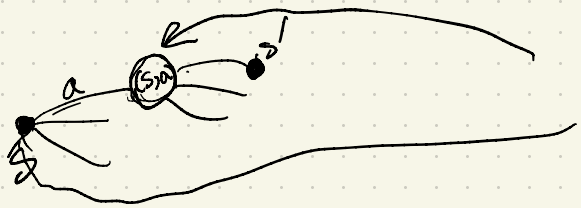


Q-value for MDP
V-value

max min chance

Learning algorithm

Episode:

$$s_0; a_1, r_1, s_1; a_2, r_2, s_2, a_3, r_3, s_3; \ldots, a_n, r_n, s_n$$

A small piece of experience:

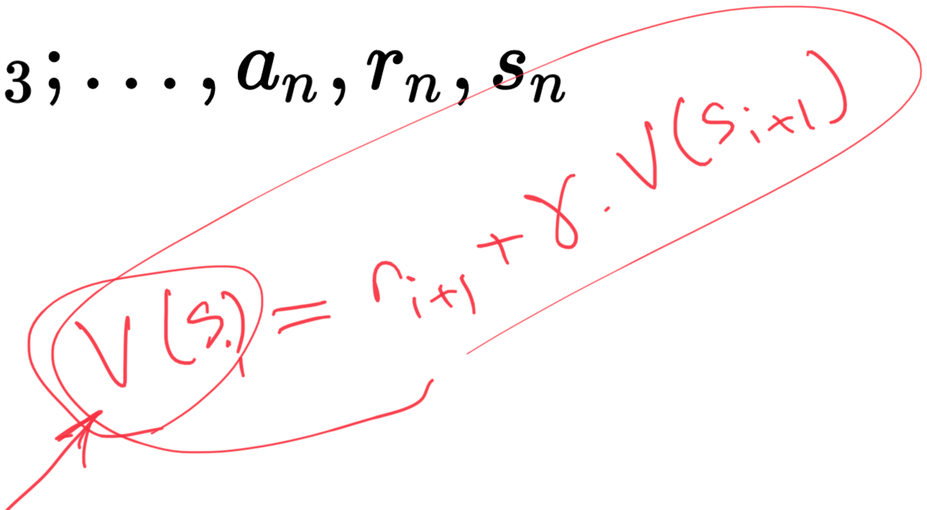
$$(s, a, r, s')$$

Prediction:

$$V(s; \mathbf{w})$$

Target:

$$r + \gamma V(s'; \mathbf{w})$$



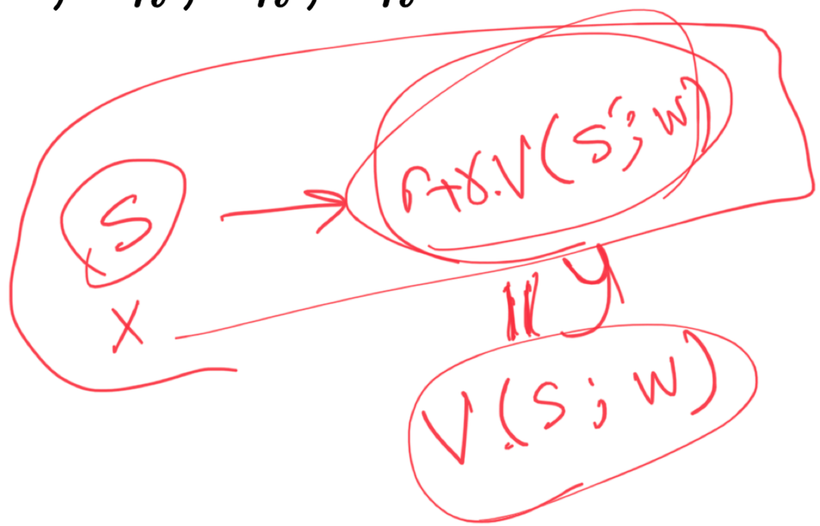
Learning algorithm

Episode:

$$s_0; a_1, r_1, s_1; a_2, r_2, s_2, a_3, r_3, s_3; \ldots, a_n, r_n, s_n$$

A small piece of experience:

$$(s, a, r, s')$$



Prediction:

