Optimising Replay for Partially Observable Domains

Georgy, lab meeting 09/03/22

talk outline

- planning
- DYNA, prioritised sweeping
- explore-exploit
- normative theory of hippocampal replay
- replay in Bayesian bandits
- replay in BAMDPs
- limitations and future directions

planning

- generally speaking, planning refers to the process of computing a policy – i.e., coming up with an action to execute
- in value-based planning, this is achieved by estimating values associated with the available actions
- planning typically requires a model of the environment
- several aproaches exist: dynamic programming (DP), sample/simulation-based

planning $\rightarrow$ dynamic programming

- Bellman optimality equation

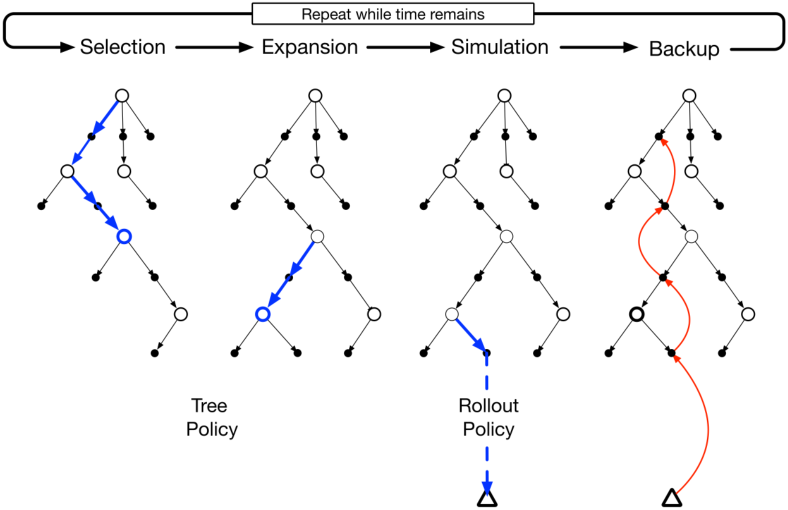
$$V^*(s) = \max_a \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V^*(s')]$$

- value iteration is one example value-based planning algorithm

$$V_k(s) \leftarrow \max_a \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V_{k-1}(s')]$$

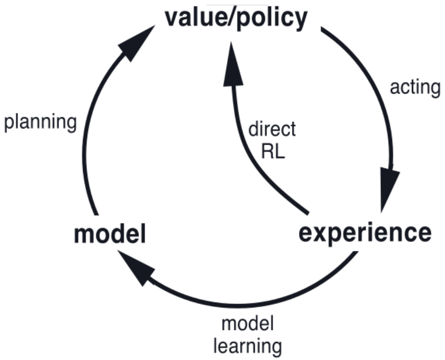
- assumes a known model of the environment $\mathcal{P}^a_{ss'}$
- performs synchronous sweeps over all states; computationally expensive; time-consuming

planning → MCTS



- MCTS is a simulation-based planning algorithm
- Tree and rollout policies
- Difficult to balance exploration and exploitation

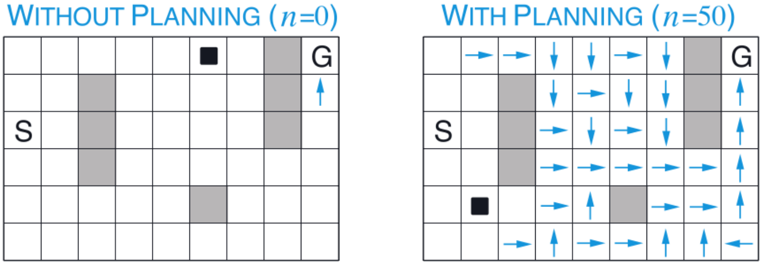
DYNA



Sutton (1990)

- DYNA is an integrated architecture
- combines a *reactive* MF policy and a *deliberate* MB system
- MB system is used offline to provide additional training to MF values

DYNA → example



- agent discovers online prediction erros (e.g., a goal)
- model inversion to additionally train MF values

$$Q^{MF}(s,a) \leftarrow Q^{MF}(s,a) + \alpha[R^{MB}(s') + \gamma \max_{a'} Q^{MB}(s',a') - Q^{MF}(s,a)]$$

