Efficient Bayesian Ultra-Q Learning for Multi-Agent Games

Ward Gauderis, Fabian Denoodt, Bram Silue, Pierre Vanvolsem, Andries Rosseau
AI Lab - Vrije Universiteit Brussel

Bayesian Ultra-Q Learning is an extension of the Hyper-Q Learning algorithm that is more efficient in discovering mixed Nash equilibria in adaptive multi-agent games.

Hyper-Q Learning (Tesauro, 2003)

Partial observability is one of the main challenges for independent Q-learning agents in multi-agent settings. Hyper-Q learning agents estimate their opponent's mixed strategy y (using EMA or Bayesian estimation) to learn the Q-value of joint mixed strategies (x, y).

Update equation

$$Q(y,x) \leftarrow Q(y,x) + \Delta Q(y,x),$$

$$\Delta Q(y,x) = \alpha(t)\left[r + \gamma \max_{x'} Q(y',x') - Q(y,x)\right]$$

Greedy policy

$$\hat{x}(s,y) = \arg\max_{x} Q(s,y,x)$$

Bayesian Ultra-Q Learning (ours)

While Hyper-Q agents merely update the Q-table corresponding to a single state (x, y), Bayesian Ultra-Q leverages the information that **similar states most likely result in similar rewards**. Therefore, the Q-value of **every state** (x, y) is updated, weighted by the probability of opponent strategy y given the history of observed actions H and the similarity measure between x and the true agent strategy.

Update equation

Cosine similarity *Bayesian belief probability*

$$\Delta Q(y,x) = \alpha \langle x, x_{real} \rangle \; P(y|H) \Big[r + \gamma \max_{x'} \sum_{y'} P(y'|H') Q(y',x') - Q(y,x) \Big]$$

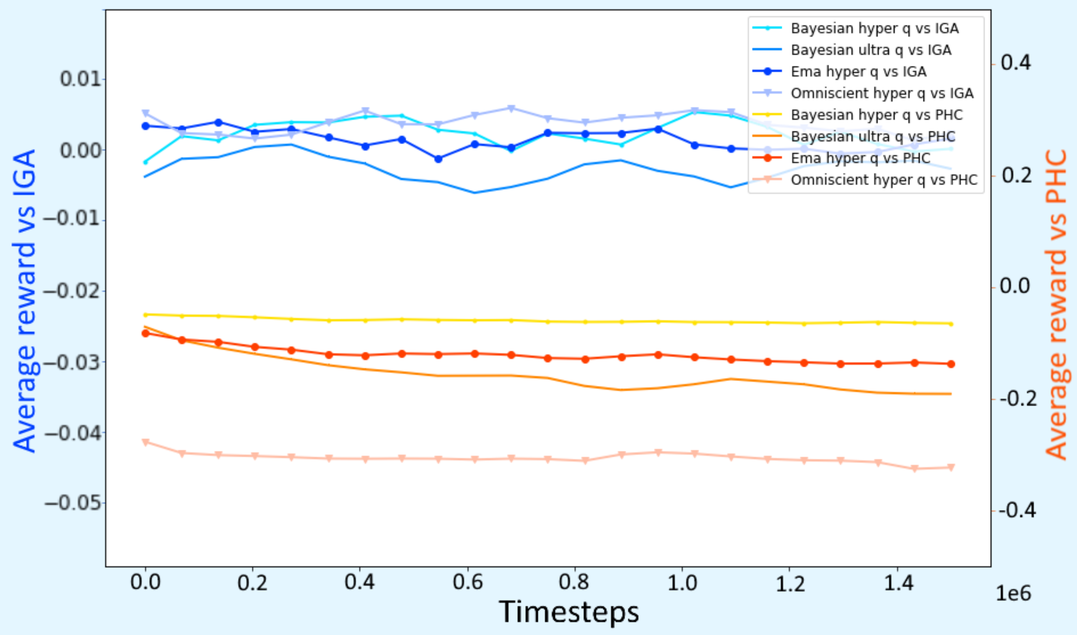
Greedy policy

$$\hat{x} = \arg\max_{x} \sum_{y} P(y|H) Q(y,x)$$

Alternatively, this similarity measure could be replaced with the probability $P(a|x)$, the likelihood of the action a being played when the mixed strategy x would be used.

