Imitation learning from neural networks with continuous action spaces using regression trees

Author: **Tymon Cichocki (t.s.cichocki@student.tudelft.nl)**
Responsible professor: **Anna Lukina**
Supervisor: **Daniël Vos**

1. Introduction

- Deep Neural Network models (DNN) are commonly used in the Reinforcement Learning tasks
- They raise safety concerns because an expert cannot verify a decision-making process
- Small decision trees can be used to mimic NN behaviour

2. Research Objective

The aim of this research was to study whether decision tree policies trained using the DAGGER algorithm yield comparable results to the NN policies. The following Research questions were answered:

1. *How does increasing the number of DAGGER iterations impact the results?*
2. *How does limiting the decision tree nodes affect performance?*
3. *How does observation space dimensionality impact the trained policy results?*

3. Methodology

DAGGER algorithm

- Trains a policy imitating an expert
- At each iteration adds new state action pairs to the dataset
- States that are visited depend on the latest policy, and an expert says what actions should be taken

$D \leftarrow \varnothing$
$N \leftarrow$ number of DAGGER iterations
$T \leftarrow$ number of trajectory rollouts
for i in N **do**
  $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$
  for j in T **do**
    $T_i^j \leftarrow$ sample trajectory with π_i
    For each state s in T_i^j add $(s, \pi^*(s))$ to D_i
  end for
  $D \leftarrow D \cup D_i$
  Learn $\hat{\pi}_{i+1}$ on D
end for
return best $\hat{\pi}_i$ on validation

Figure 1: DAGGER algorithm – used to train decision tree policies

4. Results

- **Early termination** environments are those where simulation stops once the status becomes *unhealthy* (e.g., robot falls on the ground)
- There exist **critical states** in those environments – states where making a mistake leads to an early termination
- Learning curve in the early termination environments shows significant performance drops, which might be caused by the underrepresentation of the critical states in the training dataset (Figure 2).

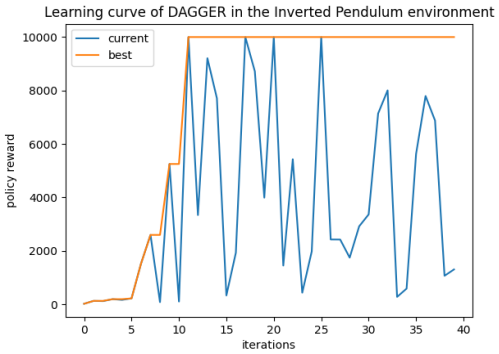

Figure 2: Learning curve of DAGGER in the environment with early termination

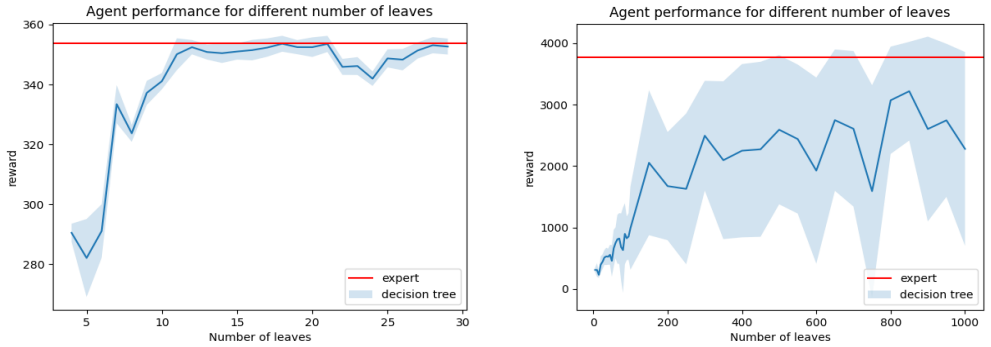

Figure 3: Performance for the trees with different sizes. Red line is the expert performance. Right environment is more complex than the Left

- Size of a decision tree does have an impact on the performance
- For the environments with many observations, it is not possible to replicate a NN policy

(Figure 3)

5. Conclusion

- Further research is needed to improve the performance in the early termination environments
- It is possible to use DAGGER to produce interpretable policies for the less complex environments

TU Delft