Multi-agent RL Models

Adam Gleave
adam@gleave.me
18th October 2017

University of California Berkeley

Papers

Presentation based on:

- Joel Z. Leibo et al. (2017). "Multi-agent Reinforcement Learning in Sequential Social Dilemmas". In: *AAMAS*
- Julien Pérolat et al. (2017). "A multi-agent reinforcement learning model of common-pool resource appropriation". In: *CoRR* abs/1707.06600. Accepted to NIPS '17

Preview: Sequential Social Dilemmas

- Leibo et al. (2017) introduce sequential social dilemmas (SSD), a type of multi-player Markov game.
- Generalization of one-shot social dilemmas matrix games, e.g. Prisoner's dilemma.
- Agents represented by a deep Q-network (DQN).
- Two grid-world environments:
  - Gathering: two players gathering 'apples', can attack opponent.
  - Wolfpack: two 'wolves' hunting a 'prey'.
- Experiments vary parameters of the environments, and the agents.

Preview: Common Pool Resources

- Pérolat et al. (2017) study an N-player variant of Gathering.
- Rate at which 'apples' regenerate depends on stock level.
- Proposes metrics for measuring the social outcome of the game.
- Experiments again vary parameters of both the environment and the agents.

Sequential Social Dilemmas

Social dilemmas

- Social dilemmas: conflict between collective and individual rationality.
- Canonical example is the Prisoner's dilemma matrix game:

	C	D
C	$3, 3$	$0, 4$
D	$4, 0$	$1, 1$

Matrix Game Social Dilemma

Macy and Flache (2002) define a matrix game social dilemma (MGSD) as a two-player matrix game:

	C	D
C	R, R	S, T
D	T, S	P, P

with:

- $R > P$: mutual cooperation preferred to mutual defection.
- $R > S$: mutual cooperation preferred to exploitation by a defector.
- $2R > T + S$: mutual cooperation preferred to equal probability of unilateral cooperation and defection.
- either greed, $T > R$, or fear, $P > S$.

Example MGSDs

	C	D
C	$3,3$	$1,4$
D	$4,1$	$0,0$

(a) Chicken

	C	D
C	$4,4$	$0,3$
D	$3,0$	$1,1$

(b) Stag Hunt

	C	D
C	$3,3$	$0,4$
D	$4,0$	$1,1$

(c) Prisoner's Dilemma

satisfy the MGSD conditions:

	C	D
C	R, R	S, T
D	T, S	P, P

where:

- $R > P$.
- $R > S$.
- $2R > T + S$.
- either greed, $T > R$, or fear, $P > S$.

Limitations of MGSDs

Matrix games ignore many aspects we may want to model:

- Cooperation or defection are labels for sequences of actions in a temporally extended game, not a one-shot decision.
- Cooperativeness can be a graded rather than binary quantity.
- Decisions to cooperate or defect occur only quasi-simultaneously.
- Imperfect information: the state of the world and other player's action only partially observable.

Leibo et al. (2017) propose a Sequential Social Dilemma model to better capture these aspects.

Partially-observable Markov Game

An N-player partially observable Markov game $\mathcal{M}$ is defined by:

- a set of states $\mathcal{S}$.
- a set of actions $\mathcal{A}_i$ for each player i.
- a transition function, determined by the current state and actions of each player, of the form $\tau : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_N \to \Delta(\mathcal{S})$, where $\Delta(S)$ is the set of discrete probability distributions over $\mathcal{S}$.
- a reward function for each player i, $r_i : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_N \to \mathbb{R}$.
- an observation function $O : \mathcal{S} \times \{1, \ldots, N\} \to \mathbb{R}^d$, where $O(s, i)$ gives player i's view of state s.

This paper only considers the two-player case ($N = 2$).

