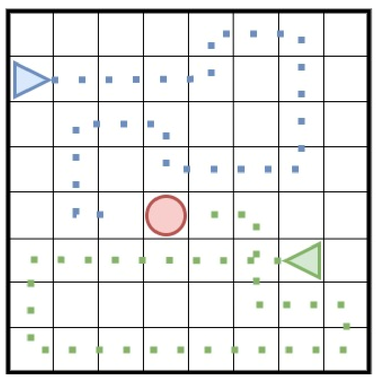
Ensemble Value Functions for Efficient Exploration in Multi-Agent Reinforcement Learning

Lukas Schäfer, Oliver Slumbers, Stephen McAleer, Yali Du, Stefano V. Albrecht, David Mguni

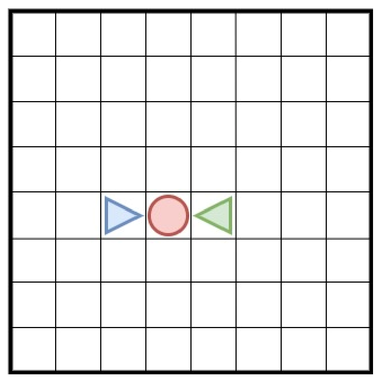
See paper for more details!

Problem Setting

Problem: Random exploration is very inefficient in discovering cooperation in multi-agent reinforcement learning (MARL)



Independent Exploration



Cooperative Exploration

Question: How to focus exploration on states that require coordination?

Summary and Contributions

Idea: Rewards in cooperative states vary depending on the actions of other agents → **Use variability of value estimates** to focus exploration on cooperative states and actions

Ensemble Value Functions for Multi-Agent Exploration (EMAX)
1. Disagreement of value estimates across the ensemble to guide exploration towards cooperative states
2. Average value estimates as robust target values

Plug-and-play extension for value-based MARL algorithms. Across 21 tasks, EMAX improves the final evaluation returns of IDQN, VDN, and QMIX by 53%, 36%, and 498%, respectively.

Ensemble Value Functions for Multi-Agent Exploration

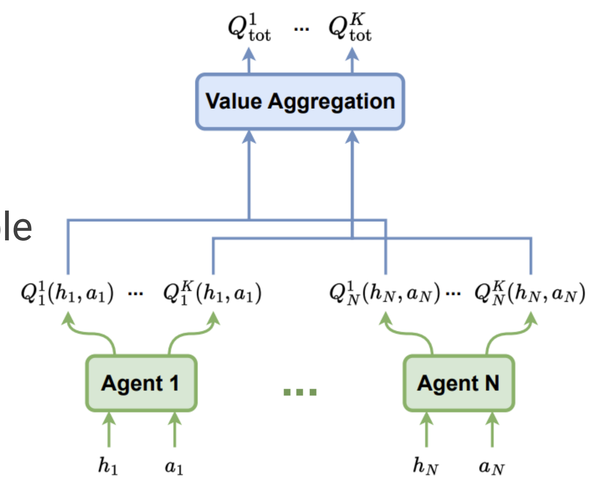
Agent i trains **ensemble of K value functions** $\{Q_i^k\}_{k=1}^K$

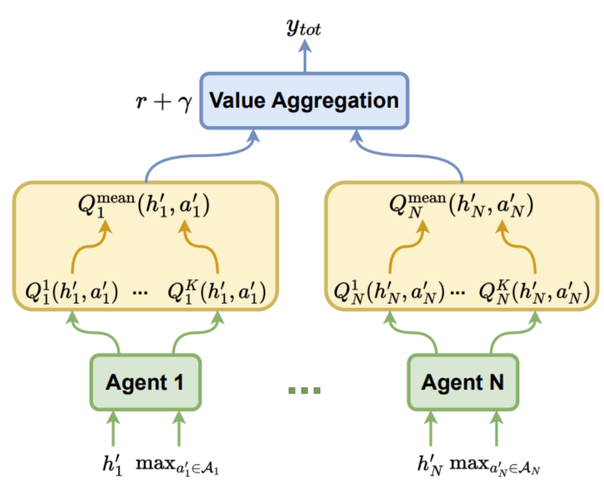
Exploration policy: $\pi_i^{expl}(h_i) \in \operatorname{argmax}_a Q_i^{mean}(h_i, a) + \beta Q_i^{std}(h_i, a)$

