Belief Reward Shaping in Reinforcement Learning: Supplementary Material

Ofir Marom,[1] Benjamin Rosman [1 , 2]
[1]University of the Witwatersrand, Johannesburg, South Africa
[2]Council for Scientific and Industrial Research, Pretoria, South Africa

A Backgammon

This section describes the full details for the belief clusters and parameters settings used for this experiment in the main paper.

Notation	Description
dc	degree of contact
p1_bl	player 1 number of blots
p2_bl	player 2 number of blots
p1_ipt	player 1 number of inner points
p2_ipt	player 2 number of inner points
p1_pt	player 1 number of points
p2_pt	player 2 number of points
p1_pr	player 1 maximum prime
p2_pr	player 2 maximum prime
p1_iopp	player 2 checkers in player 1 inner board / on bar
p2_iopp	player 1 checkers in player 2 inner board / on bar

Table 1: Backgammon: description of notation used in tables 2 and 3.

Belief Cluster Criterion	μ_0	λ
dc>0, p1_bl>0	-0.5	3000
dc>0 ,p2_bl>0	0.5	3000
dc>0, p1_ipt>1	0.5	3000
dc>0, p2_ipt>1	-0.5	3000
dc>0, p1_pt>4	0.5	3000
dc>0, p2_pt>4	-0.5	3000
dc>0, p1_pr>1	0.5	3000
dc>0, p2_pr>1	-0.5	3000

Table 2: Backgammon: simple set of prior beliefs.

For the simple set of prior beliefs in table 2 the criterions are chosen so that the positions are "improvements" over the starting position as when we start a game each side has 0 blots, 1 inner point, 4 points and a maximum prime of 1. The complex set of priors are more advanced because they scale the reward based on the position (i.e. having 5 blot is considered worse than 2 blots) and restrict certain rewards to situations when the position is beneficial (i.e. there is an opponent checker in the player's inner board or on the bar). For the complex set of priors each i in table 3 is a separate belief cluster with its own prior mean and prior mean pseudo-count.

Belief Cluster Criterion	μ_0	λ	Range
dc>0, p1_bl=i	-1.5 $\frac{Min(i,5)}{15}$	3000	$i \in [0, 15]$
dc>0, p2_bl=i	1.5 $\frac{Min(i,5)}{15}$	3000	$i \in [0, 15]$
dc>0, p1_iopp>0, p1_ipt=i	0.5 $\frac{i}{6}$	3000	$i \in [0, 6]$
dc>0, p2_iopp>0, p2_ipt=i	-0.5 $\frac{i}{6}$	3000	$i \in [0, 6]$
dc>0, p1_pt=i	0.5 $\frac{i}{7}$	3000	$i \in [0, 7]$
dc>0, p2_pt=i	-0.5 $\frac{i}{7}$	3000	$i \in [0, 7]$
dc>0, p1_iopp>0, p1_pr=i	0.5 $\frac{i}{7}$	3000	$i \in [0, 7]$
dc>0, p2_iopp>0, p2_pr=i	-0.5 $\frac{i}{7}$	3000	$i \in [0, 7]$

Table 3: Backgammon: complex set of prior beliefs.

Furthermore, note that that the neural network has 4 output nodes to account for gammon / backgammon wins. It should be understood from the table above that $\mu_0 = c$ is actually a vector $[c, c, -c, -c]$ where the first two elements represents wins for player 1 and the last two elements represent wins for player 2.