DeepMind

Temporal Difference and Return Optimism in Cooperative Multiagent Reinforcement Learning

Mark Rowland, Shayegan Omidshafiei, Daniel Hennes,
Will Dabney, Andrew Jaegle, Paul Muller, Julien Perolat, Karl Tuyls

AAMAS ALA Workshop 2021

Setting and overview of contributions

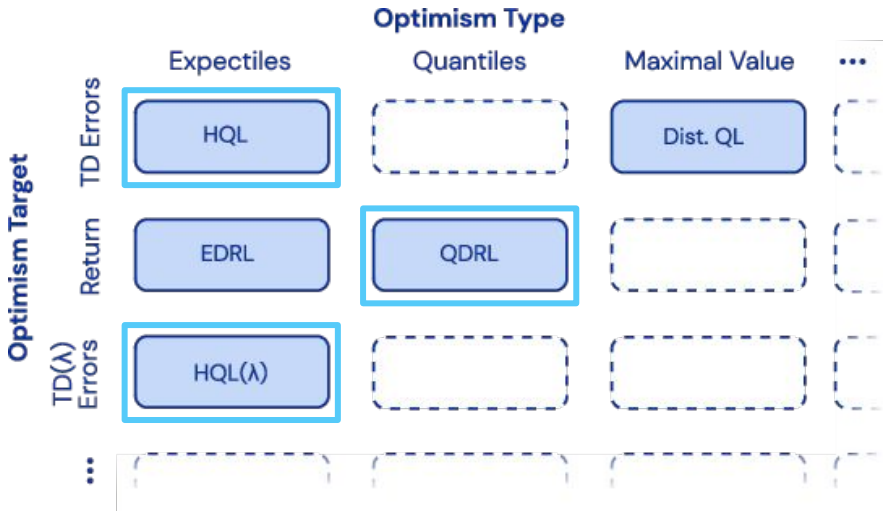
Setting: Cooperative multi-agent reinforcement learning, with independent value-based learners.

Optimistic variants of Q-learning (distributed, hysteretic, lenient, …) have long been used to induce cooperation in this settings.

Recently, variants of **distributional reinforcement learning** have also been used very effectively.

Contributions of this paper:

- Introduce a unifying framework that encompasses both families of approaches.
- Study similarities and differences of existing approaches.
- Identify environment properties that emphasise the difference between these different approaches.
- Identify as-yet unexplored algorithms within the framework.



The problem

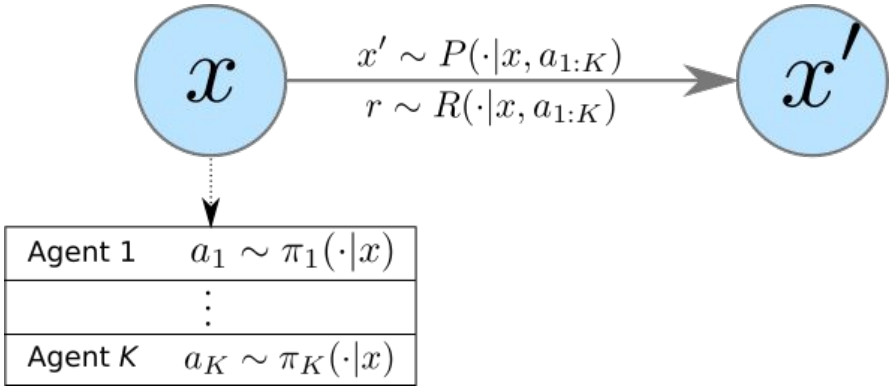
Multiagent Markov decision process

State space $\mathcal{X}$

Action space $\mathcal{A} = \prod_{i=1}^{K} \mathcal{A}_i$

Transition kernel $P : \mathcal{X} \times \prod_{i=1}^{K} \mathcal{A}_i \to \mathscr{P}(\mathcal{X})$

Rewards $R : \mathcal{X} \times \prod_{i=1}^{K} \mathcal{A}_i \to \mathbb{R}$



Aim: Compute optimal policies in a **decentralised** manner (agents do not see each other's actions).

Note: Cooperative case → no randomisation required → potentially straightforwardly use value-based learning.

However, environment is **non-stationarity** → cannot directly use single-agent RL algorithms.