Evaluation policy: majority vote of greedy actions across the ensemble

Independent target computation: $r + \gamma \max_{a_i'} Q_i^{mean}(h_i', , a_i')$

Value decomposition: Aggregate kth value function of all agents to joint state-action value estimate Q_{tot}^k and target values with the aggregation of $Q_1^{mean}(h_1', a_1'), \dots, Q_N^{mean}(h_N', a_N')$





Value Estimates

$$Q_i^{mean}(h, a) = \frac{1}{K} \sum_{k=1}^{K} Q_i^k(h, a)$$

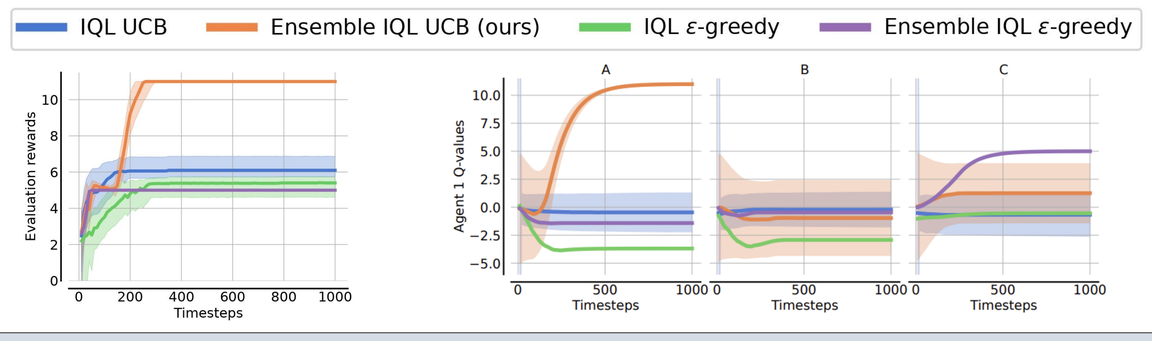
Target Estimates

$$Q_i^{std}(h, a) = \sqrt{\frac{\sum_{k=1}^{K} \left(Q_i^k(h, a) - Q_i^{mean}(h, a) \right)^2}{K}}$$