$$V(s; \mathbf{w})$$

Target:

$$r + \gamma V(s'; \mathbf{w})$$

Learning algorithm

Episode:

$$s_0; a_1, r_1, s_1; a_2, r_2, s_2, a_3, r_3, s_3; \ldots, a_n, r_n, s_n$$

A small piece of experience:

$$(s, a, r, s')$$

Prediction:

$$V(s; \mathbf{w})$$

Target:

$$r + \gamma V(s'; \mathbf{w})$$

Example of TD learning

Step size $\eta = 0.5$, discount $\gamma = 1$, reward is end utility

Example: TD learning

S1		**r:0**	**S4**		**r:0**	**S8**		**r:1**	**S9**
$\phi:\begin{pmatrix}0\\1\end{pmatrix}$			$\phi:\begin{pmatrix}1\\0\end{pmatrix}$			$\phi:\begin{pmatrix}1\\2\end{pmatrix}$			$\phi:\begin{pmatrix}1\\0\end{pmatrix}$
$\mathbf{w}:\begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:0 p-t:0		$\mathbf{w}:\begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:0 p-t:0		$\mathbf{w}:\begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:1 p-t:-1		$\mathbf{w}:\begin{pmatrix}0.5\\1\end{pmatrix}$

S1		**r:0**	**S2**		**r:0**	**S6**		**r:0**	**S10**
$\phi:\begin{pmatrix}0\\1\end{pmatrix}$			$\phi:\begin{pmatrix}1\\0\end{pmatrix}$			$\phi:\begin{pmatrix}0\\0\end{pmatrix}$			$\phi:\begin{pmatrix}1\\0\end{pmatrix}$
$\mathbf{w}:\begin{pmatrix}0.5\\1\end{pmatrix}$	p:1 t:0.5 p-t:0.5		$\mathbf{w}:\begin{pmatrix}0.5\\0.75\end{pmatrix}$	p:0.5 t:0 p-t:0.5		$\mathbf{w}:\begin{pmatrix}0.25\\0.75\end{pmatrix}$	p:0 t:0.25 p-t:-0.25		

$s \to s'$

$p = w \cdot \phi(s) = V(s;w)$

$t = r + V(s';w)$

$= r + w \cdot \phi(s')$

Example of TD learning

$\text{Loss} = \frac{1}{2}(\text{pred}(w) - \text{target})^2$

Step size $\eta = 0.5$, discount $\gamma = 1$, reward is end utility

Example: TD learning

S1	**r:0**	**S4**	**r:0**	**S8**	**r:1**	**S9**	
$\phi: \begin{pmatrix}0\\1\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$		$\phi: \begin{pmatrix}1\\2\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$	
$\mathbf{w}: \begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:0 p-t:0	$\mathbf{w}: \begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:0 p-t:0	$\mathbf{w}: \begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:1 p-t:-1	$\mathbf{w}: \begin{pmatrix}0.5\\1\end{pmatrix}$	
S1	**r:0**	**S2**	**r:0**	**S6**	**r:0**	**S10**	
$\phi: \begin{pmatrix}0\\1\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$		$\phi: \begin{pmatrix}0\\0\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$	
$\mathbf{w}: \begin{pmatrix}0.5\\1\end{pmatrix}$	p:1 t:0.5 p-t:0.5	$\mathbf{w}: \begin{pmatrix}0.5\\0.75\end{pmatrix}$	p:0.5 t:0 p-t:0.5	$\mathbf{w}: \begin{pmatrix}0.25\\0.75\end{pmatrix}$	p:0 t:0.25 p-t:-0.25		

