# Empirical convergence analysis

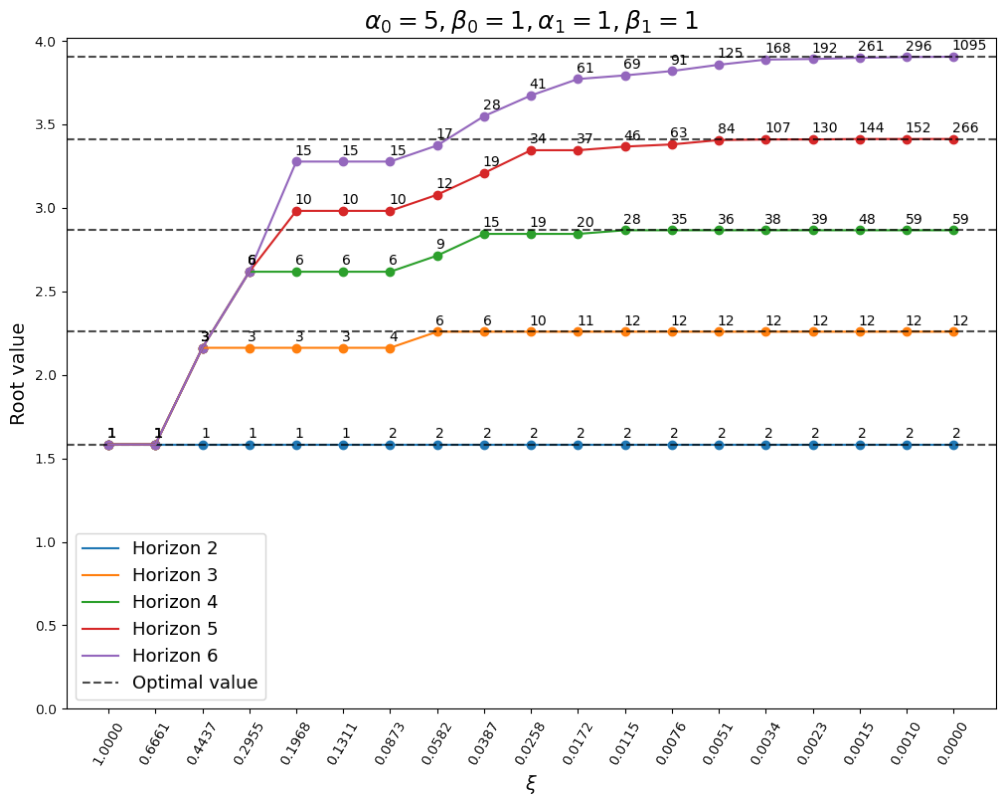
I am still puzzled by the choice of policy *inside* the tree. The options are, for instance, to set all $Q$-values for the intermediate nodes within the tree to 0 – in which case the policy would be uniform; or, alternatively, those $Q$-values could be set to the immediate expected reward according to those intermediate belief states. The former choice will force the replays to be initiated at the leaves of the tree (i.e., at the final horizon reached), since only the leaf values would have non-zero values to propagate. As for the other option, the replay patterns would be very different because of the Need term which discounts states that are further away.