A solution

Algorithm: Distributed Q-learning (Lauer & Riedmiller, 2000).

$$Q_i(x,a) \leftarrow \max(Q_i(x,a), r + \gamma \max_{a' \in \mathcal{A}_i} Q_i(x',a'))$$

Intuition: Forgive teammates' mistakes, and only pay attention to good outcomes → "maximally optimistic".

Convergence guarantees: Converges to optimal Q-values/policy if rewards and transitions are **deterministic**, with some additional algorithmic details.

Stochastic environments: Optimism about teammates' actions becomes conflated with misplaced optimism about transitions/rewards, and so algorithm does not generally converge. Also unstable with function approximation.

Optimism as a heuristic in stochastic environments

Can interpret distributed Q-learning as temporal difference learning with asymmetric learning rates:

$$Q_i(x,a) \leftarrow \max(Q_i(x,a), r + \gamma \max_{a' \in \mathcal{A}_i} Q_i(x',a'))$$

$$Q_i(x,a) \leftarrow Q_i(x,a) + \begin{cases} \alpha\delta & \text{if } \delta > 0 \\ \beta\delta & \text{otherwise} \end{cases} \qquad \delta = r + \gamma \max_{a' \in \mathcal{A}_i} Q_i(x',a') - Q_i(x,a)$$

$$\alpha = 1 \qquad \beta = 0$$

Approaches such as **Hysteretic Q-learning (HQL)** (Matignon et al., 2007) aim to address the problems raised by stochasticity by:

1. Reducing these rates to anneal the noise in the TD errors.
2. Making β non-zero, to soften the level of optimism applied.

HQL update rule:

$$Q_i(x,a) \leftarrow Q_i(x,a) + \begin{cases} \alpha\delta & \text{if } \delta > 0 \\ \beta\delta & \text{otherwise} \end{cases} \qquad \delta = r + \gamma \max_{a' \in \mathcal{A}_i} Q_i(x',a') - Q_i(x,a)$$

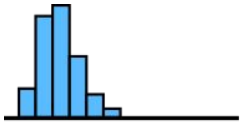
$$\alpha > \beta > 0$$

Distributional reinforcement learning

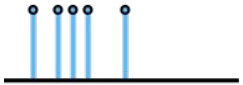
Learn **distribution** of returns, not just expected value.

Random return: $Z^\pi(x,a) = \sum_{t=0}^{\infty} \gamma^t R_t$

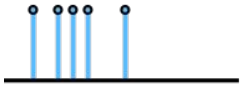
Categorical Distributional RL (Bellemare et al., 2017)



Quantile Distributional RL (Dabney et al., 2018)



Expectile Distributional RL (Rowland et al., 2019)



See also references in paper for further methods

An expectile of a distribution μ is parameterised by $\tau \in [0,1]$, and defined by

$$\arg\min_{e \in \mathbb{R}} \mathbb{E}_{X \sim \mu}[(X-e)^2[\tau \mathbb{1}_{X \geq e} + (1-\tau)\mathbb{1}_{X<e}]]$$

Effective in single-agent deep RL, acting greedily with respect to distribution means.

Recently applied in cooperative MARL, by acting greedily with respect to optimistic summary of distribution, based on e.g. quantiles (Lyu & Amato, 2020); see paper for more references.

A first result

Proposition
In stateless environments, **hysteretic Q-learning** and **expectile distributional RL** are identical algorithms.

Sketch proof

HQL update:

$$Q_i(a) \leftarrow Q_i(a) + \begin{cases} \alpha\delta & \text{if } \delta > 0 \\ \beta\delta & \text{otherwise} \end{cases}$$

In the stateless case:

$$\delta = r - Q_i(a)$$

So update corresponds to gradient descent on the following loss:

$$\alpha \mathbb{1}_{r>Q_i(a)}(r - Q_i(a))^2 + \beta \mathbb{1}_{r<Q_i(a)}(r - Q_i(a))^2$$

Which in expectation is proportional to the expectile loss with optimism parameter:

$$\tau = \frac{\alpha}{\alpha + \beta}$$

A unifying framework

Approaches like hysteretic Q-learning apply optimism:

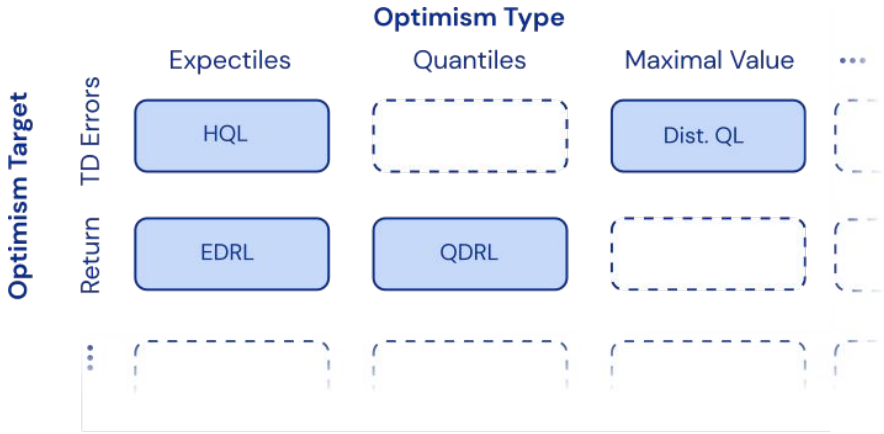
- At the level of each TD error.
- Using asymmetric learning rates, leading to expectile optimism.

Approaches like quantile distributional RL:

- Apply optimism at the level of the return.
- Using a quantile as an optimistic summary.

General framework:

- Select a **target** for optimism.
- Select a **type** of optimism.

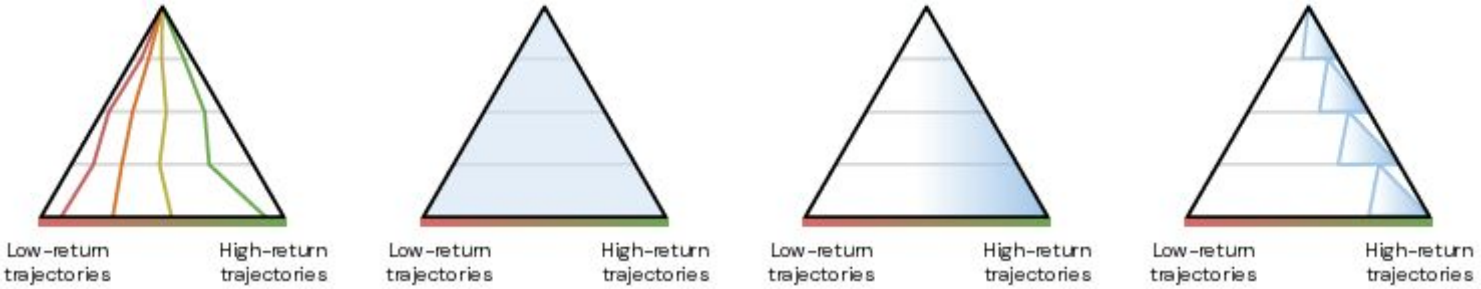


Similarities and differences

Proposition
In stateless environments, **hysteretic Q-learning** and **expectile distributional RL** are identical algorithms.

More generally: TD-optimistic and distributional RL approaches coincide in stateless environments.

What are the differences in stateful environments?



Raw trajectories **Risk-neutral** **Return-optimistic** **TD-optimistic**

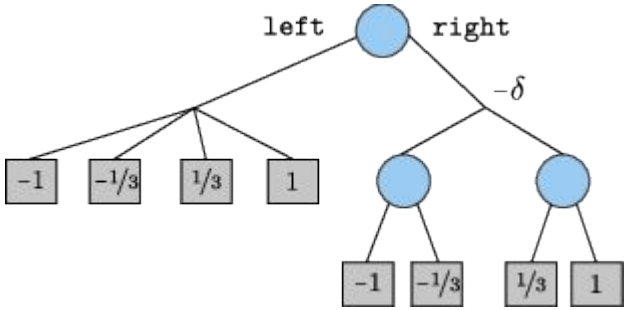
TD-optimistic approaches **compound optimism** at each state along a trajectory.

