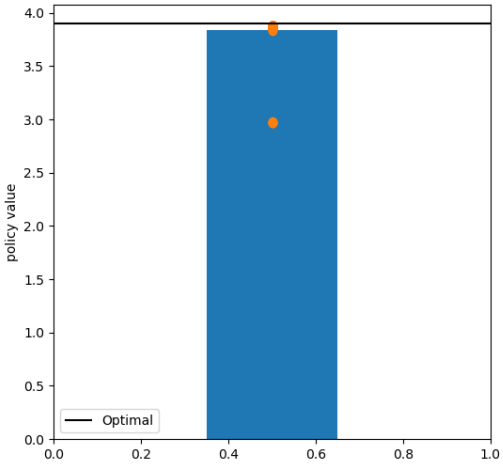
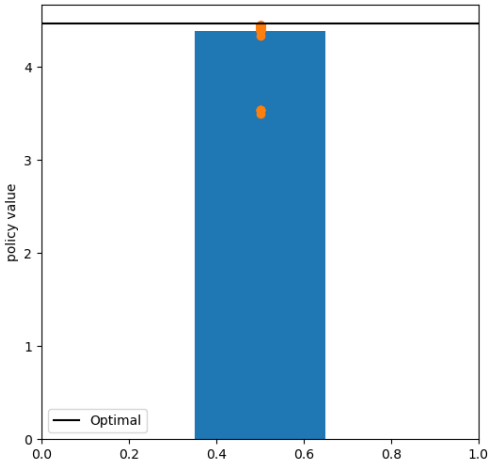
I simulated 200 trees with each tree's initial Q^{init}-values being initialised as a random (independent for each belief state) fraction of the true Q^{DP}-values (computed by a full DP solution): that is, each $Q^{init}(b,\cdot)$ was initialised as $uQ^{DP}(b,\cdot)$, where $u \sim U[0,1]$. The plot below shows the average policy value as a result of replay in each of these trees.

Horizon 4



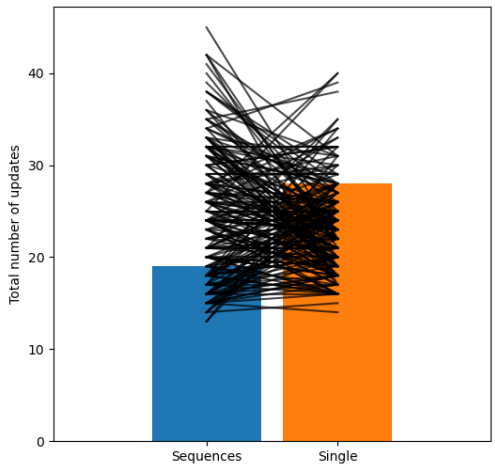
Horizon 5



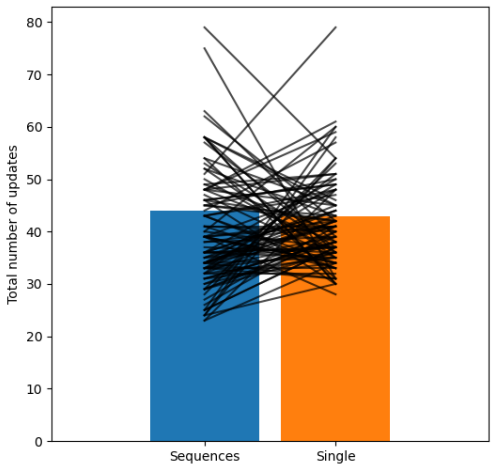
So, on average, the replay yields a near-optimal policy (which depends on the Q-value initialisation, as well as the softmax and ξ parameters which were fixed in these simulations).

The next plot shows the average number of executed replays, both with sequences (total number of state-action updates, so accounting for sequence lenghts), and without sequences. The difference doesn't seem to be significant, but it will be interesting to see the 'time' equivalent – i.e., what we discussed before with inter- vs inta-replay intervals.

