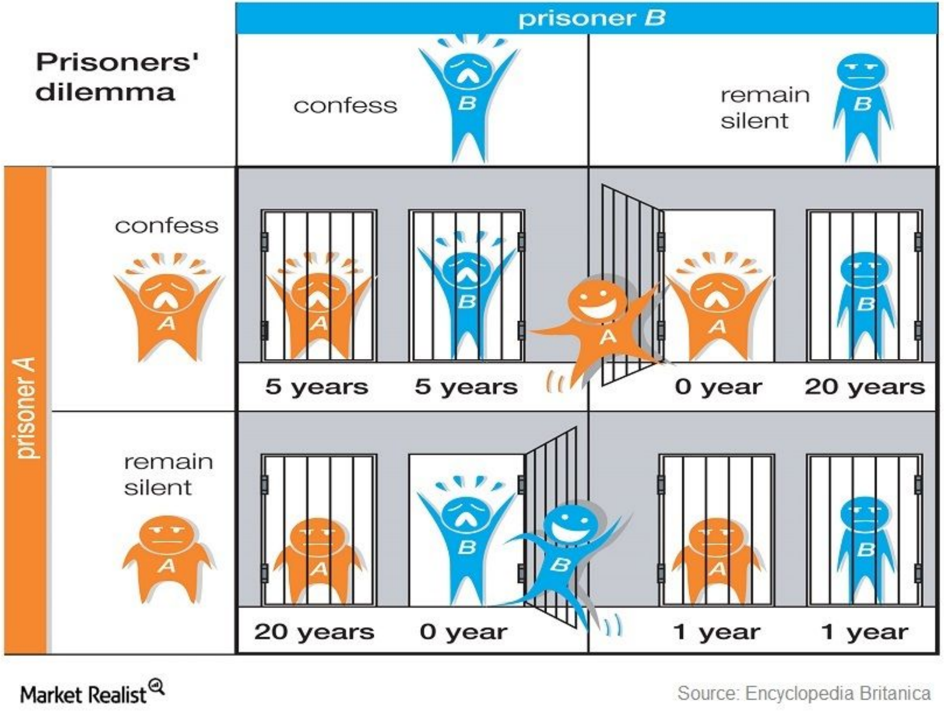
# Game Theory: Prisoner's Dilemma

Two suspects 웃 and 웃 arrested, questioned by police

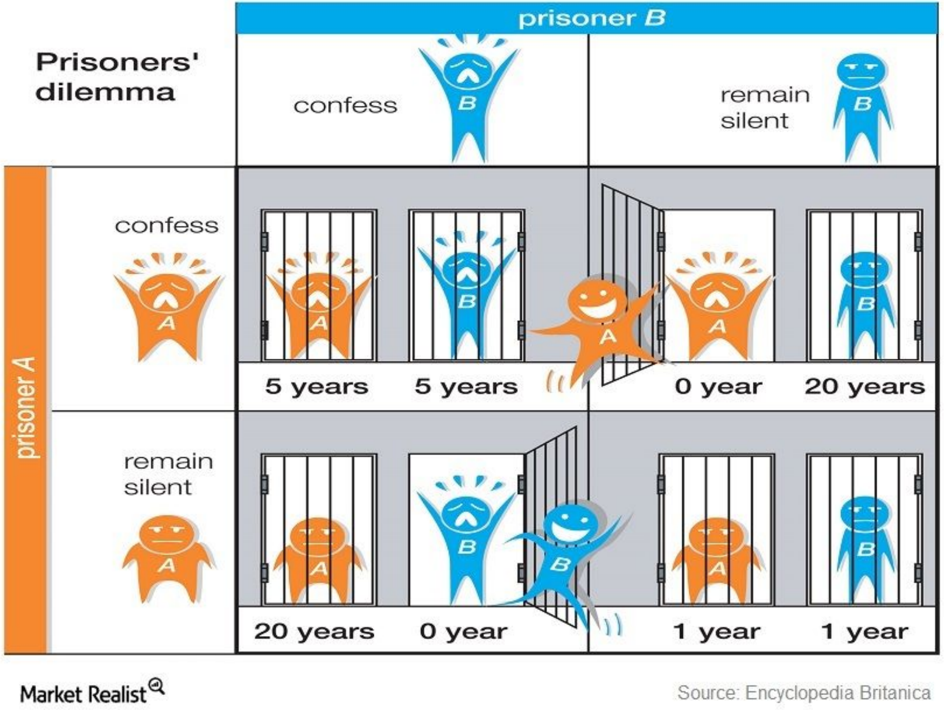
1. If no one confesses: 1 year jail each
2. If one confesses, the other keeps quiet:
   - leniency for the rat: 0 years
   - severe punishment for silence: 20 years
3. If both confess:
   - both punished equally: 5 years

Best strategy? To confess or not to confess?