Value in a Matrix Game

Define, for discount factor $\gamma \in [0, 1)$, the payoff for player i under joint policy $\vec{\pi} = (\pi_1, \pi_2)$ as:

$$V_i^{\vec{\pi}} = \mathbb{E}_{\vec{a}_t \sim \vec{\pi}(O(s_t)), s_{t+1} \sim \tau(s_t, \vec{a}_t)} \left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, \vec{a}_t) \right].$$

Can represent a choice between two policies π^C and π^D, starting in state s, as the matrix game:

	C	D
C	$R(s), R(s)$	$S(s), T(s)$
D	$T(s), S(s)$	$P(s), P(s)$

where:

$$R(s) = V_1^{\pi^C, \pi^C}(s) = V_2^{\pi^C, \pi^C}(s), \qquad P(s) = V_1^{\pi^D, \pi^D}(s) = V_2^{\pi^D, \pi^D}(s),$$
$$S(s) = V_1^{\pi^C, \pi^D}(s) = V_2^{\pi^D, \pi^C}(s), \qquad T(s) = V_1^{\pi^D, \pi^C}(s) = V_2^{\pi^C, \pi^D}(s).$$

Sequential Social Dilemma

A sequential social dilemma is a tuple $(\mathcal{M}, \Pi^C, \Pi^D)$ where:

- $\mathcal{M}$ is a Markov game with state space $\mathcal{S}$.
- Π^C and Π^D are disjoint sets of policies representing 'cooperation' and 'defection'.
- There exists $s \in \mathcal{S}$ and $\pi^C \in \Pi^C, \pi^D \in \Pi^D$ for which the induced matrix game is an MGSD.

Simulation Environment

- Game engine: 2D deterministic gridworld.
- Observations $O(s, i) \in \mathbb{R}^{3 \times 16 \times 21}$ are RGB bitmap of window 15 squares ahead of players and 10 grid squares from side to side.
- Eight agent-centered actions: step forward/backward/left/right, rotate left/right, use beam and stand still.
- Player blue in local view, light-blue teammate view, red in opponent's view.
- Each episode lasts 1,000 steps.

Agents

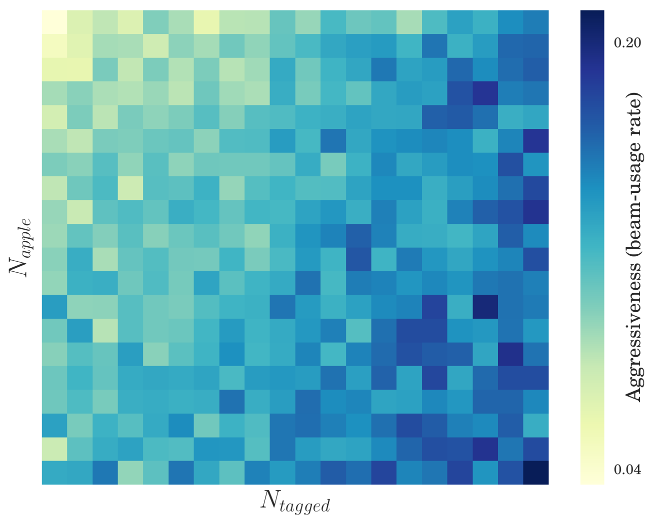
- Each agent is represented by a deep Q-network (DQN).
- The DQN for agent i represents a function $Q_i : \mathcal{O}_i \times \mathcal{A}_i \rightarrow \mathbb{R}$.
- During learning, take optimal action according to Q_i with probability $1 - \epsilon$ and a random action with probability ϵ.
- DQN updated based on a batch of the 10^5 last observations.
- Trained through gradient descent on mean squared Bellman residual, taken over transitions uniformly sampled from the batch.
- Two hidden layers with 32 units, interleaved with rectified linear layers projecting to the output layer with 8 units.
- ϵ decays linearly over time from 1.0 to 0.1.
- Per-step time discount rate of $\gamma = 0.99$.
- No theory of mind.

Environment: Gathering

- Two players, competing over scarce resources (‘apples’).
- Player receives reward 1 if moves to a square containing an apple. Apple is removed from the game and respawns after a constant N_{apple} steps.
- Players can shoot a beam in a straight line along their current orientation.
- If a player is hit twice by a beam, removed from the game for N_{tagged} frames.[1]

[1] Respawns from the far left.