Horizon 4

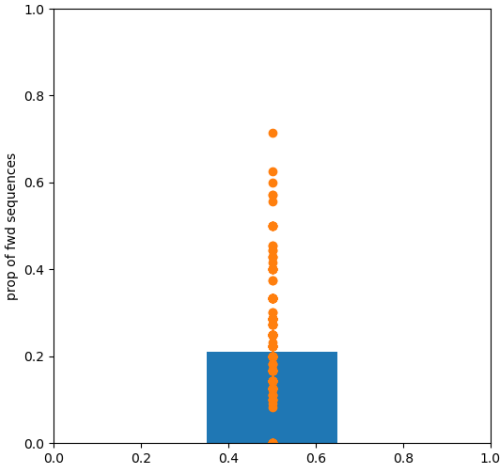


Horizon 5

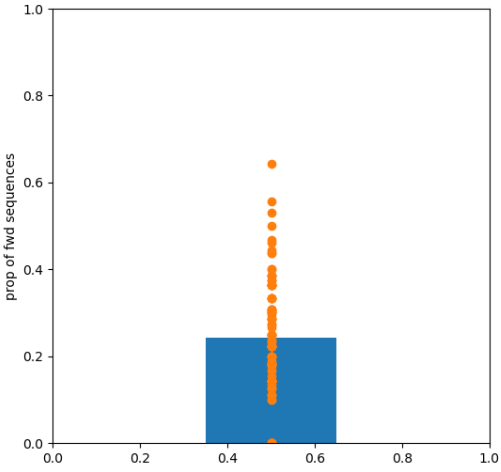


The next plot shows the average relative proportion of forward vs reverse sequences (perhaps there's a better way of showing this since here I'm treating all $>$ 2-step sequences equally).

Horizon 4



Horizon 5



And finally I also looked at whether there is any decipherable systematic relationship between the initial Q^{init}-values and the proportion of forward and reverse sequences – this is shown in the next figure. It shows the difference between the average Q^{init}-values of the trees with the proportion of forward replays above the mean and the average Q^{init}-values of the trees below the mean (based on the bar plot above). For clarity, I'm showing this in a tree with horizon 3.

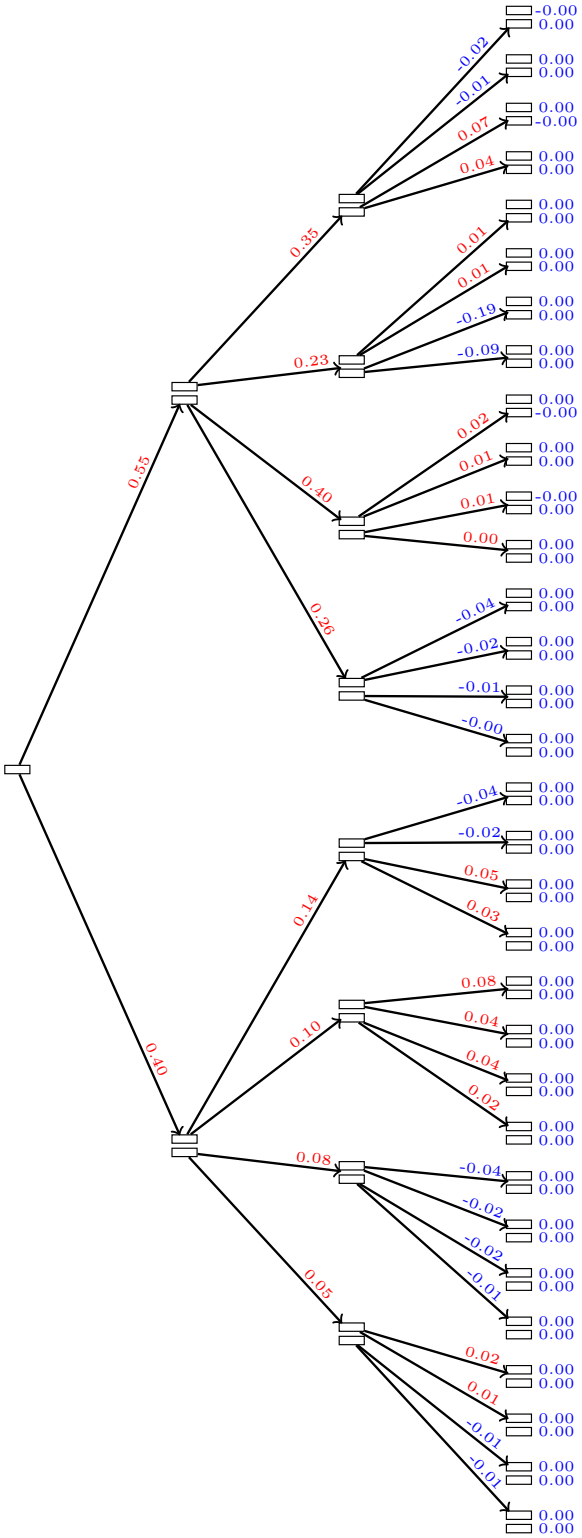


Figure 1: Average difference in Q^{init}-values between the trees in which the proportion of forward seuences was above the mean and the trees in which it was below the mean. Red values are positive and blue values are negative.

I think the main takeaway is exactly what we thought – that a fraction of rewards must be known, but not too high (since the average difference is rather small for most Q^{init}-values), because otherwise Need at the deeper horizons is too high which initiates reverse sequences.