Example of TD learning

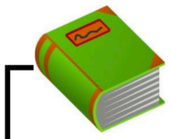
$V(s;w) \simeq V(s';w)+r$

Step size $\eta = 0.5$, discount $\gamma = 1$, reward is end utility

Example: TD learning

S1	**r:0**	**S4**	**r:0**	**S8**	**r:1**	**S9**	
$\phi: \begin{pmatrix}0\\1\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$		$\phi: \begin{pmatrix}1\\2\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$	
$\mathbf{w}: \begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:0 p-t:0	$\mathbf{w}: \begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:0 p-t:0	$\mathbf{w}: \begin{pmatrix}0\\0\end{pmatrix}$	p:0 t:1 p-t:-1	$\mathbf{w}: \begin{pmatrix}0.5\\1\end{pmatrix}$	
S1	**r:0**	**S2**	**r:0**	**S6**	**r:0**	**S10**	
$\phi: \begin{pmatrix}0\\1\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$		$\phi: \begin{pmatrix}0\\0\end{pmatrix}$		$\phi: \begin{pmatrix}1\\0\end{pmatrix}$	
$\mathbf{w}: \begin{pmatrix}0.5\\1\end{pmatrix}$	p:1 t:0.5 p-t:0.5	$\mathbf{w}: \begin{pmatrix}0.5\\0.75\end{pmatrix}$	p:0.5 t:0 p-t:0.5	$\mathbf{w}: \begin{pmatrix}0.25\\0.75\end{pmatrix}$	p:0 t:0.25 p-t:-0.25		

Game evaluation



Definition: game evaluation

The **value** of the game if player A follows π_{A} and player B follows π_{B} is

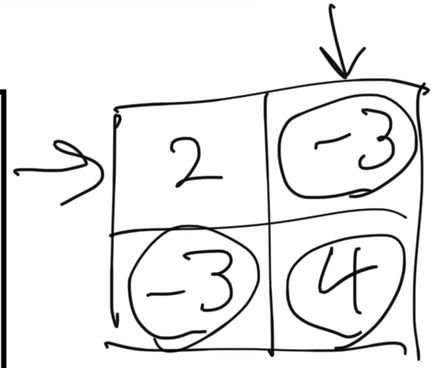
$$V(\pi_{\mathrm{A}}, \pi_{\mathrm{B}}) = \sum_{a,b} \pi_{\mathrm{A}}(a)\pi_B(b)V(a,b)$$

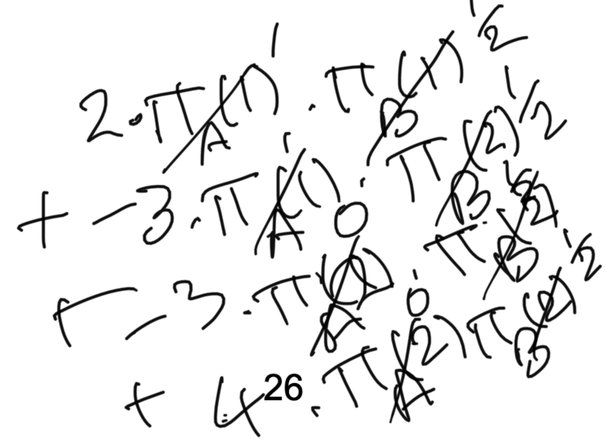
Example: two-finger Morra

Player A always chooses 1: $\pi_{\mathrm{A}} = [1, 0]$

Player B picks randomly: $\pi_{\mathrm{B}} = [\frac{1}{2}, \frac{1}{2}]$

Value: $\boxed{-\frac{1}{2}}$





[whiteboard: matrix]

Mixed strategies

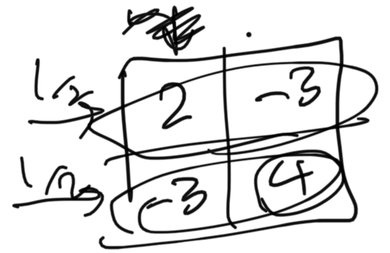
Example: two-finger Morra

Player A reveals: $\pi_{\mathrm{A}} = [\frac{1}{2}, \frac{1}{2}]$

Value $V(\pi_{\mathrm{A}}, \pi_B) = \pi_B(1)(-\frac{1}{2}) + \pi_B(2)(+\frac{1}{2})$

Optimal strategy for player B is $\pi_B = [1, 0]$ (**pure!**)

Mixed strategies



Example: two-finger Morra

Player A reveals: $\pi_{\mathrm{A}} = [\frac{1}{2}, \frac{1}{2}]$

Value $V(\pi_{\mathrm{A}}, \pi_B) = \pi_B(1)(-\frac{1}{2}) + \pi_B(2)(+\frac{1}{2})$

