Strategic Value – Normal Form, 2-player games

2-player zero-sum games

- As we’ve seen, the minimax value of each player is the value each player receives in all Nash equilibria
- Can be interpreted as the strategic value of each player
- Q. Can we generalize this to general strategic games? And what is the right generalization?

Motivating Example

		You: reach	You: climb
Friend	don't boost	$0, 2$	$0, 0$
	boost	$0, 2$	$0, 4$

Figure 1. A banana-picking game.

Motivating Example

		You	
		reach	climb
Friend	don't boost	$0, 2$	$0, 0$
	boost	$0, 2$	$0, 4$

Figure 1. A banana-picking game.

Q. How to encourage cooperation?

Q. How to define the strategic "strength" of each player?

Motivating Example

		You	
		reach	climb
Friend	don't boost	$0, 2$	$0, 0$
	boost	$0, 2$	$0, 4$

Figure 1. A banana-picking game.

If you agree to give your friend a “side payment”, both of you can improve your payoff. But how much?

Definition of Value

$$
\begin{aligned}
&\mathrm{Coco}(U, \overline{U}) \\
&\quad = \quad \mathrm{maxmax}((U + \overline{U})/2) + \mathrm{minmax}((U - \overline{U})/2).
\end{aligned}
$$

For banana game, this recommends (1,3) split of the four bananas.

Desirable Properties

- Pareto efficiency
- Shift invariance
- Monotonicity in actions
- Payoff dominance
- Invariance to redundant strategies

Desirable Properties

- Pareto efficiency
- Shift invariance
- Monotonicity in actions
- Payoff dominance
- Invariance to redundant strategies

Theorem: COCO is unique value satisfying all of these properites.

Strategic Values – Stochastic Games (2 players)

Stochastic Games

Definition 6.2.1 (Stochastic game) *A* stochastic game *(also known as a* Markov game*) is a tuple* (Q, N, A, P, r)*, where:*

- Q *is a finite set of games;*
- N *is a finite set of* n *players;*
- $A = A_1 \times \cdots \times A_n$*, where* A_i *is a finite set of actions available to player* i*;*
- $P : Q \times A \times Q \mapsto [0, 1]$ *is the transition probability function;* $P(q, a, \hat{q})$ *is the probability of transitioning from state* q *to state* $\hat{q}$ *after action profile* a*; and*
- $R = r_1, \ldots, r_n$*, where* $r_i : Q \times A \mapsto \mathbb{R}$ *is a real-valued payoff function for player* i*.*

How to generalize COCO definition to stochastic games (with discounted payoff?)

Generalized Q-learning

$$\langle s, a, \overline{a}, r, \overline{r}, s' \rangle$$

$$\begin{aligned} Q'_s &= Q_s + \alpha(r + \gamma \bigotimes(Q_{s'}, \overline{Q}_{s'}) - Q_s), \\ \overline{Q}'_s &= \overline{Q}_s + \alpha(\overline{r} + \gamma \bigotimes(\overline{Q}_{s'}, Q_{s'}) - \overline{Q}_s). \end{aligned}$$

If $\bigotimes$ is a contraction $|\otimes f - \otimes f'| \leq \mathrm{maxmax}|f - f'|$, guaranteed to converge to solution:

$$Q_s(a, \overline{a}) = R_s(a, \overline{a}) + \gamma \sum_{s'} T(s, a, \overline{a}, s') \otimes (Q_{s'}, \overline{Q}_{s'}),$$

Definition of Value

$$\langle s, a, \overline{a}, r, \overline{r}, s' \rangle$$

$$Q'_s = Q_s + \alpha_{a,\overline{a}}(r + \gamma \,\mathrm{Coco}(Q_{s'}, \overline{Q}_{s'}) - Q_s);$$

$$\overline{Q}'_s = \overline{Q}_s + \alpha_{a,\overline{a}}(\overline{r} + \gamma \,\mathrm{Coco}(\overline{Q}_{s'}, Q_{s'}) - \overline{Q}_s);$$

But COCO is not a contraction...

