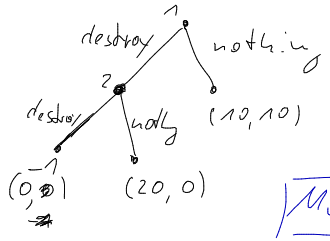
MAD (mutually assured destruction)

- If player 1 does nothing, then everybody lives happily ever after.
- If player 1 destroys the country of player 2, then player 2 has a threat to retaliate.
- However, no soldier has any incentive to do so after his country is destroyed.



"destroy" by player 2 is a non-credible threat,
→ Make "destroy" an automatic response.

Movie: Dr. Strangelove

	nothing	destroy
nothing	10, 10	10, 10
destroy	20, 0	0, -1

Subgame-perfect Equilibrium

Let $\Gamma = \langle N, A, H, P, (u_i) \rangle$ be an EGWPI.

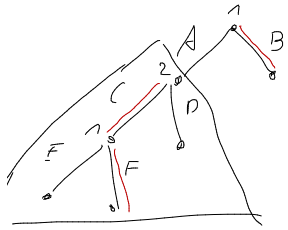
Def (Subgame)

The subgame of Γ rooted at history h is the EGWPI $\Gamma(h) = \langle N, A, H|_h, P|_h, (u_i|_h) \rangle$ where

$$H|_h = \{ h' \mid (h, h') \in H \}$$

$$P|_h(h') = P((h, h'))$$

$$u_i|_h(h') = u_i((h, h')) \text{ for all } (h, h') \in Z$$



$h = (A)$

Strategies relativized to histories.

For each strategy s_i in Γ, let $s_i|_h(h') := s_i((h, h'))$

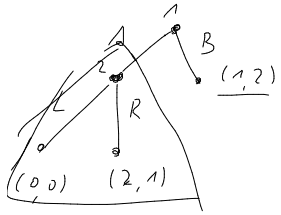
The outcome function of $\Gamma(h)$ is denoted by O_h.

Def (SPE)

A subgame-perfect equilibrium (SPE) of a EGWPI Γ is a strategy profile $s^* = (s_i^*)_{i \in N}$ such that for each history $h \in H$:

$$s^*|_h := (s_i^*|_h)_{i \in N}$$

is a NE of $\Gamma(h)$.



	L	R
A	0,0	2,1
B	1,2	1,2

$s = (A, R)$

$s = (\{\emptyset \mapsto A\}, \{(A) \mapsto R\})$

$s|_{(A)} = (\{\emptyset \to A\}|_{(A)}, \{(A) \mapsto R\}|_{(A)})$

$= (\{\}, \{\emptyset \mapsto R\})$

$H = \{\emptyset, (A), (B), (A, L), (A, R)\}$

$H|_{(A)} = \{\emptyset, (L), (R)\}$

$\Rightarrow s = (A, R)$

$h = \emptyset$, s is a NE

$h = (A)$, $s|_{(A)}$ is a NE

$\Rightarrow$ s is a SPE

$s = (B, L)$

$h = \emptyset$ s is a NE

$h = (A)$ s is not NE

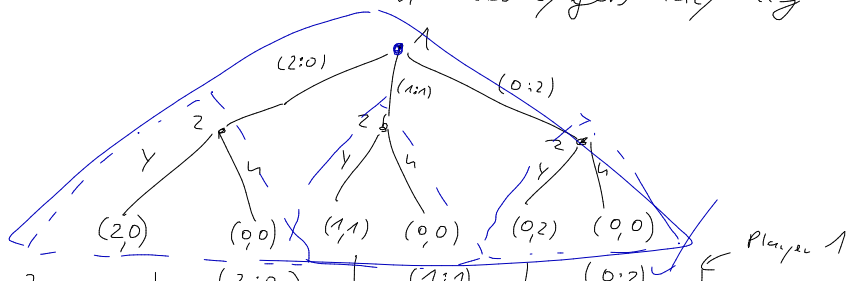
$\rightarrow$ s is not a SPE!

Example (Sharing game)

Actions for player 1: (2:0), (1:1), (0:2)

Actions for player 2: y - agree to split

n - nobody gets anything



Player 2	(2:0)	(1:1)	(0:2)
yyy	SPE	X	X
yyn	—	—	—
yny	—	—	—
ynn	—	—	—
nyy	X	SPE	X
nyn	—	—	—
nny	—	—	—
nnn	—	—	—

Questions

- Does an SPE always exist?
- Under which conditions?
- How to compute it?
- What is the complexity?