# Game Theory: Prisoner's Dilemma



# Game Theory: Prisoner’s Dilemma




| 웃\웃 | Confess | Defect |
|---|---|---|
| Confess | (–5,–5) | (0,–20) |
| Defect | (–20,0) | (–1,–1) |

# Game Theory: Prisoner's Dilemma

Payoff matrix:

1. if 웃 confesses:
confess gets 웃 less years

2. if 웃 defects
confess gets 웃 less years

'confess' vs 'defect'
the only possible outcome is: ???

Q: Why didn't they go for ???

| 웃\웃 | Confess | Defect |
|---|---|---|
| Confess | (–5,–5) | (0,–20) |
| Defect | (–20,0) | (–1,–1) |

# Game Theory: Prisoner’s Dilemma

- ‘confess’ strictly dominates ‘defect’
- **rational** players never play strictly dominated strategies
- the only possible outcome is
  - <Confess, Confess> → (-5,-5)

Why didnt they go for <Defect,Defect> → (-1,-1)?

| 웃\웃 | Confess | Defect |
|---|---|---|
| Confess | (–5,–5) | (0,–20) |
| Defect | (–20,0) | (–1,–1) |

# Game Theory: Prisoner's Dilemma

Why not <Defect,Defect>?

- even if they had agreed to both not confess (i.e. defect) once in interrogation room, its in their best interest to confess irrespective of what the other party does
- so, its **not a stable** outcome, each of them can get better payoff by flipping their strategies, so **not equilibrium**

| 웃\웃 | Confess | Defect |
|---|---|---|
| Confess | (–5,–5) | (0,–20) |
| Defect | (–20,0) | (–1,–1) |

# Nash Equilibrium

Stable outcome:

- the only possible outcome: <Confess, Confess> → (-5,-5)

For this outcome,

- each player can't do any better by changing his/her strategy/action, so no incentive to change ➠Nash equilibrium
  - since 웃 confesses:
    changing from confess to defect …
  - since 웃 confesses:
    changing from confess to defect …

| 웃\웃 | Confess | Defect |
|---|---|---|
| Confess | (–5,–5) | (0,–20) |
| Defect | (–20,0) | (–1,–1) |

# Nash Equilibrium

Stable outcome/state,

- where no player can gain any better payoff by changing his/her strategy
- given that other players have chosen their action

Pure strategy Nash Equilibrium

- each player only one (pure) strategy: {Confess,Defect}

| 웃\웃 | Confess | Defect |
|---|---|---|
| Confess | (–5,–5) | (0,–20) |
| Defect | (–20,0) | (–1,–1) |

**Non-cooperative games**

J Nash - Annals of mathematics, 1951 - JSTOR

… prove that a finite **non-cooperative game** always has at least … **non-cooperative game** and prove a theorem on the geometrical structure of the set of equilibrium points of a solvable **game**. …

Save Cite Cited by 14413 Related articles All 35 versions

# Matching Coins game

Matrix game: the Matching Coins game

- 2 players 웃 and 웃
- actions: {Heads, Tails}
- Payoff matrix:

| 웃\웃 | Heads | Tails |
|---|---|---|
| Heads | (1,–1) | (–1,1) |
| Tails | (–1,1) | (1,–1) |

Best strategy: Heads or Tails?

# Penalty Kick game

Matrix game: the Penalty Kick game

- 2 players Goalie 웃 and Striker 웃
- actions: {Left, Right}
- Payoff matrix:

| 웃\웃 | Left | Right |
|---|---|---|
| Left | (1,–1) | (–1,1) |
| Right | (–1,1) | (1,–1) |

Best strategy: Left or Right?

