Review of some technical concepts from class

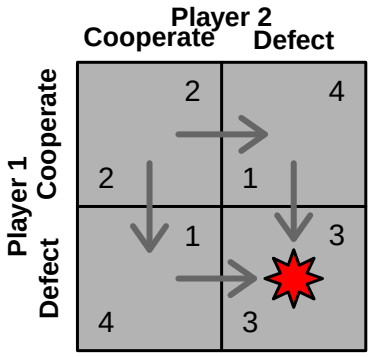
PSYC 162: Evolution of Cooperation

Classical game theory

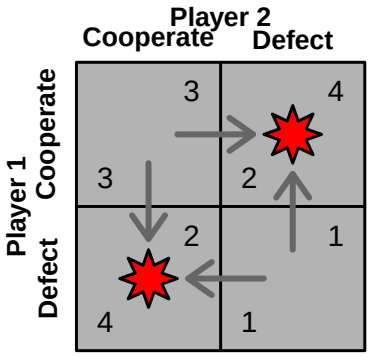
- Two strategies and two players. The players **simultaneously** decide which strategy to select.
- The **best response** is the strategy which produces the most favorable outcome for a player, *taking the other player's strategy as given.*
- A set of (pure) strategies (C, C) is a **Nash Equilibria** if player 1 is playing a best response against player 2, and player 2 is playing a best response against player 1!
- $U_1(C, C) > U_1(D, C)$ **and** $U_2(C, C) > U_2(C, D)$
- In a NE, neither player has an incentive to switch strategies!
- Once we get here, we are stuck. Note this framework doesn't tell us *how* we might get to NE, especially when we have multiple equilibria.

Common game types and their NE,

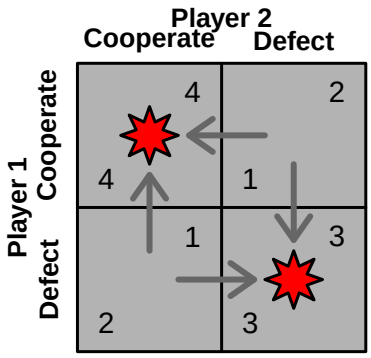
Social Dilemma



Anti−coordination



Coordination



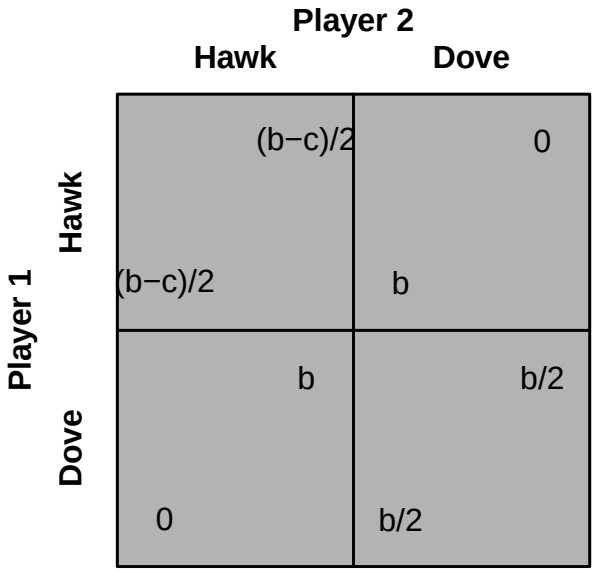
Evolutionary game theory

- Nash equilibrium from classic game theory is a **static concept**.

- Evolutionary game theory allows *populations* of agents to evolve and equilibria predictions are stochastic rather than deterministic.
- Fraction p of the population plays Strategy 1 and fraction $(1-p)$ plays Strategy 2.
- Strategy A **risk dominates** strategy B if $\pi_A > \pi_B$ at $p = 1/2$.
  - The risk dominant strategy has the largest **basin of attraction**
- When $\pi_A = \pi_B$, solve for p to get the **interior equilibrium**, p^*.
- An **Evolutionary Stable Strategy (ESS)** is a strategy which, if adopted by a population, cannot be invaded by a mutant strategy that is initially rare.