We show:

- It is easy to verify that a profile is an SPE
  $\Rightarrow$ "one deviation property" (for finite horizon games)
- For finite games, we can easily compute the SPE by "backward induction" (Kuhn's Theorem)

Notation: If Γ is an EGWPI then $\ell(\Gamma)$ denotes the length of the longest history in Γ.

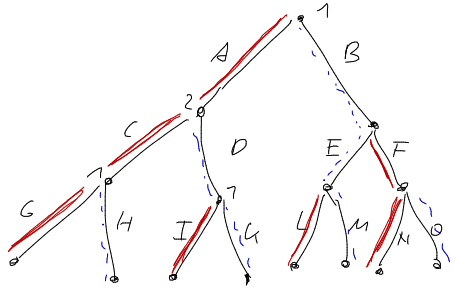
Lemma (one deviation property)

Let $\Gamma = \langle N, A, H, P, (u_i) \rangle$ be a finite horizon EGWPS. Then a strategy profile s^* is an SPE of Γ if and only if for every player $i \in N$ and every history $h \in H$ for which $P(h) = i$, we have:

$$u_i|_h \left(O_h \left(s^*_{-i}|_h, s^*_i|_h \right) \right) \geq u_i|_h \left(O_h \left(s^*_{-i}|_h, s_i \right) \right)$$

for every strategy s_i of player i in the subgame $\Gamma(h)$ that differs from $s^*_i|_h$ only in the action after the initial history of $\Gamma(h)$.

(without the underlined parts, it is just the def of SPE)



Red denote a strategy profile

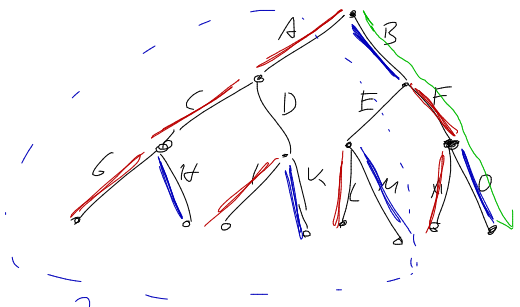
---- denotes the branches we have to check in order to verify that we an SPE

Proof

$\Rightarrow$: obvious.

$\Leftarrow$: By contradiction

Suppose that s^* is not an SPE.
Then there is a history h and a player i such that s_i is a profitable deviation for player i in the subgame $\Gamma(h)$.
WLOG the number of histories h' with $s_i(h') \neq s^*_i|_h(h')$ is at most $\boxed{\ell(\Gamma(h))}$ and hence finite (finite horizon assumption!), since deviations not on the resulting outcome path are irrelevant.



original strategy profile

deviation: BCFHLNO

Outcome

irrelevant part

There is another deviation:

BCFGILO

Choose profitable deviation s_i in $\Gamma(h)$ with minimal number of deviation points. Let h^* be longest history in $\Gamma(h)$ with $s_i(h^*) \neq s_i^*|_h(h^*)$, i.e. "deepest" deviation point for s_i.

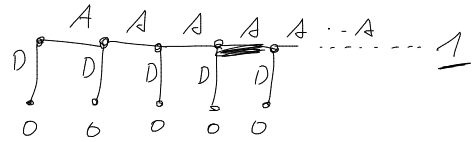
Then in $\Gamma((h,h^*))$, $s_i|_{h^*}$ differs from $s_i^*|_{(h,h^*)}$ only in the initial history. Moreover, $s_i|_{h^*}$ is a profitable deviation in $\Gamma(h,h^*)$, since h^* is the longest history in $\Gamma(h)$ with $s_i(h^*) = s_i^*|_h(h^*)$.

So $\Gamma(h,h^*)$ is the desired subgame where a one-step deviation is sufficient to improve the utility.

$\square$

The corresponding proposition for <u>infinite horizon games</u> does not hold.

Counter-example



Strategy s_i with $s_i(h) = D$ for all $h \in H \setminus Z$

- satisfies "one deviation property", but
- is not an SPE, since it is dominated by s_i^* with $s_i^*(h_1) = A$ for all $h \in H \setminus Z$