# Matching Coins game

1. < Heads,Heads >: 웃 chose Heads
웃 prefers to have chosen Tails
so <Heads,Heads> not equilibrium

2. < Heads,Tails >: 웃 chose Tails
웃 prefers to have chosen Tails
so <Heads,Tails> not equilibrium

Recall… Payoff matrix:

| 웃\웃 | Heads | Tails |
|---|---|---|
| Heads | 1.(1,–1) | 2.(–1,1) |

| 웃\웃 | Tails |
|---|---|
| Heads | 2.(–1,1) |
| Tails | 3.(1,–1) |

| 웃\웃 | Heads | Tails |
|---|---|---|
| Heads | 1.(1,–1) | 2.(–1,1) |
| Tails | 4.(–1,1) | 3.(1,–1) |

# Matching Coins game

3. < Tails,Tails >: 웃 chose Tails
웃 prefers to have chosen Heads
so <Heads,Tails> not equilibrium

4. < Tails,Heads >: 웃 chose Heads
웃 prefers to have chosen Heads
so <Tails,Heads> not equilibrium

No pure strategy equilibrium
Q: can equilibrium be achieved?

| 웃\웃 | Heads | Tails |
|---|---|---|
| Tails | 4.(–1,1) | 3.(1,–1) |

| 웃\웃 | Heads |
|---|---|
| Heads | 1.(1,–1) |
| Tails | 4.(–1,1) |

| 웃\웃 | Heads | Tails |
|---|---|---|
| Heads | 1.(1,–1) | 2.(–1,1) |
| Tails | 4.(–1,1) | 3.(1,–1) |

# Nash Equilibrium

Stable outcome/state,

- where no player can gain any better payoff by changing his/her strategy
- given that other players have chosen their action

| 웃\웃 | Confess | Defect |
|---|---|---|
| Confess | (–5,–5) | (0,–20) |
| Defect | (–20,0) | (–1,–1) |

Pure strategy Nash Equilibrium

- each player only one (pure) strategy: {Confess,Defect}

**Non-cooperative games**

J Nash - Annals of mathematics, 1951 - JSTOR

… prove that a finite **non-cooperative game** always has at least … **non-cooperative game** and prove a theorem on the geometrical structure of the set of equilibrium points of a solvable **game**. …

☆ Save 99 Cite Cited by 14413 Related articles All 35 versions

# Nash Equilibrium

**Pure strategy Nash Equilibrium**

- each player only one (pure) strategy
- e.g. H or T, L or R, Go or Stop, Confess or Defect

Q: What if can't achieve pure strategy Nash Equilibrium? but one must exist …

- **Nash's Existence Theorem**
    - every game with a finite number of players who can choose from finitely many pure strategies, has at least one Nash equilibrium

# Mixed Strategy Nash Equilibrium

Based on Nash's Existence Theorem, if no equilibrium for pure strategies, must have for mixed

Mixed strategy

- probability distribution over multiple pure strategies: each one may be chosen based on some probability P( )

# Matching Coins game

Q: what if playing against a mind reader? how to not always lose?
A: ?

| 웃\웃 | Heads<br>0.5 | Tails<br>0.5 |
|---|---|---|
| Heads<br>0.5 | (1,–1) | (–1,1) |
| Tails<br>0.5 | (–1,1) | (1,–1) |

# Matching Coins game

Q: what if playing against a mind reader? how to not always lose?
A: just **flip the coin** (i.e. **random**)

- at best, opponent wins only half the time
- he can't do anything to change outcome
- neither player can change strategy & expect to do better
- mixed strategy Nash equilibrium

| 웃\웃 | Heads 0.5 | Tails 0.5 |
|---|---|---|
| Heads 0.5 | (1,–1) | (–1,1) |
| Tails 0.5 | (–1,1) | (1,–1) |

# Mixed Strategy Nash Equilibrium

Mixed strategy

- each player chooses an action with probability 0.5
- Payoff for each player for choosing an action:
  - payoff웃(H) = 0.5(1)+0.5(–1) = 0
  - payoff웃(T) = 0.5(–1)+0.5(1) = 0
  - payoff웃(H) = 0.5(–1)+0.5(1) = 0
  - payoff웃(T) = 0.5(1)+0.5(–1) = 0

| 웃\웃 | Heads 0.5 | Tails 0.5 |
|---|---|---|
| Heads 0.5 | (1,–1) | (–1,1) |
| Tails 0.5 | (–1,1) | (1,–1) |

vs pure strategy

- payoff either 1 or –1

# Battle of the Sexes game

- 2 players: man 웃 and woman 웃
  - get together for night for entertainment, but no communication
  - actions: choose ballet or watch fight
  - man 웃 prefers fight, woman 웃 prefers ballet
  - both prefer together vs alone

Payoff matrix

| 웃\웃 | Ballet | Fight |
|---|---|---|
| Ballet | (1,2) | (0,0) |
| Fight | (0,0) | (2,1) |

Q: any Nash equilibrium?