COCO Definition

$$Z_s = (Q_s - \overline{Q}_s)/2 \qquad C_s = (Q_s + \overline{Q}_s)/2.$$

$$\langle s, a, \overline{a}, r, \overline{r}, s' \rangle$$

$$Z'_s = Z_s + \alpha_{a,\overline{a}}((r - \overline{r})/2 + \gamma \,\mathrm{minmax}(Z_{s'}) - Z_s);$$

$$C'_s = C_s + \alpha_{a,\overline{a}}((r + \overline{r})/2 + \gamma \,\mathrm{maxmax}(C_{s'}) - C_s).$$

Claim: If $Z = (Q - \overline{Q}) / 2$ and $C = (Q + \overline{Q}) / 2$, then relationship is preserved after update.

Proof of Claim

$$
\begin{aligned}
&(Q'_s - \overline{Q}'_s)/2 \\
&= (Q_s + \alpha_{a,\overline{a}}(r + \gamma\,\mathrm{Coco}(Q_{s'}, \overline{Q}_{s'}) - Q_s))/2 - \\
&\quad (\overline{Q}_s + \alpha_{a,\overline{a}}(\overline{r} + \gamma\,\mathrm{Coco}(\overline{Q}_{s'}, Q_{s'}) - \overline{Q}_s))/2 \\
&= (Q_s - \overline{Q}_s)/2 + \alpha_{a,\overline{a}}((r - \overline{r})/2 + \\
&\quad \gamma\,\mathrm{minmax}((Q_{s'} - \overline{Q}_{s'})/2) - (Q_s - \overline{Q}_s)/2) \\
&= Z_s + \alpha_{a,\overline{a}}((r - \overline{r})/2 + \gamma\,\mathrm{minmax}(Z_{s'}) - Z_s) \\
&= Z'_s.
\end{aligned}
$$

And Z, C sequences converge (minmax, maxmax are contractions!). Therefore Q values converge.

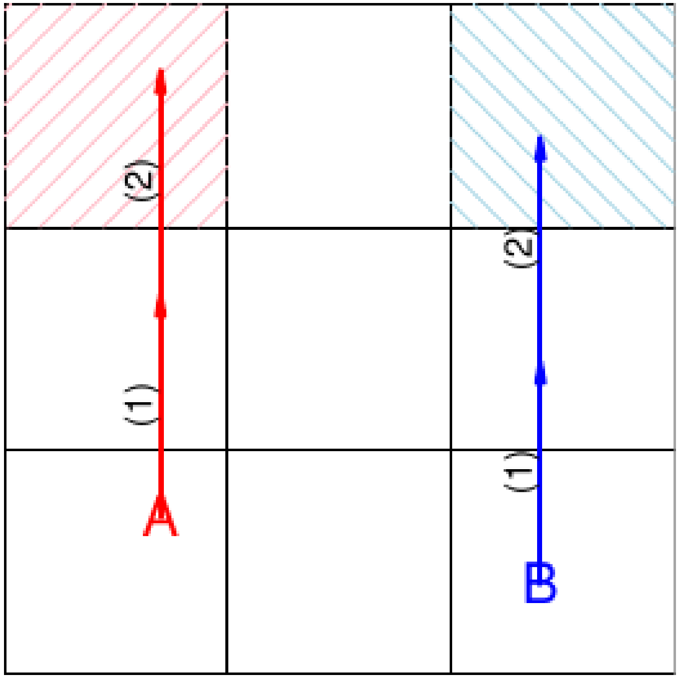
Grid Games

- Played on *m x k* squares
- Each agent has a designated starting square, and a set of goal squares where rewards are received
- Agents observe their own position, location of walls, semi-walls, and other agents
- All agents simultaneously choose an action from {up,down,right,left,stick}
- Every action except stick incurs a step cost

Grid Games

- If an agent's selected move is unimpeded, the agent moves in the direction selected
- If trying to move through a wall, or to a spot occupied by an agent who sticks, move fails
- If agent tries to move through semi-wall, it succeeds with probability *p*
- If multiple agents try to move to same square, a uniformly selected agent succeeds
- Game ends if one agent reaches a goal