Illustration (solving evolutionary games):

Consider the Hawk Dove Game,



When two Doves meet, they divide the benefit, b, in half and share so each get payoffs $b/2$. When two Hawks meet they win against the other with equal probability of $1/2$, and incur a cost of c for fighting. Let p denote the proportion of the population playing Hawk so $(1-p)$ is the proportion playing Dove. Assume $b > c$ so that $(b-c) > 0$.

Step 1: always calculate the NE (the only NE here is [Hawk,Hawk]) and the expected payoffs for each strategy:

$$\pi_H = p\left[\frac{(b-c)}{2}\right] + (1-p)b$$
$$\pi_D = 0 + (1-p)\frac{b}{2}$$

Risk dominance: Find the risk dominant strategy by determining which strategy has the largest expected payoff at $p = 0.5$. [Note: we don't need to assume $p = 0.5$ for this to work out and under time pressure you might skip this.]

$$\pi_H = \frac{b-c}{4} + \frac{b}{2}$$
$$\pi_D = \frac{b}{4}$$

It's clear that $\pi_H > \pi_D$ because 1) $b - c > 0$ by assumption and $\frac{b}{2} > \frac{b}{4}$ for any $b > 0$. Since Hawk is risk dominant, $p = 1$ is more likely to evolve than $p = 0$. Another way of saying this same thing is that Hawk has a larger basin of attraction, so a larger fraction of initial conditions will lead to $p = 1$ than $p = 1$.

Interior equilibrium: Find the interior equilibrium by equating expected payoffs and solving for p^*.

$$p\left[\frac{(b-c)}{2}\right] + (1-p)b = (1-p)\frac{b}{2}$$
$$\cdots$$
$$p^* = \frac{b}{c}$$

Recall that $p \in [0, 1]$ so if $p^* = 0.5$ the interior equilibrium is 50% Hawk and 50% Dove, in which case the basin of attraction is the same size for each strategy. If $p^* > 0.5$ then Hawk has the larger basin of attraction and risk dominantes Dove. Likewise, if $p^* < 0.5$ then Dove has the larger basin of attraction. Here, we see that as long as $b > c$ (which is our assumption) Hawk has the larger basin of attraction. In fact, Hawk has the *entire* basin of attraction, since $p > 1$ when $b > c$, so no matter what value of p^* we pick, the population will evolve to all Hawks. So we have a clear prediction about what will happen here.

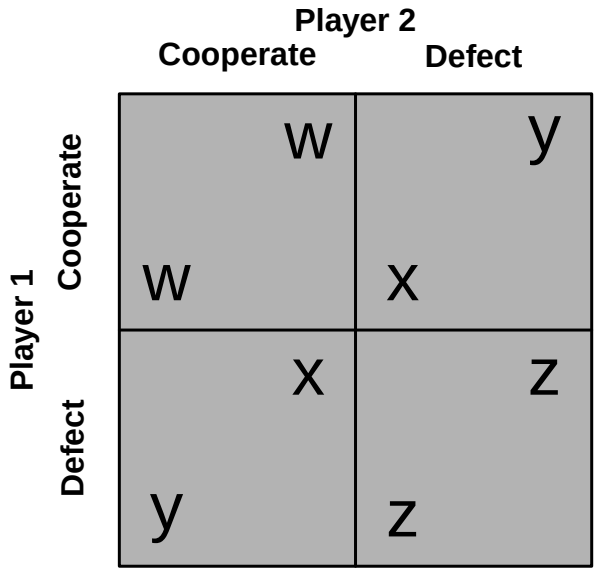
Stability: selection may or may not push the population away from the interior equilibrium toward $p = 0$ or $p = 1$. For illustration, let's consider the Hawk-Dove game where $c < b$, so the cost of fighting is less than the benefit (verify that in order for Dove to Risk dominate Hawk, it must be the case that $c > 2b$). If the interior equlibrium (p^*) is **stable** then coexistence is a possibility. To test for stability, we ask what happens when p^* increases by just a little bit. Consider the relative advantage of Hawk over Dove: $\pi_H - \pi_D$. We can rearrange this with some algebra to be $(b - pc)/2$. At the interior equilibrium p^*, the relative payoff is, as expected, zero since $\frac{1}{2}[b - \frac{b}{c}c] = 0$. If mutation occurs in favor of Hawks and we increase p^* by just a little bit, some $\epsilon > 0$, then we see that the relative advantage of Hawk becomes negative,