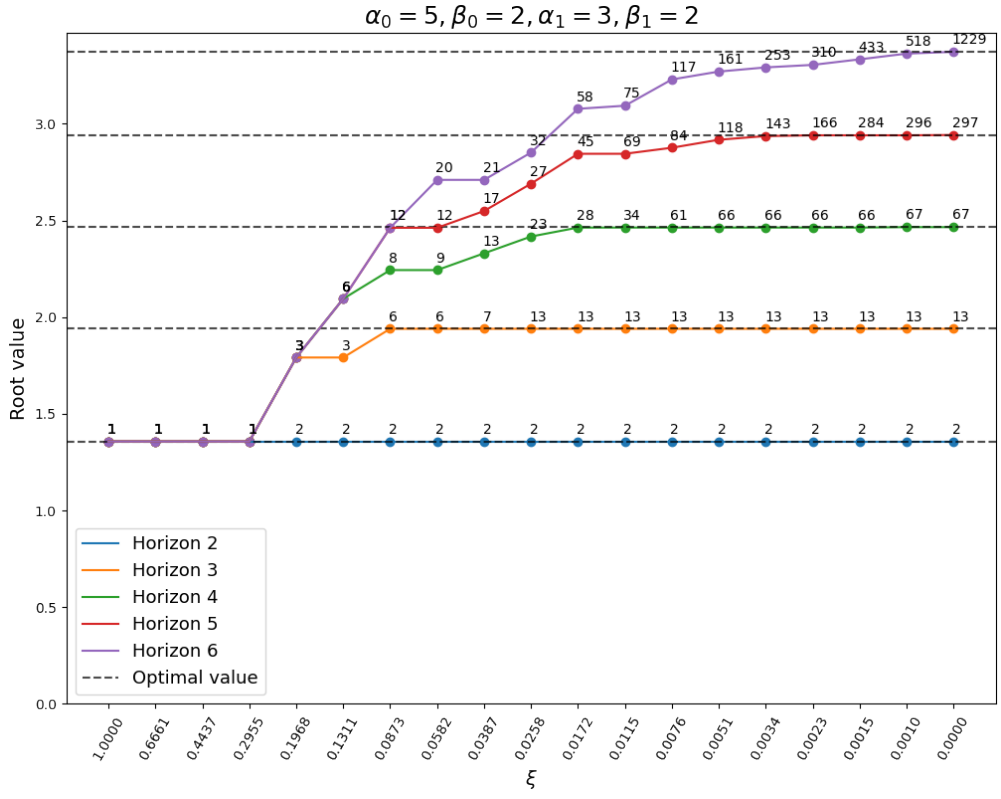
Moreover, I think that the uniform policy would make it less likely that distal beliefs for the sub-optimal action (at the root) get replayed. Indeed, the choice of the tree policy seems to affect the rate of convergence. I generated convergence plots for the two policies for comparison.