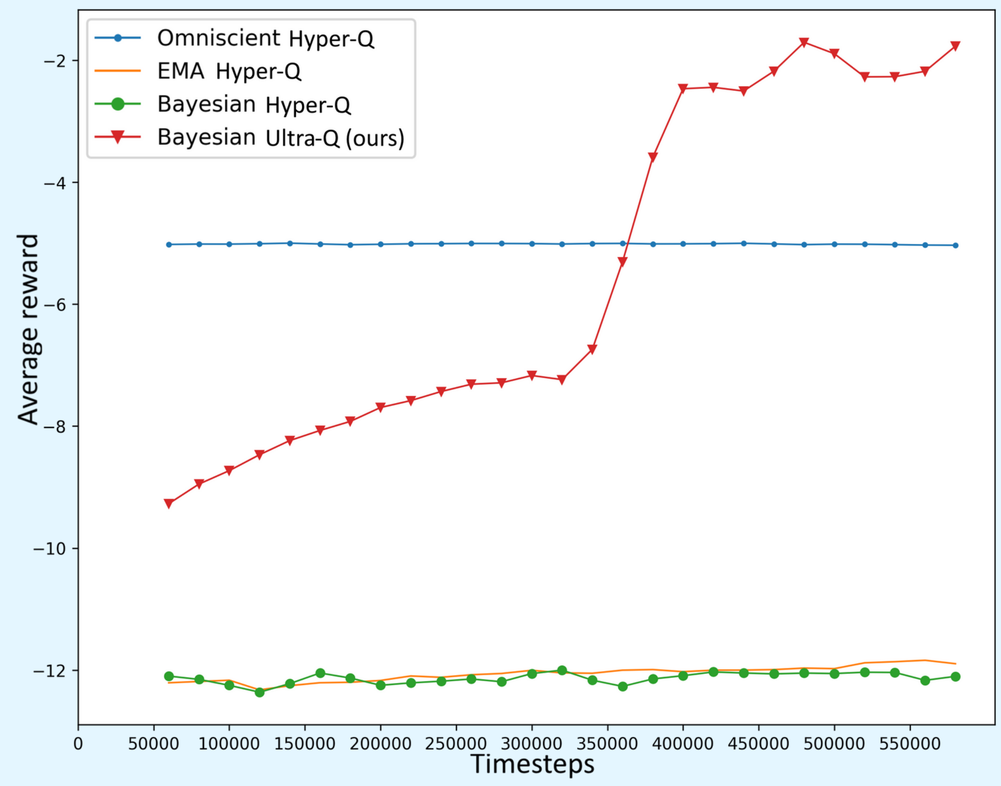
Rock-Paper-Scissors Game

We compared Bayesian Ultra-Q and Tesauro's Hyper-Q agents against dynamic opponents (IGA and PHC). When both agents adopt the Nash equilibrium strategy, evenly distributing their choices among rock, paper, and scissors, the expected reward for both agents becomes zero. They cannot improve their performance beyond this point.



Hill-climbing Game

Two agents work together to maximise shared rewards by collaborating. Successful cooperation requires accurate opponent strategy estimation. If agents fail to cooperate, they face severe punishment, resulting in low rewards for both. This game is susceptible to the issue of **relative overgeneralisation**, which occurs when agents learn to repeat a suboptimal action because other agents took the wrong action.



We compare our results with Tesauro's three Hyper-Q agents, which differ in their strategies for estimating opponents' actions:

- **Bayesian Ultra-Q surpasses Tesauro's agents** significantly in the hill climbing game, while its performance in rock-paper-scissors falls behind.
- **Bayesian Ultra-Q converges to the Nash equilibrium** in the hill climbing game, highlighting its effectiveness in achieving optimal outcomes while **addressing the issue of relative overgeneralisation**.
- We reproduce Tesauro's work behind Hyper-Q Learning and shed light on certain aspects that require further clarification, such as the definition of y' in the Bayesian update equation. The code is publicly available through GitHub.

ARTIFICIAL INTELLIGENCE RESEARCH GROUP