$$\begin{aligned}\pi_H - \pi_D &= \frac{1}{2}\left[b - \left(\frac{b}{c} + \epsilon\right)c\right] \\ &= \frac{1}{2}\left[b - b - c\epsilon\right] \\ &= -\frac{1}{2}c\epsilon < 0\end{aligned}$$

In other words, an increase in Hawks beyond the interior equilibrium will *disadvantage* Hawks relative to Doves. Therefore, selection will push the population back toward p^*. Likewise we see that if we decrease the population by a little bit, to $p^* - \epsilon$, the relative payoff to Hawk is positive so selection again pushes the population back toward p^*. This is a therefore a stable interior equilibrium and an example of **coexistence**.

By contrast, if we saw that the relative advantage of Hawks was positive at $p^* + \epsilon$ then selection would push the population toward $p = 1$. If we *also* saw that the relative advantage of Hawks was positive at $p^* - \epsilon$ (e.g. when the number of Doves increased a little bit) then we could conclude that selection always pushed the population toward $p = 1$ and this would be an example of **dominance**. Note that when $b > c$ we don't need to bother with this kind of calculation since p^* is greater than one by design. Lastly, if we saw that the relative advantage of Hawk to Dove was negative at $p^* - \epsilon$ *and* the relative advantage of Hawk to Dove was positive at $p^* + \epsilon$ then we have **bistability** – selection pushes away from the unstable equilibrium toward $p = 0$ or $p = 1$.

Relationship between ESS and NE. In an arbitrary symmetric 2x2 game with payoffs,



C is a (strict) NE if $w > y$. C is a (weak) NE if $w \geq y$. C is an ESS if 1. $w > y$ **or** 2. $w = y$ and $x > z$. So, *all (strict) NE are ESS and all ESS are (weak) NE.* ESS intuition: if there is neutrality between C and D when C is common, then C needs to beat D when D is common; otherwise a mutant D would be able to invade.

For example, in the Hawk-Dove game above, $\frac{(b-c)}{2} > 0$ when $b > c$, so Hawk is an ESS, Dove is not. However, if we instead assume that $b < c$ then neither is an ESS.

Relationship between ESS and Stability. There is a very useful relationship between stability and ESS. As we saw in the Hawk-Dove example with $c > b$, neither strategy was an ESS and we had a coexistence situation at $p^* = \frac{b}{c}$. We can in fact use ESS to predict outcomes. The table below gives a summary for the generic 2x2 payoff matrix with strategies C and D where p is the proportion playing C and $(1-p)$ is the proportion playing D,

	D is ESS	D is not ESS
C is ESS	$p^* \in (0,1)$ unstable	$p^* = 1$ stable
C is not ESS	$p^* = 0$ stable	$p^* \in (0,1)$ stable

Another way of thinking about this (discussed in Lecture 6) is to look directly at the payoffs. Using the arbitrary 2x2 matrix above we can make predictions about the three types of outcomes,

