Lecture 15 Rewards in RL

Emma Brunskill

CS234 Reinforcement Learning.

Spring 2024

Refresh Your Understanding

Select all that are true:

- Direct Preference Optimization assumes human preferences follow a Bradley Terry model
- RLHF can be used with reward models learned from preferences or reward models learned from people labeling rewards
- Asking people to provide preference pair rankings is likely to be an efficient way to learn the reward model for board games
- DPO and RLHF can be used with extremely large policy networks
- Not sure

Refresh Your Understanding

Select all that are true:

- Direct Preference Optimization assumes human preferences follow a Bradley Terry model
- RLHF can be used with reward models learned from preferences or reward models learned from people labeling rewards
- Asking people to provide preference pair rankings is likely to be an efficient way to learn the reward model for board games
- DPO and RLHF can be used with extremely large policy networks
- Not sure

Class Structure

- Last time: MCTS
- Today: Rewards in RL
- Next time: Quiz

Quiz Information

- Multiple-choice
- Covers all material up to next Wednesday
- Allowed 1 two-sided page of notes
- We will release past sample quizzes. Note the material this year is slightly different (ex. RLHF and DPO) so the past quizzes will not be a perfect representation. However we still think they will help illustrate the type of questions and much of the material does overlap.

Refresh Your Understanding 2

Select all that are true:

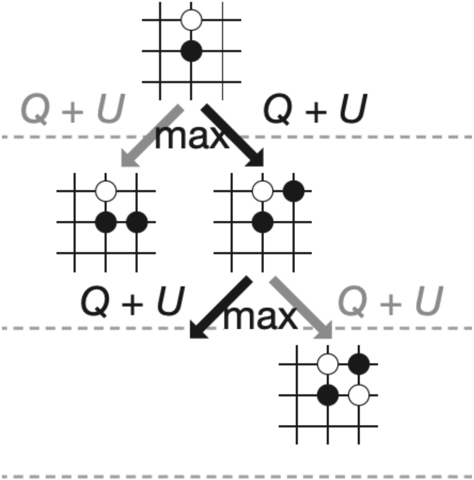
- Monte Carlo Tree Search approximates a forward search tree
- MCTS tackles the action branching function through sampling
- AlphaZero uses two networks, one to help prioritize across actions, and one to provide an estimate of the value at leaves
- Doing additional guided Monte Carlo tree search when computing an action significantly improved the test time performance of AlphaZero
- Self play provides a form of curriculum learning
- Not sure

Refresh Your Understanding 2

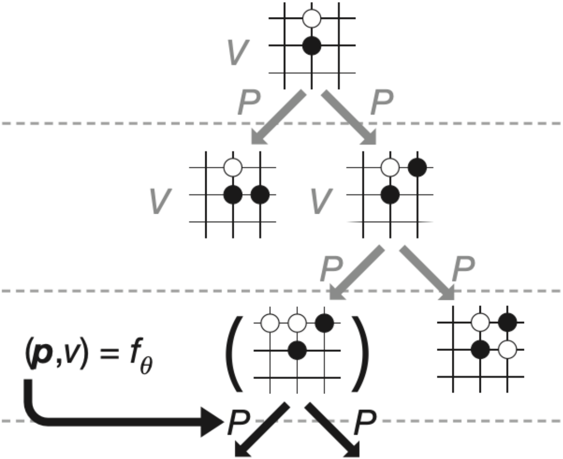
Select all that are true:

- Monte Carlo Tree Search approximates a forward search tree
- MCTS tackles the action branching function through sampling
- AlphaZero uses two networks, one to help prioritize across actions, and one to provide an estimate of the value at leaves
- Doing additional guided Monte Carlo tree search when computing an action significantly improved the test time performance of AlphaZero
- Self play provides a form of curriculum learning
- Not sure

Selecting a Move in a Single Game: Repeatedly Expand[1]



Selecting a Move in a Single Game: Note Using Network Predictions for Action Probabilities[2]


V
P
P
V
V
P
P
$(\mathbf{p}, v) = f_\theta$
P
P