Distributional Monte Carlo Tree Search for Risk-Aware and Multi-Objective Reinforcement Learning

Conor F. Hayes [1] Mathieu Reymond [2] Diederik M. Roijers [2, 3] Enda Howley [1] Patrick Mannion [1]

[1]National University of Ireland Galway, Ireland. [2]Vrije Universiteit Brussel, Belgium. [3]HU Univeristy of Applied Science Utrecht, Netherlands.

Introduction

- In many risk-aware and multi-objective reinforcement learning (MORL) settings the utility of a user is derived from the single execution of a policy.
- In such settings the expected return, or value, does not provide sufficient critical information about the potential positive or adverse effects a decision may have.
- In this case, it is essential to replace the expected value with a posterior distribution over the expected utility of the returns (ESR).
- We propose a novel algorithm, Distributional Monte Carlo Tree Search (DMCTS), which learns a posterior distribution over the expected utility of the returns.
- We implement and demonstrate DMCTS for both risk-aware and multi-objective problems under the ESR criterion.

A full version of the paper can be found at the following link: https://arxiv.org/abs/2102.00966.

Distributional Monte Carlo Tree Search

- To compute the distribution we first calculate the accrued returns, $\mathbf{R}_t^-$. The accrued returns is the sum of rewards received during the execution phase as far as timestep, t, where $\mathbf{r}_t$ is the reward received at each timestep,

$$\mathbf{R}_t^- = \sum_{0}^{t-1} \mathbf{r}_t.$$

- Secondly, we must calculate future returns, $\mathbf{R}_t^+$. The future returns is the sum of the rewards received when traversing the search tree during the learning phase and Monte Carlo simulations from timestep, t, to a terminal node, t_n,

$$\mathbf{R}_t^+ = \sum_{t}^{t_n} \mathbf{r}_t.$$

- The cumulative returns, $\mathbf{R}_t$, is the sum of the accrued returns, $\mathbf{R}_t^-$ and the future returns, $\mathbf{R}_t^+$.

Algorithm 1: Update Bootstrap Distribution

```
Input: i ← Node in the tree
Input: R_t ← Cumulative Returns
J ← node.bootstrapDistribution
for j, ..., J bootstrap replicates do
    Sample d_j from Bernoulli(1/2)
    if d_j = 1 then
        α_ij = α_ij + u(R_t)
        β_ij = β_ij + 1
    end
end
```

- We use a bootstrap distribution to approximate the posterior [2]. To update the bootstrap distribution at each node we use Algorithm 1.
- The agent then executes the action, a^*, which corresponds to the following:

$$a^* = \arg\max_i \frac{\alpha_{ij}}{\beta_{ij}}.$$

Experiments

- We evaluate DMCTS in a risk-aware problem domain [4] under ESR using the following non-linear utility function:

$$u = 1 - e^{-r_t}. \tag{1}$$

- To evaluate DMCTS in the risk-aware domain, we compare DMCTS against Q-learning [5].
- To evaluate DMCTS in a multi-objective setting under ESR, we use the Fishwood problem [3] with the following non-linear utility function:

$$u = \min\left(\texttt{fish}, \left\lfloor \frac{\texttt{wood}}{2} \right\rfloor\right). \tag{2}$$

- To evaluate DMCTS in the Fishwood domain, we compare DMCTS against C51 [1], EUPG [3], and Q-learning [5].
- As shown in Figure 1 and Figure 2, DMCTS learns good policies in risk-aware settings and achieves state-of-the-art performance in MORL under ESR.

