Poisoning Attacks in Games Example

-

June 14, 2021

1 Example

Consider the repeated game with the following stage game,

Actions	Accept	Reject
Accept	θ, θ	$0, 0$
Reject	$0, 0$	$\text{-}\theta, -\theta$

Assume a common prior,

$$\theta = \begin{cases} 1 & \text{with probability } p = \frac{1}{2} \\ -1 & \text{with probability } 1 - p = \frac{1}{2} \end{cases}. \tag{1}$$

Now suppose the reward function is given by,

$$R_i^t\left(\theta = 1\right) = \begin{cases} 1 & \text{with probability } q = \frac{2}{3} \\ -1 & \text{with probability } 1 - q = \frac{1}{3} \end{cases}, \tag{2}$$

and,

$$R_i^t\left(\theta = -1\right) = \begin{cases} 1 & \text{with probability } 1 - q = \frac{1}{3} \\ -1 & \text{with probability } q = \frac{2}{3} \end{cases}. \tag{3}$$

In a world with $\theta = -1$,
In stage 1, with probability $q^2 = \frac{1}{9}$, we have $R_i^1\left(\theta = -1\right) = 1$, and the posterior belief that $\theta = 1$ is,

$$\mathbb{P}_i\left\{\theta = 1 | R_i^1 = 1\right\} = \frac{pq}{pq + (1-p)(1-q)} = \frac{2}{3} > \frac{1}{2}. \tag{4}$$

This implies that both players will select the action Accept.
In stage 2 after the history (Accept, Accept), the posterior belief of both players that $\theta = -1$ is,

$$\mathbb{P}_i\left\{\theta = 1 | R_i^2 = -1, h_1 = (\text{ Accept }, \text{ Accept })\right\} = \frac{pq^2(1-q)^1}{pq^2(1-q)^1 + (1-p)(1-q)^2 q^1} = \frac{2}{3} > \frac{1}{2}, \tag{5}$$

and,

$$\mathbb{P}_i\left\{\theta = 1 | R_i^2 = 1, h_1 = (\text{ Accept }, \text{ Accept })\right\} = \frac{pq^3(1-q)^0}{pq^3(1-q)^0 + (1-p)(1-q)^3 q^0} = \frac{8}{9} > \frac{1}{2}. \tag{6}$$

This implies that it doesn't matter what the rewards are in the second stage, both players will select the action Accept.
The same implications hold for $t > 2$. In general, we have,

$$\mathbb{P}_i\left\{\theta = 1 | R_i^t = -1, h_1 = ... = h_{t-1} = (\text{ Accept }, \text{ Accept })\right\} \geqslant \frac{pq^t (1-q)^{t-1}}{pq^t (1-q)^{t-1} + (1-p)(1-q)^t q^{t-1}} = \frac{2}{3} > \frac{1}{2}, \tag{7}$$

and,

$$\mathbb{P}_i\left\{\theta = 1 | R_i^t = 1, h_1 = ... = h_{t-1} = (\text{ Accept }, \text{ Accept })\right\} \geqslant \frac{pq^{t+1} (1-q)^{t-2}}{pq^{t+1} (1-q)^{t-2} + (1-p)(1-q)^{t+1} q^{t-2}} = \frac{8}{9} > \frac{1}{2}. \tag{8}$$

Since player i only observes the past actions of player -i, but not the actual reward, player i makes the decision based on $t-1$ Accepts, 1 negative reward in stage 1, and at most $t-1$ positive rewards. As a result, the posterior belief that $\theta = 1$ is always at least $\frac{2}{3}$, which implies that it doesn't matter what the rewards are in each stage, both players will select the action Accept.

Note that the players will always choose Accept in a world with $\theta = -1$ in which the unique Nash Equilibrium is (Reject, Reject) in every stage just because the rewards are inconsistent with the state of the world in the first stage. In a more general setting with any p and $q > \frac{1}{2}$, there exists a finite T such that if the reward is different from θ in the first T periods for both players, both players will always choose the non-NE action in all future periods after T.