# Battle of the Sexes game

Pure strategy Nash equilibrium **exists**

- since 웃 chose Ballet
  웃: no point to change from Ballet to Fight
- since 웃 chose Ballet
  웃:  no point to change from Ballet to Fight
  …

| 웃\웃 | Ballet | Fight |
|---|---|---|
| Ballet | (1,2) | (0,0) |
| Fight | (0,0) | (2,1) |

Pure strategy Nash equilibrium:

- <Ballet, Ballet> or
- <Fight, Fight>

# Battle of the Sexes game

Pure strategy Nash equilibrium exists

But how to decide? Each has his/her own preference, how to coordinate even though it can be done? Not sure which one?

Check if **mixed strategy Nash equilibrium** exists using the **Mixed Strategy Algorithm**

| 웃\웃 | Ballet | Fight |
|---|---|---|
| Ballet | (1,2) | (0,0) |
| Fight | (0,0) | (2,1) |

# Mixed Strategy Algorithm

Solve for 웃's mixed strategy:

① Target: payoff웃(Ballet) = payoff웃(Fight)

- so it won't matter to 웃 which one she choses

② payoff웃(L)= P웃(U)(2) + (1-P웃(U))(0)

③ payoff웃(R)= P웃(U)(0) + (1-P웃(U))(1)

- where P웃(U) is probably of 웃 choosing U

| 웃\웃 | L | R |
|---|---|---|
| U | (1,2) | (0,0) |
| D | (0,0) | (2,1) |

# Mixed Strategy Algorithm

Solve for 웃's mixed strategy:

① Target: payoff웃(Ballet) = payoff웃(Fight)

- so it won't matter to 웃 which one she choses

② payoff웃(U)= P웃(L)(1) + (1-P웃(L))(0)

③ payoff웃(D)= P웃(L)(0) + (1-P웃(L))(2)

- where P웃(L) is probably of 웃 choosing L

| 웃\웃 | L | R |
|---|---|---|
| U | (1,2) | (0,0) |
| D | (0,0) | (2,1) |

# Mixed Strategy Algorithm

Payoff for each player for choosing an action:

① compute the P of each outcome, using P웃 and P웃
② for each outcome, multiply P by player's payoff,  & sum these up for all outcomes

payoff웃 = 2/9(1) + 1/9(0) + 4/9(0) + 2/9(2)= ⅔
payoff웃= 2/9(2) + 1/9(0) + 4/9(0) + 2/9(1) = ⅔

Pure strategy payoffs are better; just give in & follow

| 웃\웃 | Ballet (⅔) | Fight (⅓) |
| --- | --- | --- |
| Ballet (⅓) | (1,2) P=2/9 | (0,0) P=□ |
| Fight (⅔) | (0,0) P=4/9 | (2,1) P=2/9 |

# Generative Adversarial Network

## AI Security

# GANs for Security

The Threat Landscape

- Deepfakes → AI-generated media (video, audio, text, images) that depict fabricated events or speech
- Forged Content → Broader manipulation of digital media beyond deepfakes

Key Risks:

- Misinformation: Undermines trust in news and media.
- Identity Theft & Fraud: Impersonation for financial or personal gain.
- Political Manipulation: Influences elections and public opinion.
- Defamation: Damages reputations through falsified portrayals.

# GANs for Security: Threat Detection and Prevention

GANs as a Defense Mechanism

- GANs can be trained on real vs. fake datasets to identify manipulation artifacts.
- Detects subtle inconsistencies in lighting, audio, facial expressions, motion patterns.

Application in Content Moderation

- Automatic Flagging: AI detection for suspicious media.
- Real-Time Analysis: Prevents rapid spread of harmful content.
- User Reporting Integration: Enhances system accuracy with human input.

# The Dark Side: GANs as a Security Threat

- Realistic Malware Creation:
  - GANs generate polymorphic malware that evades signature-based antivirus.
  - Subtle perturbations → difficult to classify as malicious.
- Adversarial Evasion Attacks:
  - GANs create adversarial inputs that mislead ML-based security systems.
  - Exploits classifier vulnerabilities.
- Data Poisoning:
  - GAN-synthetic data injected into training pipelines.
  - Corrupts learning process, weakening long-term detection accuracy.

Reference: Bringing a GAN to a Knife-Fight: Adapting Malware Communication to Avoid Detection, Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN, Mal-D2GAN: Double-Detector based GAN for Malware Generation, VagueGAN: A GAN-Based Data Poisoning Attack Against Federated Learning Systems