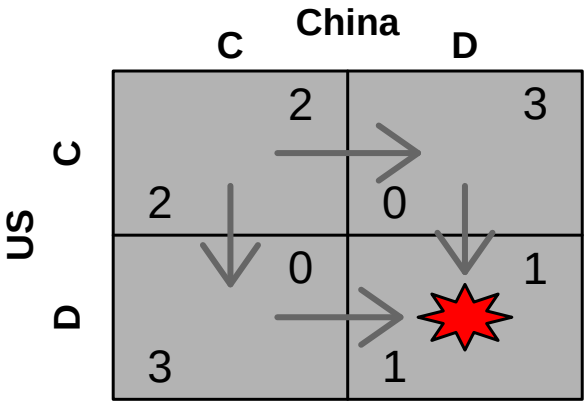
1) **Dominance** (for C): y < w and z < x
2) **Coexistence**: y > w and x > z
3) **Bistability**: y < w and x < z

Repeated games

- **Folk Theorem**: any level of cooperation can be achieved as a NE in repeated game provided players are "patient enough".
- Applied to *any* normal form game, not just the Prisoner's Dilemma.
- **Common student mistakes**: errors in calculating the continuation probability.

Typically, we let $w \in [0,1]$ denote the **probability of the next round occuring**. We assume it is **fixed** and **independent** across rounds. A math trick lets us express the **expected number of rounds** as $1/(1-w)$.

Illustration (greenhouse gas emissions):



Always cooperate v. Always defect:

- ALLC: C C C C C C = 0
- ALLD: D D D D D D = 18

Tit for Tat v. Always defect:

- TFT: C D D D D D = 5
- ALLD: D D D D D D = 8

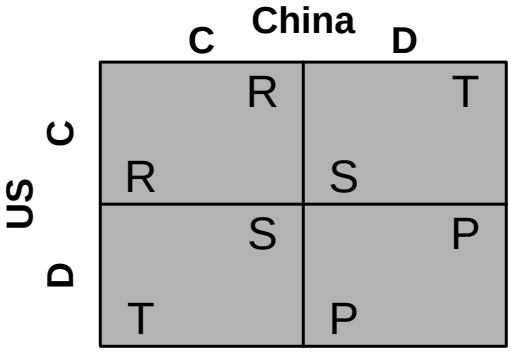
Tit for Tat v. Tit for Tat:

- TFT: C C C C C C = 12
- TFT: C C C C C C = 12

Grim Trigger v. Grim Trigger:

- *Expected payoffs to cooperation*: $\frac{2}{1-w}$
- *Expected payoffs for defection*: $3 + 1 \times \left(\frac{w}{1-w}\right)$
- NE if $\frac{2}{1-w} > 3 + \left(\frac{w}{1-w}\right)$ (if $w > 1/2$).
- More generally, we need $w > \frac{T-R}{T-P}$.

Generic Example: both play Tit-for-Tat (TFT)



- *Expected payoffs to cooperation*: $\frac{R}{1-w}$
- *Expected payoffs for defection in all periods*: $T + w\frac{P}{1-w}$

Social preferences

The *homo-economicus* story says humans only care about their individual payoffs. The **social preferences** story says humans also care about what happens to others (other regarding) and why these things happen (process regarding). Examples include altruism, fairness, reciprocity, and inequity aversion.

General setup:

- P1's utility: $U_1 = \pi_1 - \delta_1 \max(\pi_2 - \pi_1, 0) - \alpha_1 \max(\pi_1 - \pi_2, 0)$.
  - Note: $\max(3-2,0) = \max(-1,0) = 0$. [Why do we need this?]
- δ_1: how much P1 dislikes disadvantageous ($\pi_2 - \pi_1 > 0$) differences.
  - e.g. if $\delta_1 = 1$, P1 cares only about P2's payoffs if they are *more than* her own.
- α_1: how much P1 dislikes advantageous ($\pi_1 - \pi_2 > 0$) differences.
  - e.g. if $\alpha_1 = 1$ then P1 cares only about P2's payoffs if they *fall short of* her own.
- P2's utility: $U_2 = \pi_2 - \delta_2 \max(\pi_1 - \pi_2, 0) - \alpha_2 \max(\pi_2 - \pi_1, 0)$
- When $\delta_1 \neq \delta_2$ and/or $\alpha_1 \neq \alpha_2$, things get more complicated. Typically we assume symmetry!