Example

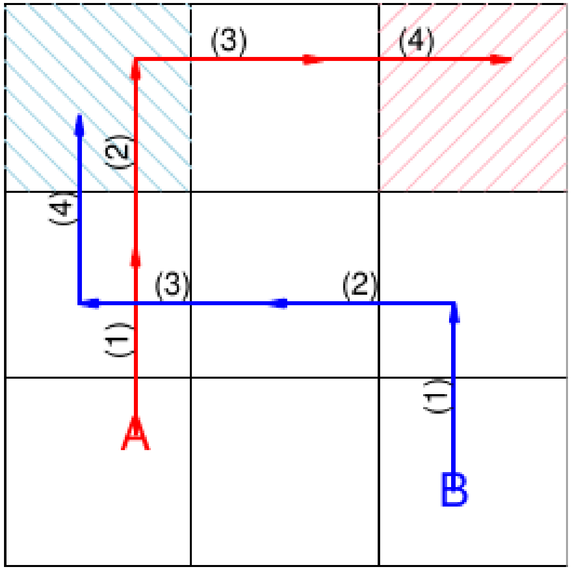


(a) An example trajectory for an example game

Coordination

- 3 x 3 grid. Player A starts in bottom left, B starts in bottom right
- Goal of A is in top right, goal of B is in top left
- Players must coordinate how to pass one another.
- Q: Does COCO policy optimally coordinate?

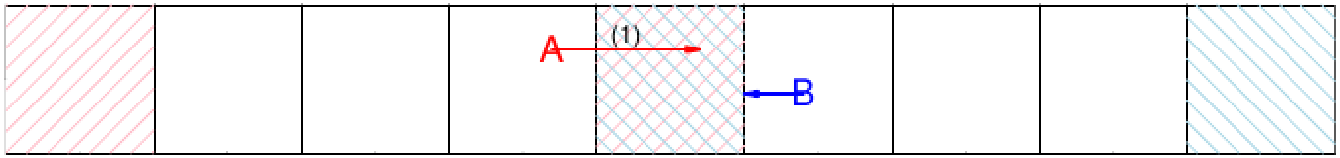
Coordination



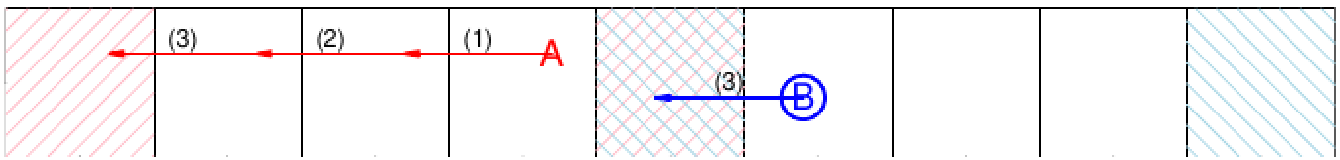
Step	P_A
1	0.00
2	−0.50
3	0.50
4	0.00

(b) COCO on Coordination. $\gamma = 1$.