Experiment: Varying Gathering Parameters

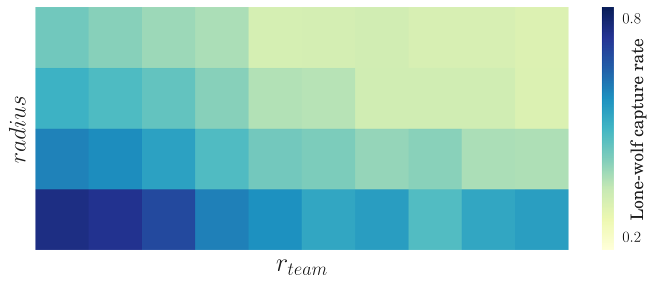


Wolfpack

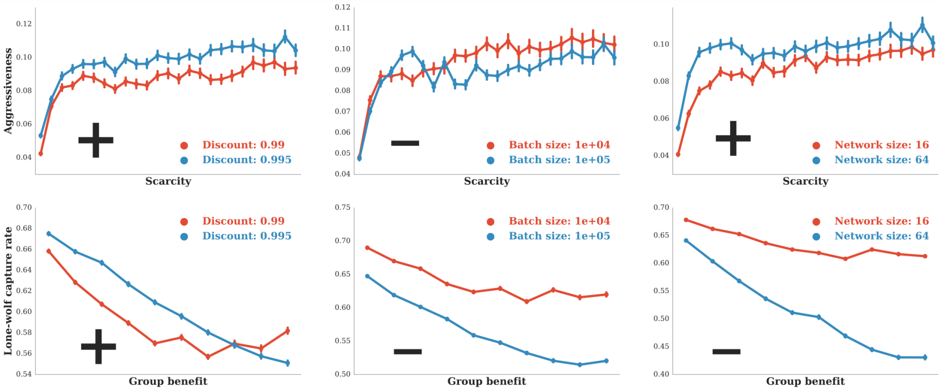
- Two players ('wolves') chase a 'prey' agent.
- Episode ends when either wolf touches the prey.
- All wolves within the *capture radius* receive reward.
- If only one wolf in capture radius, it receives r_{lone}.
- If two wolves in capture radius, they both receive r_{team}.

Wolfpack

- Two players ('wolves') chase a 'prey' agent.
- Episode ends when either wolf touches the prey.
- All wolves within the *capture radius* receive reward.
- If only one wolf in capture radius, it receives r_{lone}.
- If two wolves in capture radius, they both receive r_{team}.

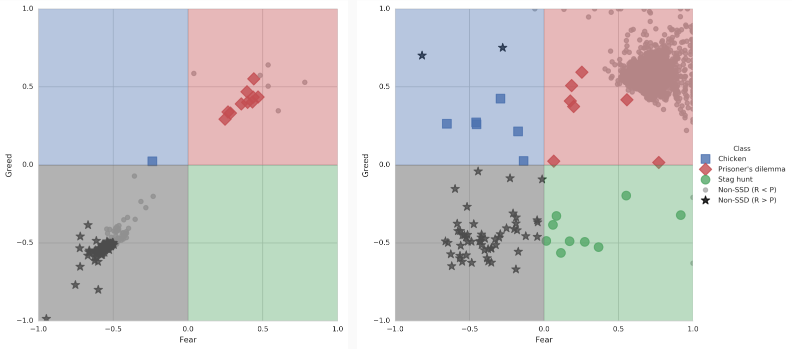


Experiment: Varying Agent Parameters



Top: Gathering. Bottom: Wolfpack.

Experiment: Induced Matrix Games



(a) Gather **(b)** Wolfpack

Common-Pool Resource Appropriation

Common-Pool Resource

A good is:

- excludable if it is possible to limit access to those who have paid for it;
- rivalrous if consumption by one person prevents simultaneous consumption by others.