Optimal strategy for player B is $\pi_B = [1, 0]$ (**pure!**)

Proposition: second player can play pure strategy

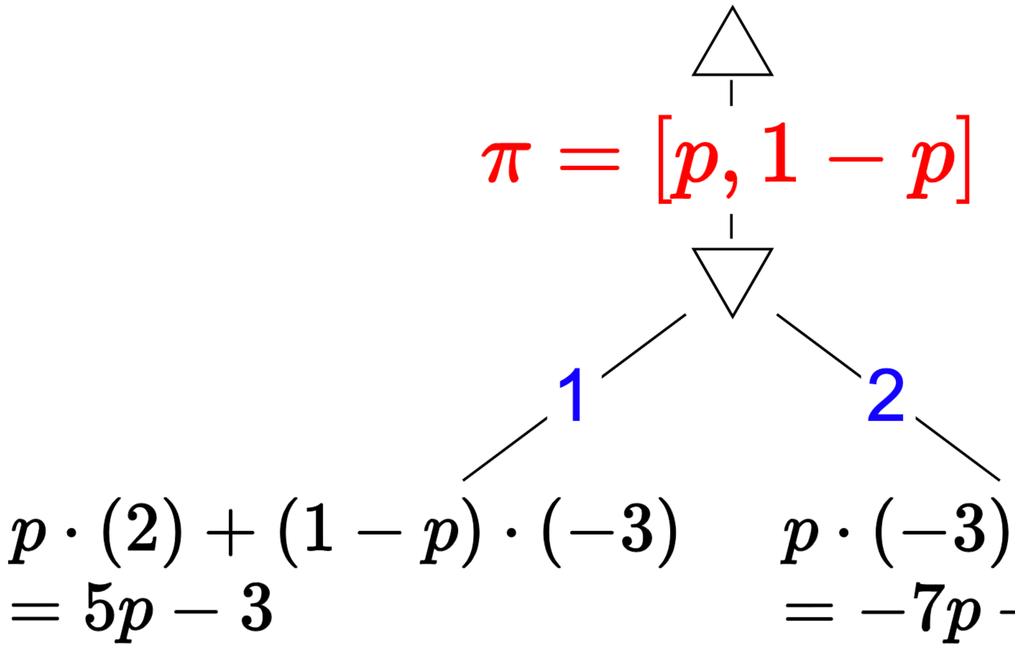
For any fixed mixed strategy π_{A}:

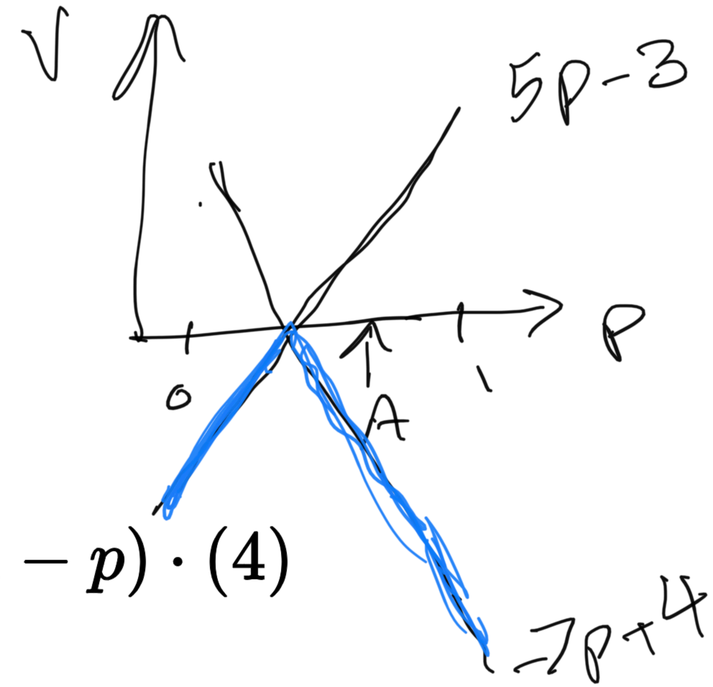
$$\min_{\pi_{\mathrm{B}}} V(\pi_{\mathrm{A}}, \pi_B)$$

can be attained by a pure strategy.