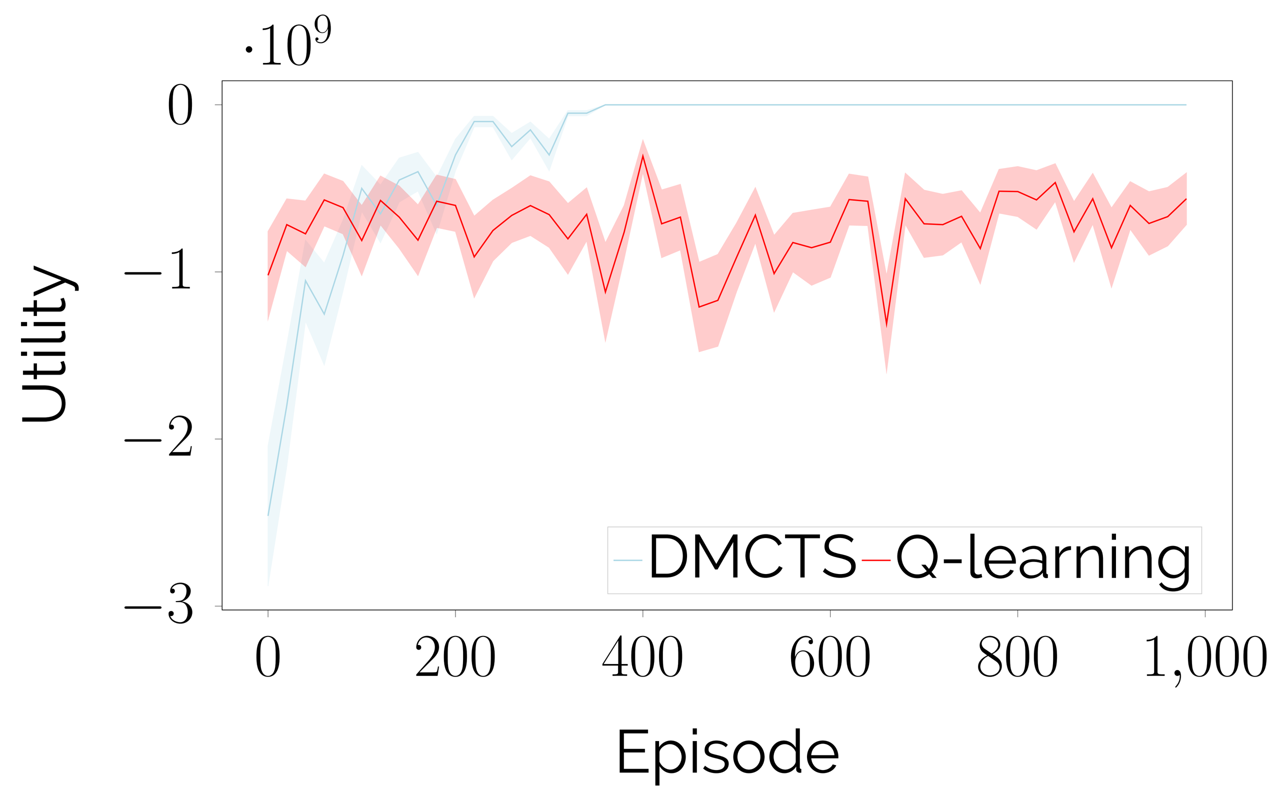

Figure 1:Results from the risk-aware environment.

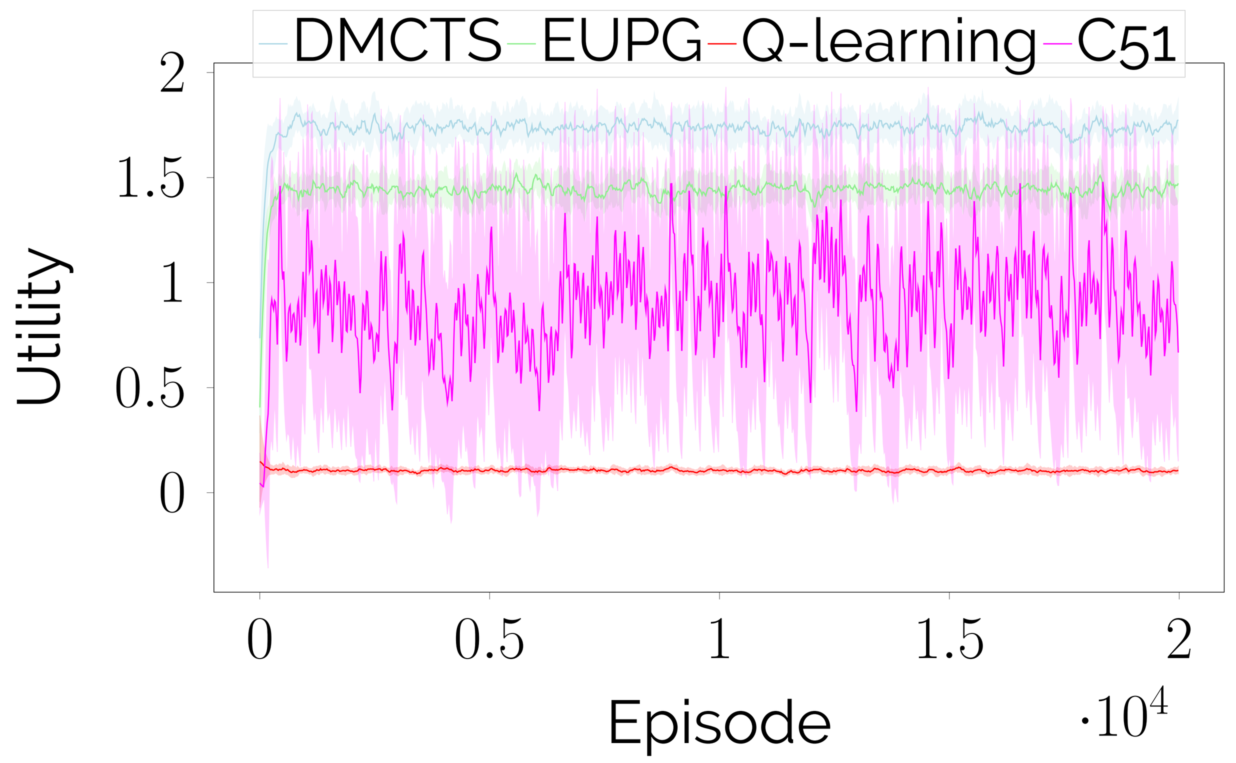

Figure 2:Results from the fishwood environment.

References

[1] Marc G. Bellemare, Will Dabney, and Rémi Munos.
A distributional perspective on reinforcement learning.
In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 449–458, Sydney, NSW, Australia, 2017. JMLR.org.

[2] Dean Eckles and Maurits Kaptein.
Thompson sampling with the online bootstrap.
CoRR, abs/1410.4009, 2014.

[3] Diederik M Roijers, Denis Steckelmacher, and Ann Nowé.
Multi-objective reinforcement learning for the expected utility of the return.
In *Proceedings of the Adaptive and Learning Agents workshop at FAIM 2018*, 2018.

[4] Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer.
Risk-sensitive reinforcement learning.
Neural Computation, 26(7):1298--1328, 2014.

[5] Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé.
Scalarized multi-objective reinforcement learning: Novel design techniques.
In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 191--199. IEEE, 2013.