Definition matrix:

	Excludable	Non-excludable
Rivalrous	**Private goods**	**Common-pool resources**
Non-rivalrous	**Club goods**	**Public goods**

New Gathering Environment

Pérolat et al. (2017) uses a variant of the Gathering environment we saw in Leibo et al. (2017). Similar to before:

- Player receives reward 1 when it collects an apple.
- Agents are DQNs trained in the same way.
- Episodes are still 1000 steps.

Different to before:

- Apple regrowth rate depends on the number of uncollected apples nearby (previously constant).
- If a player is hit by the time-out beam, it is immediately removed from the game for 25 time steps (previously needed to be hit twice).
- Any number of players N (previously $N = 2$).
- Map of the grid world specifying size and initial apple placement varies.

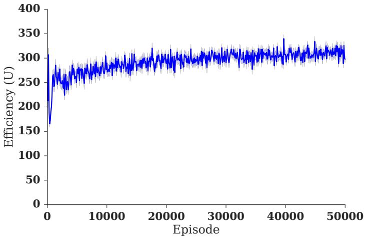
Note the last two criteria vary between experiments in the paper.

Social Outcome Metrics

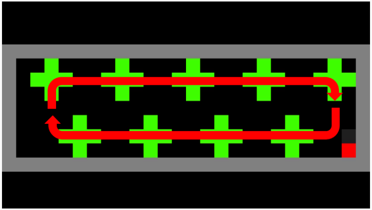
For a system with N independent agents, let $\left(r_t^i\right)_{t=1}^{t=T}$ and $\left(o_t^i\right)_{t=1}^{t=T}$ be the sequence of rewards and observations obtained by the i-th agent over an episode of duration T. Its total reward is given by $R^i = \sum_{t=1}^{T} r_t^i$. Define social outcome metrics:

Utilitarian	$U = \frac{1}{T}\mathbb{E}\left[\sum_{i=1}^{N} R^i\right]$
Equality	$E = 1 - \mathbb{E}\left[\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left\lvert R^i - R^j\right\rvert}{2N\sum_{i=1}^{N} R^i}\right]$
Sustainability	$S = \frac{1}{N}\sum_{i=1}^{N} t^i, \quad \text{where } t^i = \mathbb{E}\left[t \mid r_i^t > 0\right]$
Peace	$P = \frac{\mathbb{E}\left[NT - \sum_{i=1}^{N}\sum_{t=1}^{T} I\left(o_t^i\right)\right]}{T}, \quad \text{where } I(o) = \begin{cases} 1 & o \text{ time-out} \\ 0 & \text{otherwise} \end{cases}.$

Sanity Check: Does a Single Agent Cooperate?