- asynchronous DP

prioritised sweeping

- asynchronous DP updates can be optimised
- online discovery of a prediction error results in high offline prediction errors for the immediately preceeding states
- the idea of prioritised sweeping (Moore et al., 1993) is to execute the individual updates according to a priority queue, for instance:

$$p(s,a) = |Q^{MF}(s,a) + \alpha[R(s') + \gamma \max_{a'} Q^{MF}(s',a') - Q^{MF}(s,a)]|$$

prioritised sweeping $\rightarrow$ example

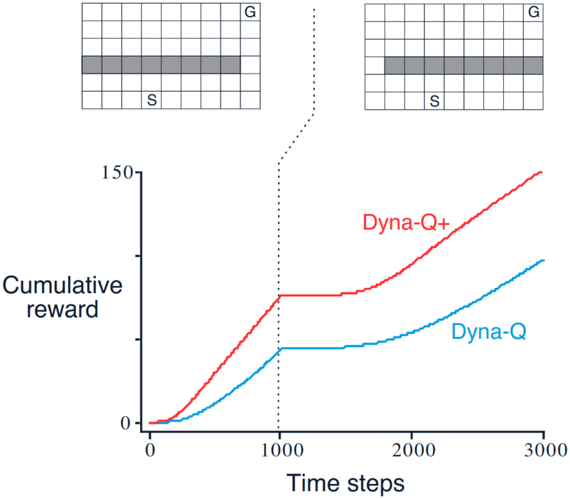
explore-exploit

- Optimal behaviour necessitates optimal balance of exploration and exploitation
- Multiple heuristics have been devised to encourage exploration
- One prominent heuristic is called an *exploration bonus*
- Sutton (1991) proposed to add an exploration bonus to the values of state-action pairs which have not been visited recently:

$$Q^{MB}(s,a) \leftarrow Q^{MB}(s,a) + \kappa\sqrt{\epsilon_{(s,a)}}$$

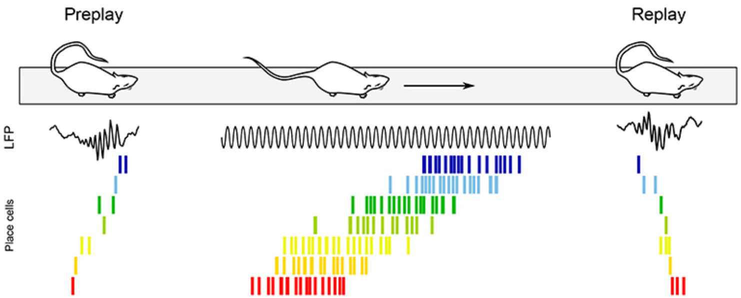
  where $\epsilon_{(s,a)}$ grows with the number of time steps elapsed since the state-action pair (s,a) was last tried, and κ controls the rate of exploration

explore-exploit $\rightarrow$ DYNA-Q+



- note that this exploration bonus is myopic
- propagates from distal locations due to the Q-learning rule
- however, it is blind towards what could be the consequences of exploration

hippocampal replay



Drieu et al. (2019)

- reinstatement of behaviourally-relevant neural activity during periods of quiet wakefullness and sleep (offline periods)
- the order of the replayed experiences is highly specific
- can proceed in forward and reverse directions
- forward replay seems to be predictive of the subsequent animal choices; reverse replay is highly sensitive to reward

hippocampal replay $\rightarrow$ normative theory

- Mattar & Daw (2018) realised that hippocampal replay might be a candidate mechanism for offline generative planning – acting in accordance with the DYNA system by supplying MB information to the animal's MF policy
- each replay experience, according to Mattar & Daw, corresponds to an update of an MF value for a state-action pair
- each replay update therefore changes the animal's policy at the state where that update is executed
- the fact that the order in which replay experiences proceed is highly specific suggests some sort of prioritisation (prioritised sweeping)

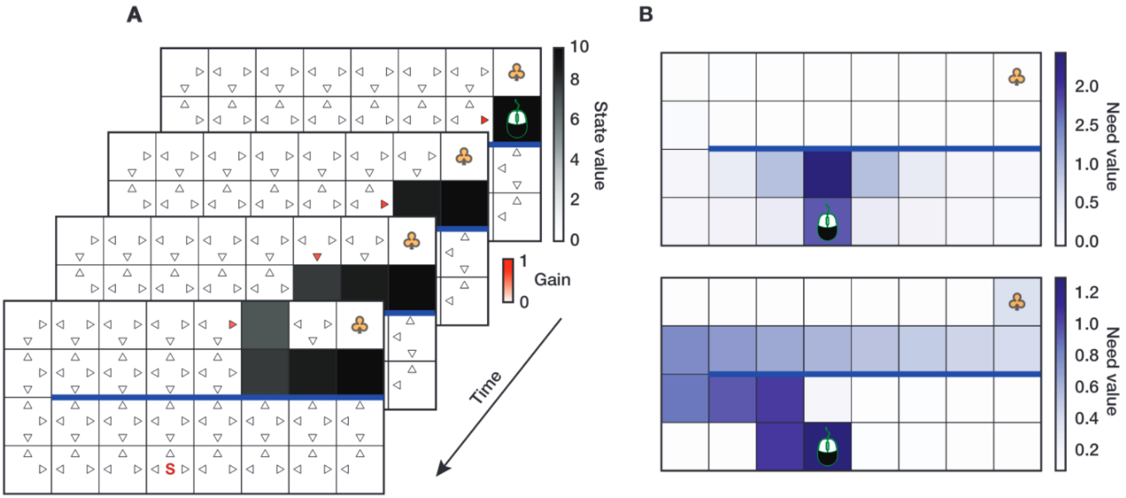
hippocampal replay $\rightarrow$ normative theory