Mixed strategies

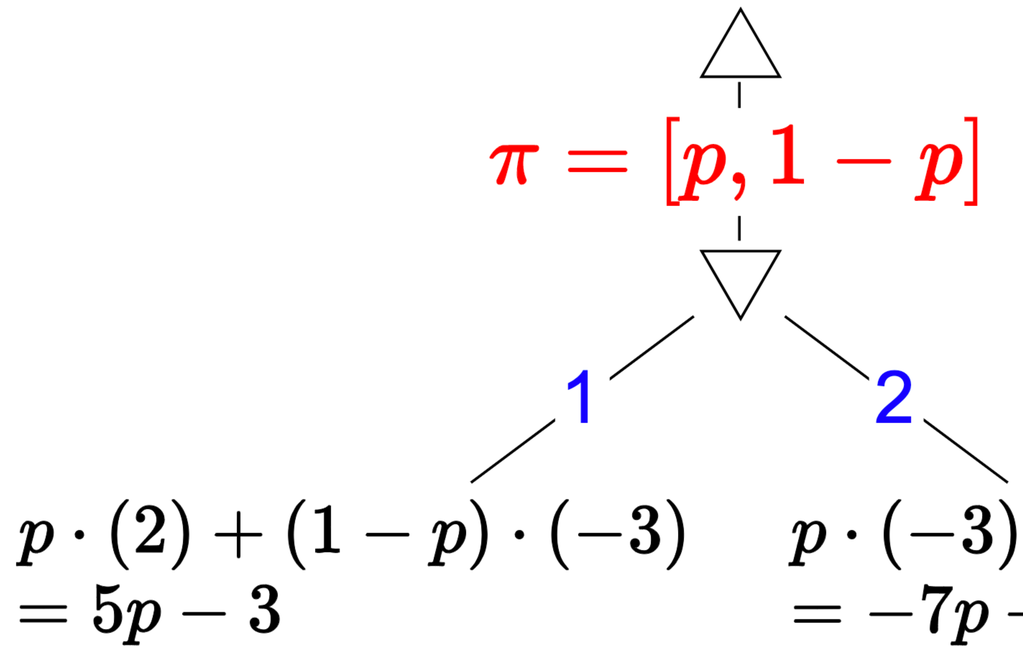
Player A first reveals his/her mixed strategy