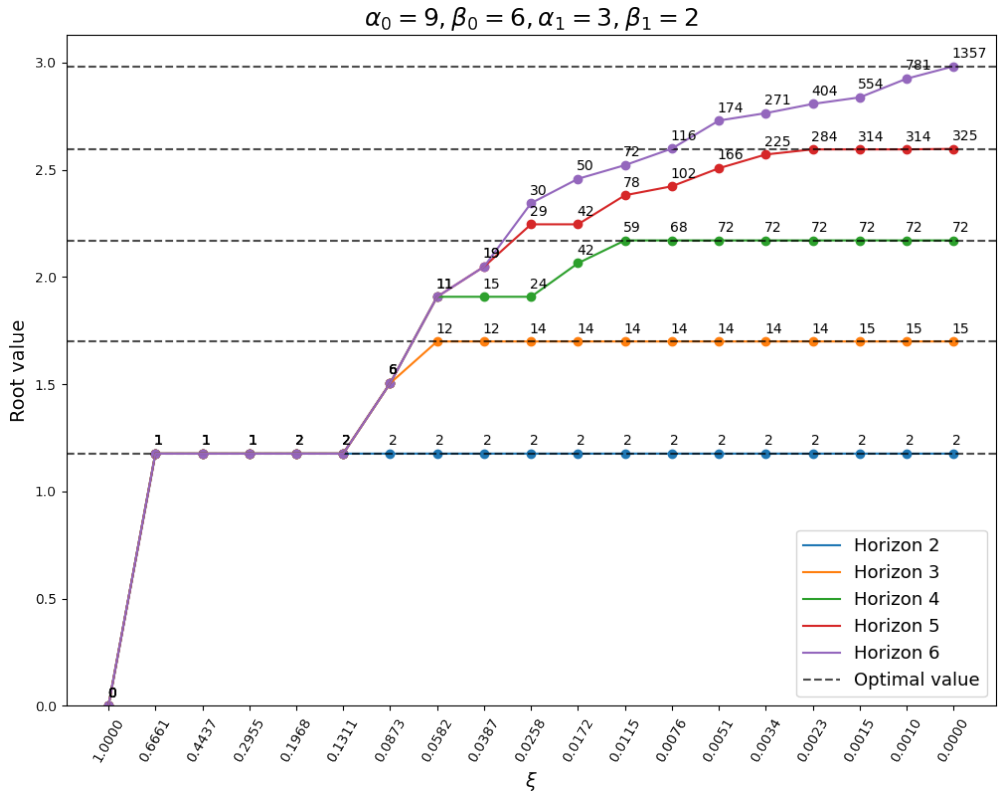
## Replay threshold

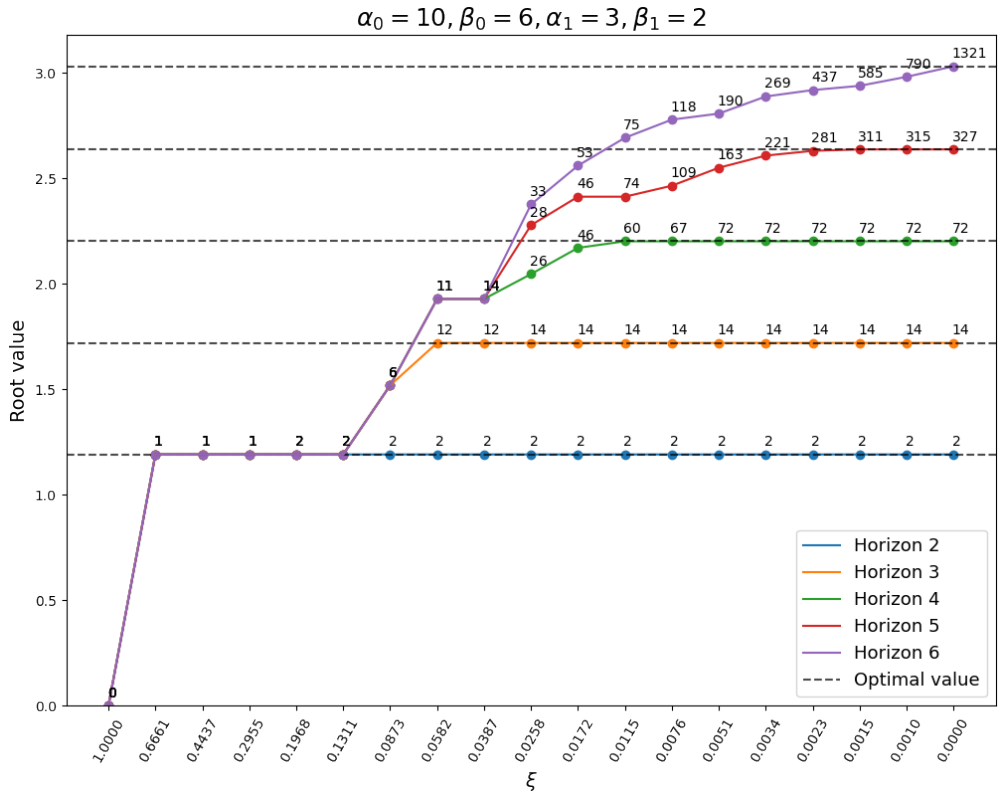
The plots below show how the value of the root state changes with the EVB threshold $\xi$ at multiple horizons and for various prior beliefs. Note that the root $Q$-values (i.e, the current MF $Q$-values) were set to 0 everywhere unless specified otherwise.

Numbers on top of each data point indicate the number of replays. Titles specify the prior beliefs at the root.


$\alpha_0 = 5, \beta_0 = 1, \alpha_1 = 1, \beta_1 = 1$
Root value
$\xi$
1.0000
0.6661
0.4437
0.2955
0.1968
0.1311
0.0873
0.0582
0.0387
0.0258
0.0172
0.0115
0.0076
0.0051
0.0034
0.0023
0.0015
0.0010
0.0000
Horizon 2
Horizon 3
Horizon 4
Horizon 5
Horizon 6
Optimal value


$\alpha_0 = 5, \beta_0 = 2, \alpha_1 = 3, \beta_1 = 2$
Root value
$\xi$
1.0000
0.6661
0.4437
0.2955
0.1968
0.1311
0.0873
0.0582
0.0387
0.0258
0.0172
0.0115
0.0076
0.0051
0.0034
0.0023
0.0015
0.0010
0.0000
Horizon 2
Horizon 3
Horizon 4
Horizon 5
Horizon 6
Optimal value




## Number of replays