- by the repeated unrolling of $v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s)$, where s is the animal's current location, M&D showed that the value of computation (i.e., replay update) performed at state s_k for action a_k can be decomposed into two terms

$$\text{EVB}(s_k, a_k) = \sum_{s' \in \mathcal{S}} \underbrace{\sum_{i=0}^{\infty} \gamma^i P(s \rightarrow s', i, \pi_{\text{old}})}_{\text{Need}} \times \underbrace{\sum_a [\pi_{\text{new}}(a \mid s') - \pi_{\text{old}}(a \mid s')] q_{\pi_{\text{new}}}(s', a)}_{\text{Gain}}$$

- Need is the expected discounted future occupancy of state s' under the animal's policy prior to the update, π_{old}
- Gain quantifies the local policy improvement at the state where the potential replay update is considered

hippocampal replay → normative theory



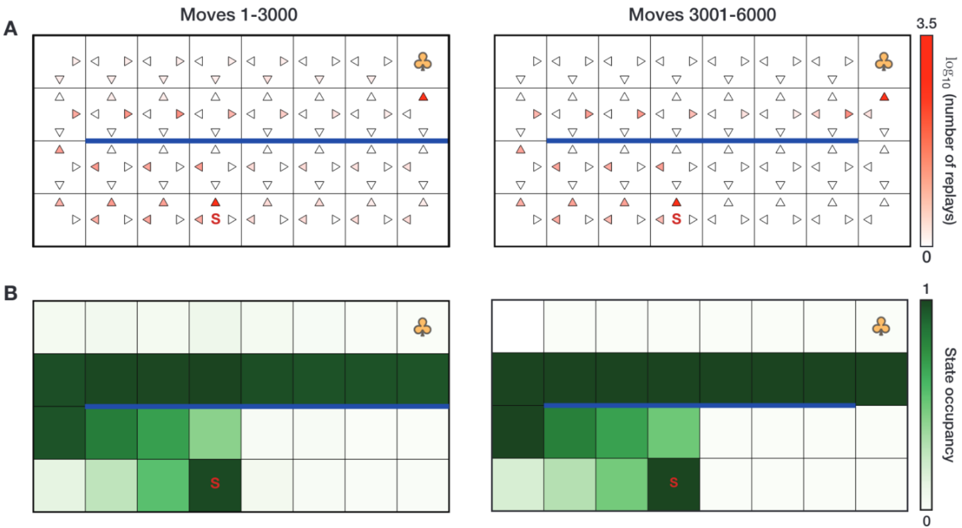
hippocampal replay $\rightarrow$ normative theory

$$\text{EVB}(s_k, a_k) = \cancel{\sum_{s' \in \mathcal{S}}} \underbrace{\sum_{i=0}^{\infty} \gamma^i P(s \rightarrow s', i, \pi_{\text{old}})}_{\text{Need}} \times \underbrace{\sum_a [\pi_{\text{new}}(a \mid s') - \pi_{\text{old}}(a \mid s')] q_{\pi_{\text{new}}}(s', a)}_{\text{Gain}}$$

Assumptions of the M&D model:

- policy updates are local, and thus M&D get rid of the first sum over $\mathcal{S}$
-
  - Note this means that the benefit of policy change at a distal state is only considered at the animal's current state
- Gain is expressed in terms of the *true* Q-values implied by the new policy, $q_{\pi_{\text{new}}}$
- the transition model P is known

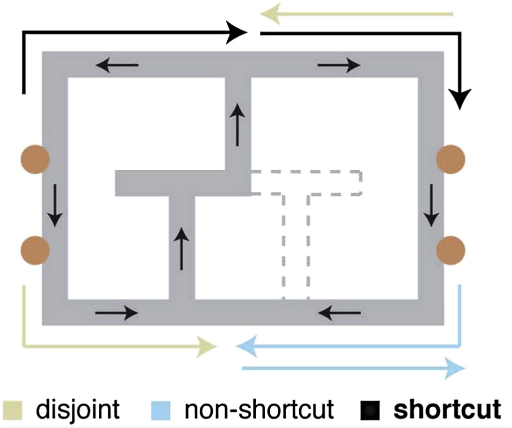
hippocampal replay → exploration?