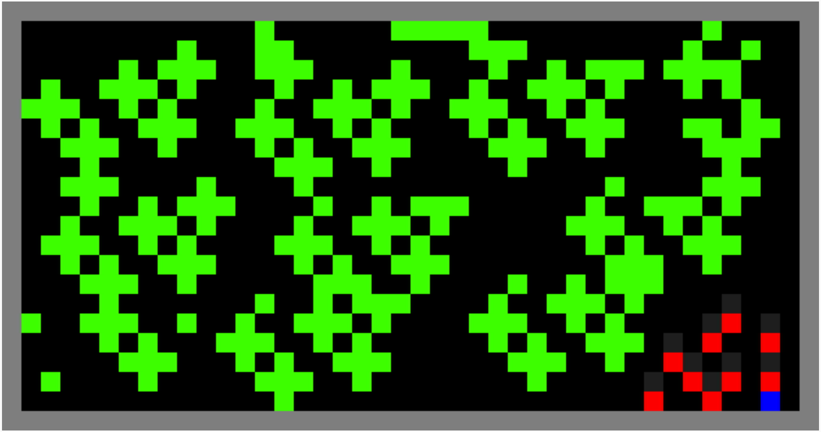
(a) Single agent return



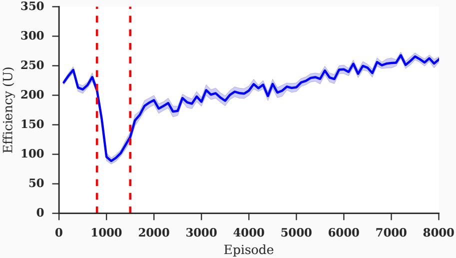
(b) Optimal path

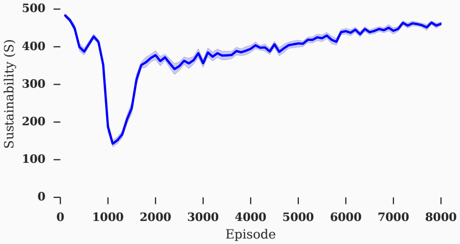
Open Map Experiment: Environment

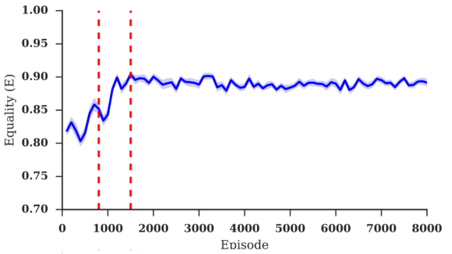
10 agents in a map with uniformly distributed apples.