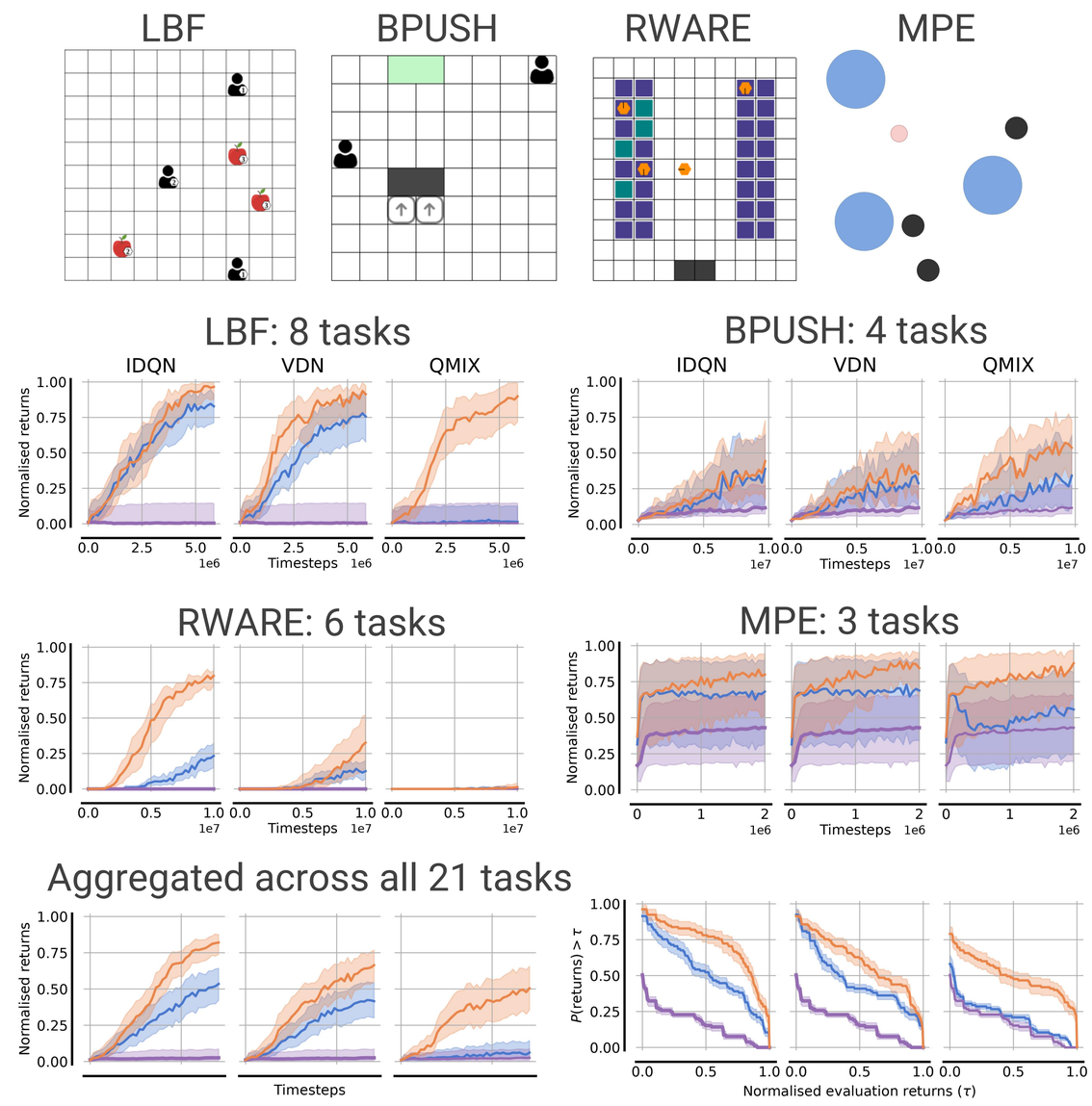
Example: Tabular Ensemble Exploration

Agent 1 \ Agent 2	A	B	C
A	11	−30	0
B	−30	7	6
C	0	0	5

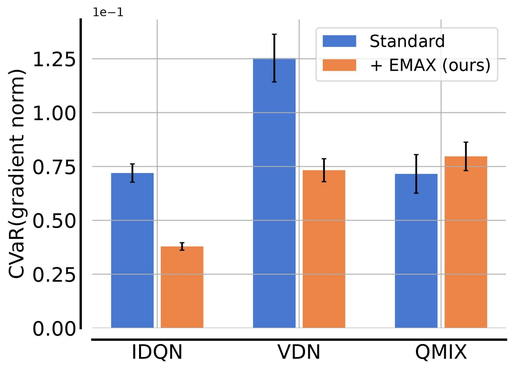
Climbing game: single-stage two-player common-reward matrix game



Evaluation



Analysis



Stability of gradients: $\nabla_t' = |\nabla_{t+1}| - |\nabla_t|$

$$CVaR(\nabla') = \mathbb{E}[\nabla' \mid \nabla' \geq VaR_{95\%}(\nabla')]$$

Training time for 10,000 training steps:

Algorithm	Baseline	$K = 2$	$K = 5$	$K = 8$
IDQN	16.80	21.29 (+27%)	33.04 (+97%)	48.06 (+186%)
VDN	16.92	21.56 (+27%)	33.25 (+97%)	48.16 (+185%)
QMIX	17.70	22.53 (+27%)	33.71 (+90%)	48.66 (+175%)

Contact: l.schaefer@ed.ac.uk & davidmguni@hotmail.com
Twitter: @LukasSchaefer96