From the perspective from the adversary, they only have to perturb the reward distribution in the first stage (or first finite number of stages) to ensure that the action pair (Accept, Accept) is chosen, the same action pair will be used in all future stages.

2 Uniqueness of Nash Equilibrium

2.1 Two by Two Games

A game with a unique Nash Equilibrium that is not a dominant strategy equilibrium.

Action	A	B
A	$(1, 1)$	$(0, 0)$
B	$(0, 1)$	$(1, 0)$

For a general two player two by two game, if it has a unique pure strategy Nash Equilibrium, then at least one player is using a dominant strategy.

Action	A	B
A	(a, b)	(c, d)
B	(e, f)	(g, h)

Proof: without loss of generality, suppose (A, A) is the unique pure strategy Nash Equilibrium. It implies that $a \geqslant e$ and $b \geqslant d$. Since (B, B) is not a Nash Equilibrium, either $c > g$ or $f > h$, and in either case A is

a dominant strategy for one of the players.

For a symmetric two player two by two game, if it has a unique pure strategy Nash Equilibrium, then it is a dominant strategy equilibrium.

For a symmetric two player two by two game, it has a unique strict pure strategy Nash Equilibrium iff it is a strictly dominant strategy equilibrium.

A game with a dominant strategy equilibrium that is not the unique Nash Equilibrium.

Action	A	B
A	$(1,1)$	$(1,0)$
B	$(1,1)$	$(0,0)$

For a general two player two by two game specified below, it has a unique (completely) mixed strategy Nash Equilibrium iff $(b-d)(f-h)<0, (a-e)(c-g)<0$ and either $(b-d)(a-e)<0$ or $(f-h)(c-g)<0$.

Action	β	$1-\beta$
α	(a, b)	(c, d)
$1-\alpha$	(e, f)	(g, h)

Proof: suppose the game has a mixed strategy Nash Equilibrium in which the row player mixes action A with probability α and the column player mixes action A with probability β, then it is given by the indifference conditions:

$$\alpha b + (1-\alpha) f = \alpha d + (1-\alpha) h,$$
$$\beta a + (1-\beta) c = \beta e + (1-\beta) g.$$

Then we have,

$$\alpha = \frac{h-f}{h-f+b-d} \in (0,1),$$
$$\beta = \frac{a-e}{a-e+g-c} \in (0,1).$$

If $h > f$, then $\alpha \in (0,1)$ implies $b > d$, and since (B, B) is not a Nash Equilibrium, $c > g$.
If $f > h$, then $\alpha \in (0,1)$ implies $d > b$, and since (B, A) is not a Nash Equilibrium, $a > e$.
If $a > e$, then $\beta \in (0,1)$ implies $g > c$, and since (A, A) is not a Nash Equilibrium, $d > b$.
If $e > a$, then $\beta \in (0,1)$ implies $c > g$, and since (A, B) is not a Nash Equilibrium, $h > f$.
Therefore, we have $(b-d)(f-h)<0, (a-e)(c-g)<0$ and $(b-d)(a-e)<0$.

For a general two player two by two game specified below, it has a unique (completely) mixed strategy Nash Equilibrium if $(b-d)(f-h)\neq 0, (a-e)(c-g)\neq 0$ and either $(b-d)(a-e)<0$ or $(f-h)(c-g)<0$.

Action	β	$1-\beta$
α	(a, b)	(c, d)
$1-\alpha$	(e, f)	(g, h)

Proof: if $(b-d)(f-h) > 0$, the column player has a strictly dominant strategy, $a \neq e$ and $c \neq g$ implies that there is a unique pure strategy Nash.
if $(a-e)(c-g) > 0$, the row player has a strictly dominant strategy, $b \neq d$ and $f \neq h$ implies that thre is a unique pure strategy Nash.
if $(b-d)(f-h) < 0$ and $(a-e)(c-g) < 0$, then there is a unique mixed strategy Nash Equilibrium.