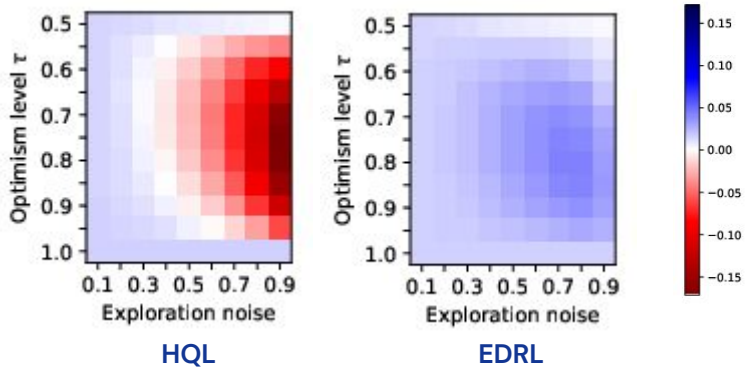
Understanding differences

What is the impact of this difference practically?

Aim to understand the effect of particular environment aspects on convergence.

Case I: Unequal trajectory lengths





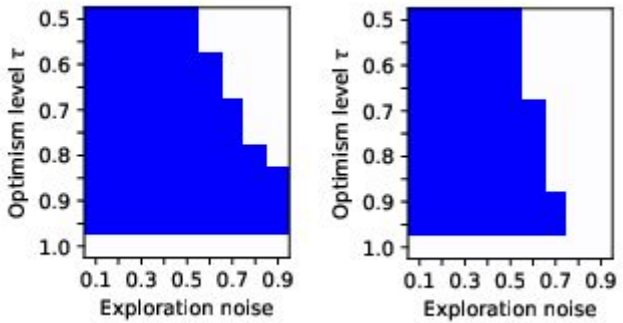
Understanding differences

What is the impact of this difference practically?

Aim to understand the effect of particular environment aspects on convergence.

Case II: Repeated state visits

Repeated partially-stochastic climbing game (Claus & Boutilier, 1998).



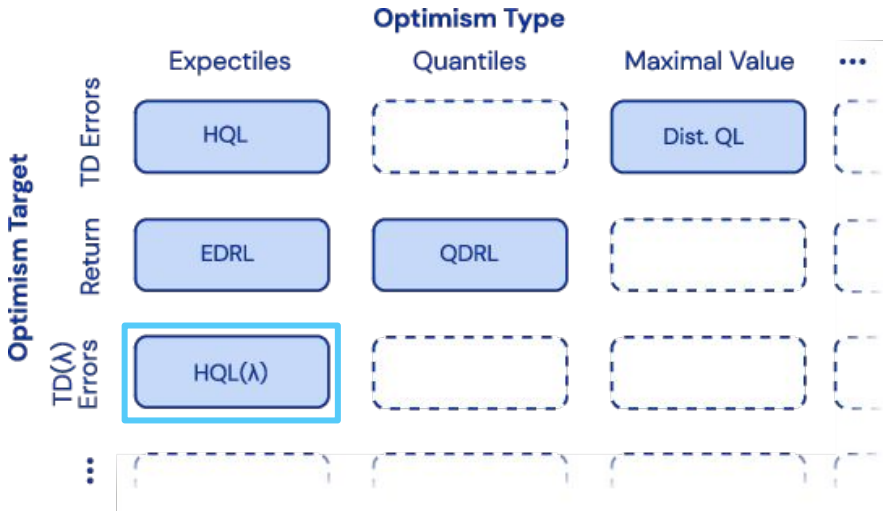
New approaches

New algorithms by selecting unexplored pair of **optimism target** and **optimism type**.

Example: HQL(λ) uses expectile optimism over λ-returns, interpolating between HQL and expectile distributional RL.

Many other approaches possible, not yet explored.

Other approaches, such as risk-neutral bootstrapping, as in (Achab, 2020), are also possible - see paper for full details.



Conclusion

Recent works use TD-optimistic approaches and distributional reinforcement learning for cooperative MARL at scale.

We introduce a unifying framework for thinking about TD-optimistic and return-optimistic algorithms, leading to:

- Theoretical comparisons between different approaches.
- Identification of unexplored algorithmic combinations.
- Fine-grained study of environment properties that lead to distinct performance between TD-optimism and return optimism.

Future work:

- Further investigation of unexplored algorithms.
- Further granular comparisons of different approaches experimentally.
- Extension to competitive, sequential-move environments.

Thank you