Policy Comparison Metric

Policy comparison for rounds of a given game played between humans, and agents learning over given examples:

$S =$ set of total states in a game

$S_H =$ set of states seen by human agents

$E =$ set of states present/ seen in both agent and human games

$\eta_S = \frac{|E|}{|S_H|} =$ state similarity metric

$B_{s,a} =$ number of times action a is taken from state $s \in E$ by the agents

$B_s =$ most frequent action from s by agents $= \arg\max_a B_{s,a}$

$H_{s,a} =$ number of times action a is taken from state $s \in E$ by human players

$H_s =$ most frequent action from s by humans $= \arg\max_a H_{s,a}$

$\eta_P = \frac{\sum_{s \in E} \mathbb{1}(B(s) == H(s))}{|E|} =$ number of times the actions with highest frequencies from a state match between humans and agents

1 Policy comparison metric for human policies and robot policies

The policies learned by human and the norm learning agent are similar if they take the same actions from similar states. We use the same number of example trajectories from human and the norm learning agent. We count the states and actions encountered in these trajectories. As the policy comparison metric we count the action taken with maximum frequency for each common state encountered in these trajectories. A common state is a state

encountered in both human and agent trajectories. If the policies are similar the number of actions taken with highest frequency from each common state would be the same, and the ratio of similar actions to common states would be high. Further if the policies learned by the agents were similar they would reach common states, hence the ratio of common states to total states reached would also be high.

2 Policy comparison metric for agent - agent policies

A policy π can be thought as a function that returns an action probability distribution given a state. If two policies are similar then the Kullback-Leibler (KL) [1] divergence of the action probabilities for each of the states encountered between the learned policy and the true policy would be low. We calculate mean KL divergence for all states, between the learned policy and the true policy action distribution.

References

[1] Kullback, Solomon and Leibler, Richard A *On information and sufficiency*, The annals of mathematical statistics, 1951.