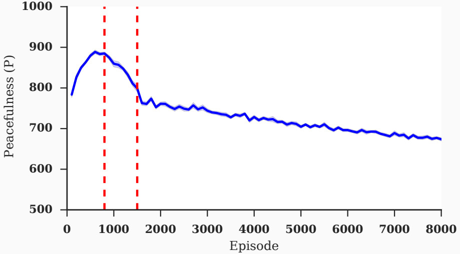


Open Map Experiment: Social Outcomes

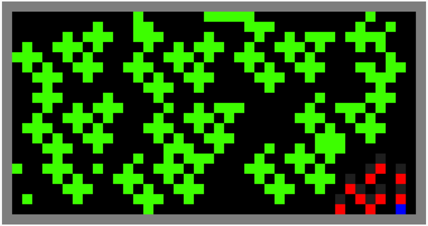








Open Map Experiment: Videos



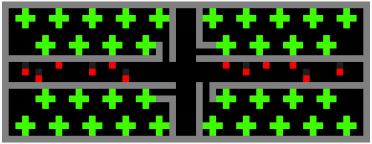
(a) Map

(b) Naivety

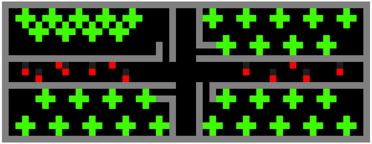
(c) Tragedy

(d) Maturity

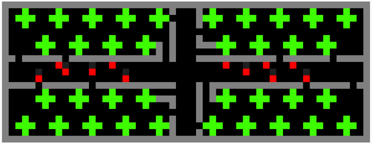
Walled Map Experiment: Environment



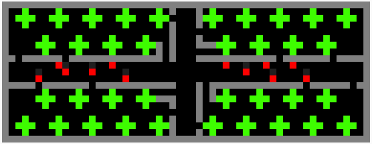
(a) Basic single entrance



(b) Unequal single entrance

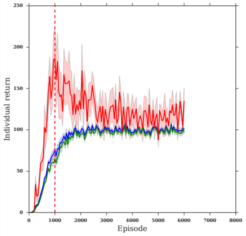


(c) Multiple entrances

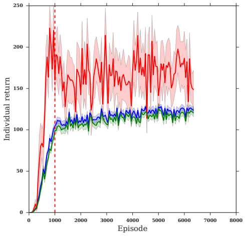


(d) No walls

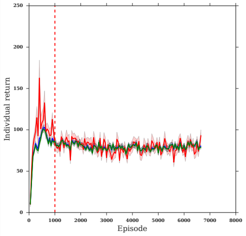
Walled Map Experiment: Expected Return



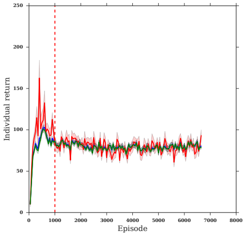
(a) Basic single entrance



(b) Unequal single entrance

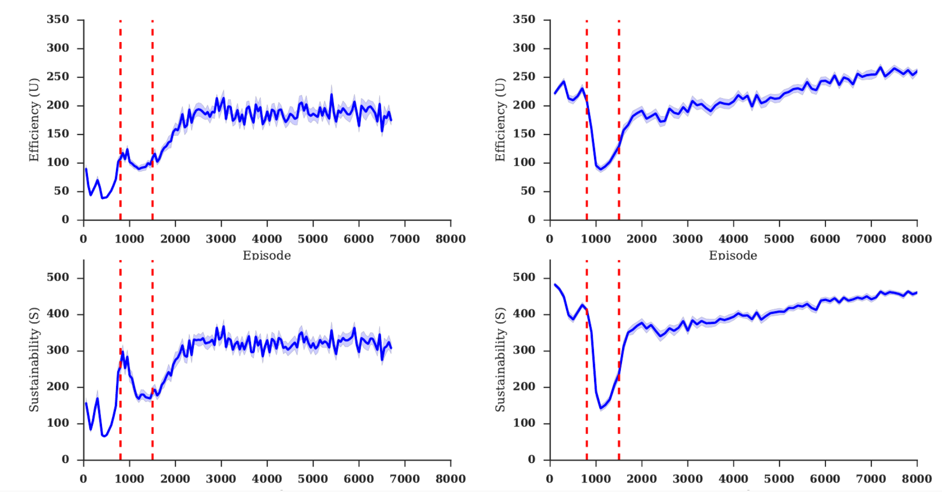


(c) Multiple entrances

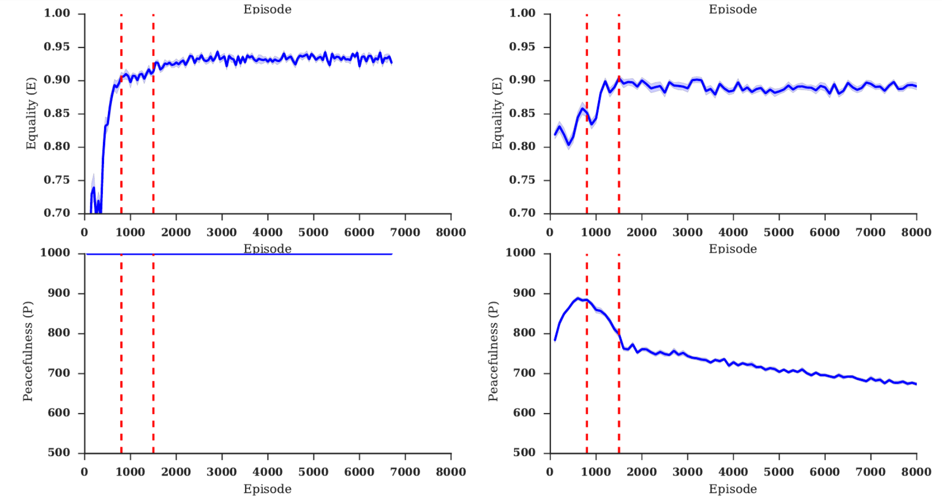


(d) No walls

Removing Tagging: Efficiency and Sustainability



Removing Tagging: Equality and Peacefulness



References

Slides: `goo.gl/EHGA5e`.

Leibo, Joel Z., Vinícius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel (2017). "Multi-agent Reinforcement Learning in Sequential Social Dilemmas". In: *AAMAS*.

Macy, Michael W. and Andreas Flache (2002). "Learning dynamics in social dilemmas". In: *Proceedings of the National Academy of Sciences* 99.suppl 3, pages 7229–7236.

Pérolat, Julien, Joel Z. Leibo, Vinícius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel (2017). "A multi-agent reinforcement learning model of common-pool resource appropriation". In: *CoRR* abs/1707.06600. Accepted to NIPS '17.