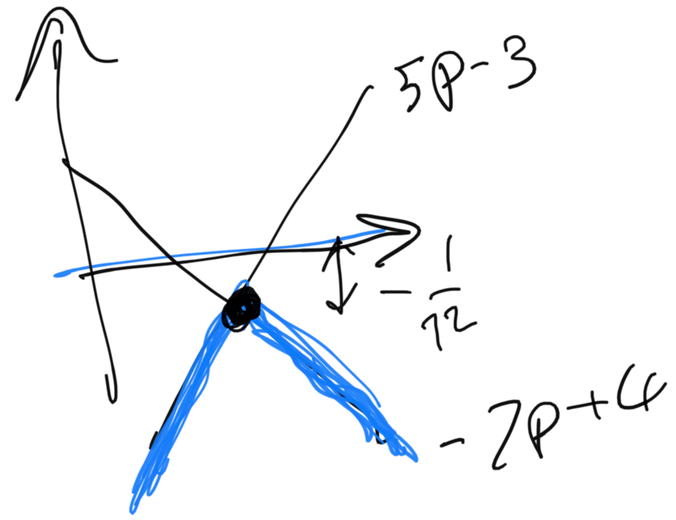




Mixed strategies

Player A first reveals his/her mixed strategy



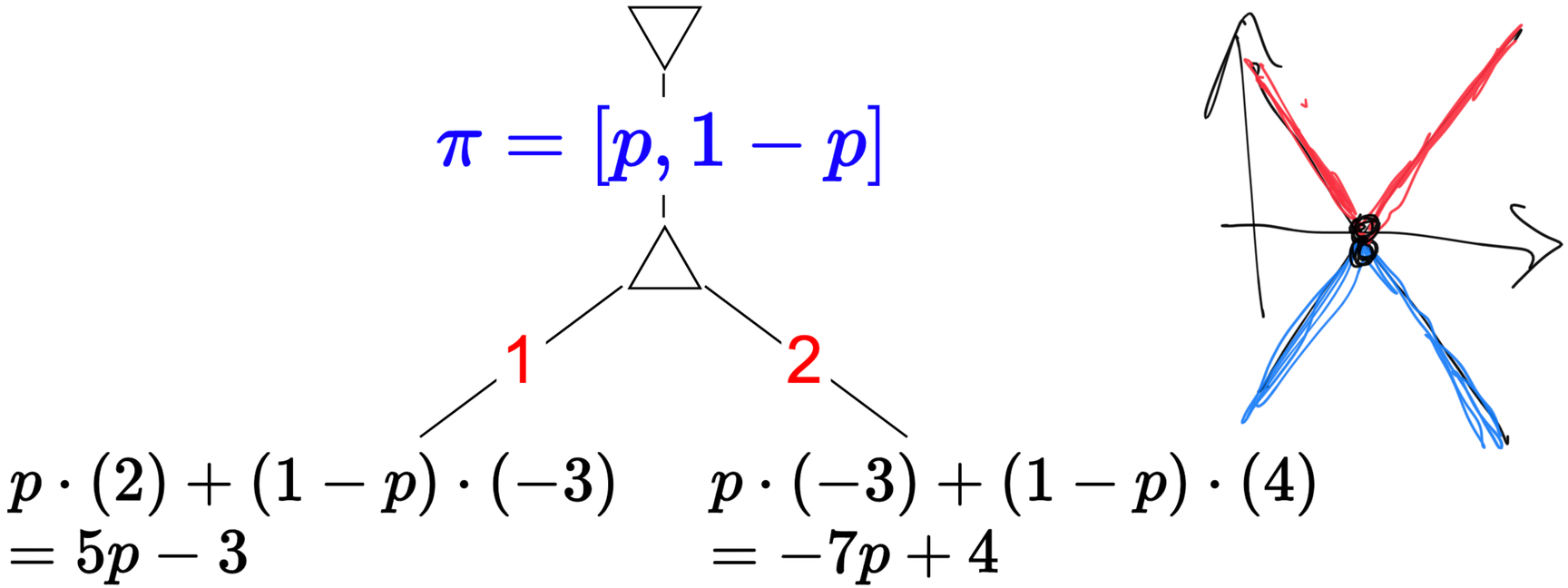


Minimax value of game:

$$\max_{0 \le p \le 1} \min\{5p - 3, -7p + 4\} = \boxed{-\frac{1}{12}} \text{ (with } p = \frac{7}{12}\text{)}$$

Mixed strategies

Player B first reveals his/her mixed strategy



Minimax value of game:

$$\min_{p \in [0,1]} \max\{5p - 3, -7p + 4\} = \boxed{-\frac{1}{12}} \text{ (with } p = \frac{7}{12}\text{)}$$

General theorem

Theorem: minimax theorem [von Neumann, 1928]

For every simultaneous two-player zero-sum game with a finite number of actions:

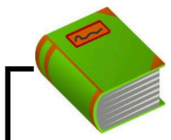
$$\max_{\pi_A}\left(\min_{\pi_B} V(\pi_A, \pi_B)\right) = \min_{\pi_B}\left(\max_{\pi_A} V(\pi_A, \pi_B)\right),$$

where π_A, π_B range over **mixed strategies**.

Prisoner's dilemma

Example: payoff matrix

B \ A	testify	refuse
testify	$A = -5, B = -5$	$A = -10, B = 0$
refuse	$A = 0, B = -10$	$A = -1, B = -1$



Definition: payoff matrix

Let $V_p(\pi_{\mathrm{A}}, \pi_{\mathrm{B}})$ be the utility for player p.

General theorem

Theorem: minimax theorem [von Neumann, 1928]

For every simultaneous two-player zero-sum game with a finite number of actions:

$$\max_{\pi_A} \min_{\pi_B} V(\pi_A, \pi_B) \leq \min_{\pi_B} \max_{\pi_A} V(\pi_A, \pi_B),$$

where π_A, π_B range over **mixed strategies**.

mixed

pure

Upshot: revealing your optimal mixed strategy doesn't hurt you!

Proof: linear programming duality