Prisoners' Dilemma



(a) Correlated-VI on Prisoner

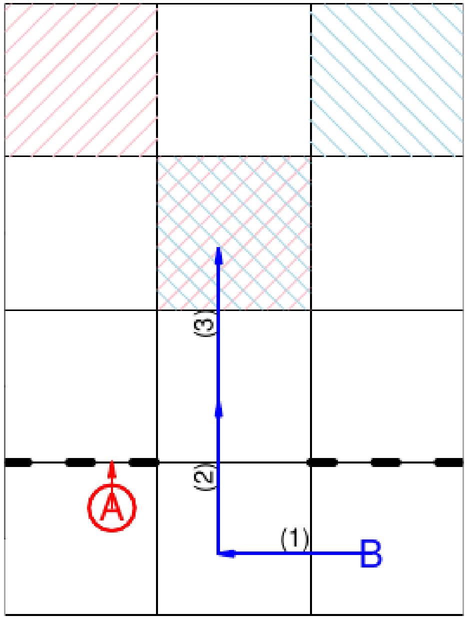


Step	P_A
1	50.00
2	−49.00
3	0.00

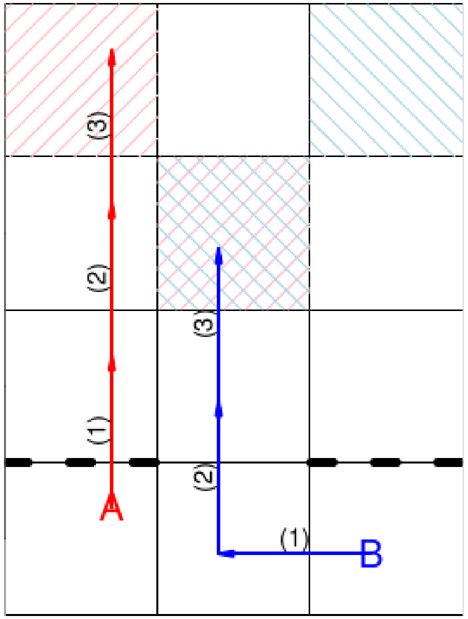
(b) COCO on Prisoner. $\gamma = 1$.

Figure 5. A Correlated-VI trajectory and the unique COCO trajectory in the Prisoner grid game.

Turkey



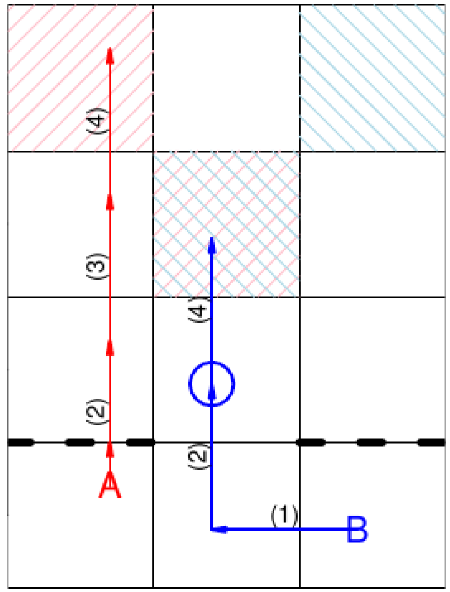
(a) Correlated-VI on Turkey. $V_A = 43.2$, while $V_B = 87.4$.



Step	P_A
1	22.10
2	0.00
3	0.00

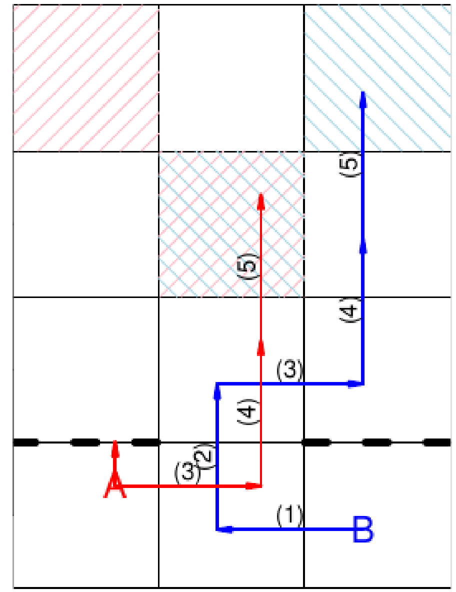
(b) COCO on Turkey, trajectory 1

Turkey



Step	P_A
1	22.10
2	0.50
3	−49.00
4	0.00

(c) COCO on Turkey, trajectory 2



Step	P_A
1	22.10
2	0.50
3	−49.50
4	0.00
5	0.00

(d) COCO on Turkey, trajectory 3

Turkey

Trajectory	Probability	A	B
(b)	0.5	109.5	65.3
(c)	0.25	60.4	104.6
(d)	0.25	54.8	99.0
Expected Value	—	83.55	83.55

Summary

- COCO learns sensible policies (in fact, optimal collaborative policy)
- Side payments incentivize the collaboration in an interpretable way on a move-by-move basis
- Players with symmetric position get equal expected payoff