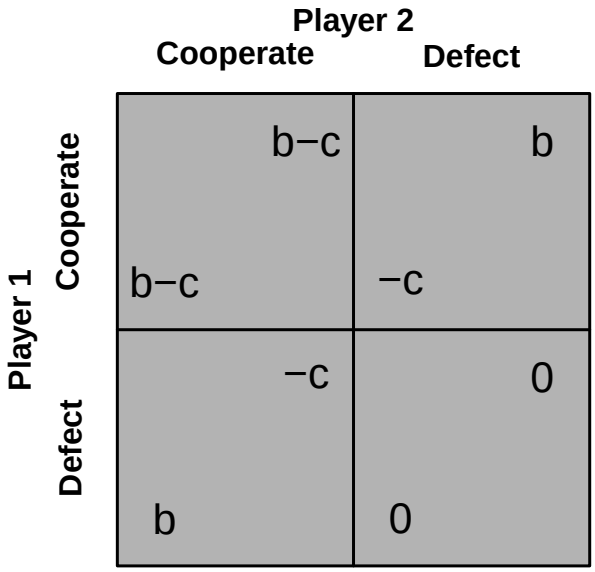
Disadvantageous Inequity Aversion: your utility decreases when someone else gets more than you. E.g. $U_1 = \pi_1 - \delta_1 \max(\pi_2 - \pi_1, 0)$

Advantageous Inequity Averson: your utility decreases when you get more than someone else. E.g. $U_1 = \pi_1 - \alpha_1 \max(\pi_1 - \pi_2, 0)$

Illustration (classic game theory):

Suppose $\alpha,\ \delta$ are same for both players. Therefore,

- P1's utility: $U_1 = \pi_1 - \delta \max(\pi_2 - \pi_1, 0) - \alpha \max(\pi_1 - \pi_2, 0)$.
- P2's utility: $U_2 = \pi_2 - \delta \max(\pi_1 - \pi_2, 0) - \alpha \max(\pi_2 - \pi_1, 0)$ If $\alpha = \delta = 0$ then players only care about their individual payoffs. Consider the game below,



Generic payoff structure for both players (by symmetry):

$$
\begin{aligned}
U(C,C) &= (b-c) - \delta \max\{(b-c)-(b-c),0\} - \alpha \max\{(b-c)-(b-c),0\} \\
&= b-c \\
U(D,C) &= b - \delta \max\{-c-b,0\} - \alpha \max\{b-(-c),0\} \\
&= b - \alpha(b+c) \\
U(C,D) &= -c - \delta \max\{b-(-c),0\} - \alpha \max\{-c-b,0\} \\
&= -c - \delta(b+c) \\
U(D,D) &= 0 - \delta \max\{0,0\} - \alpha \max\{0,0\} \\
&= 0
\end{aligned}
$$

Q: What do we need for CC to be NE?

$$1.\ \ U_1(C,C) > U_1(D,C);\ \ 2.\ \ U_2(C,C) > U_2(C,D)$$

Condition 1. implies [Note: we don't need to calculate 2. here, why?],

$$
\begin{aligned}
b - c > b - \alpha(b+c) &= -c > -\alpha b - \alpha c \\
&= \alpha(b+c) > c \\
&= \alpha > \frac{c}{b+c}
\end{aligned}
$$

So only **advantageous inequity aversion** (α) matters here: if players don't like getting more than each other (α sufficiently large) then we can have cooperation in social dilemma. Social preferences (in this case inequity aversion) can change Social Dilemma to Coordination Game.