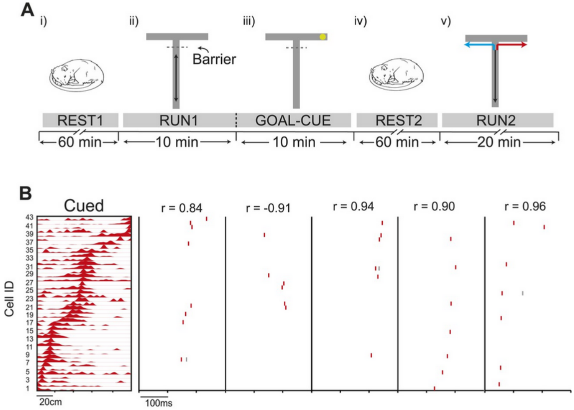
hippocampal replay $\rightarrow$ exploration?

- turns out the model of M&D doesn't explore well
- weird because the original intention of DYNA was to encourage exploration
- Need acts as a regulariser – once a policy is learnt it biases the selection of replay updates towards those states which the current policy already expects to visit (pure exploitation!)
- Gain (in fact, Need as well) doesn't account for the potential information which can be learnt and utilised in the future

hippocampal replay → exploration?



Gupta et al. (2010)



Ólafsdóttir et al. (2015)

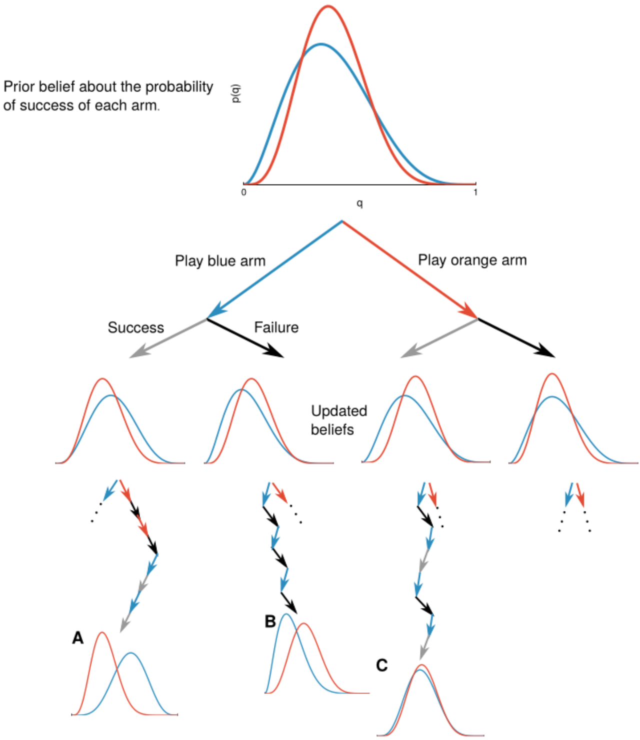
explore-exploit revisited

- in Bayesian RL, one often assumes some prior belief b over the unknown parameters
- in Bayes-adaptive MDPs [BAMDPs; Duff (1995)], the prior is over the unknown transition model parameters:

$$P(s' \mid s, a) = \int_{\Theta} P(s' \mid s, a, \theta) b(\theta) d\theta$$

- once integrated out, the model becomes known; moreover, it incorporates the agent's epistemic uncertainty – i.e., the 'known unknowns'
- the resulting policies are known as Bayes-adaptive policies, since they optimally trade-off exploration and exploitation

- planning in bandits can be visualised as belief trees



Guez (2015)

- optimal solution is known as the Gittins indices (Gittins, 1979), which correspond to the DP solution in the belief space

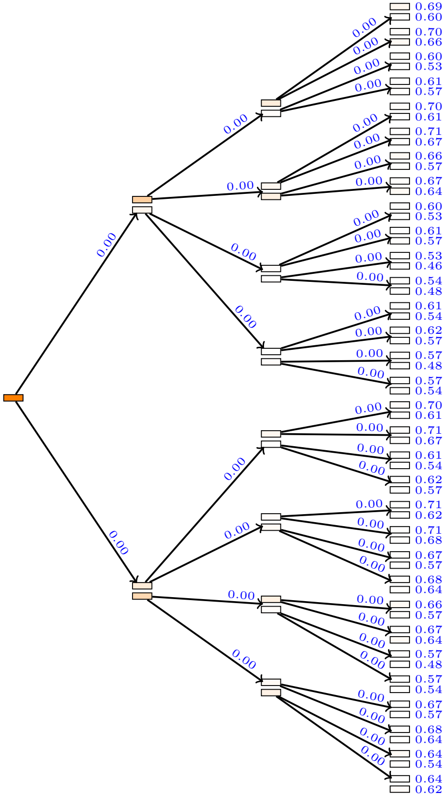
replay in Bayesian bandits

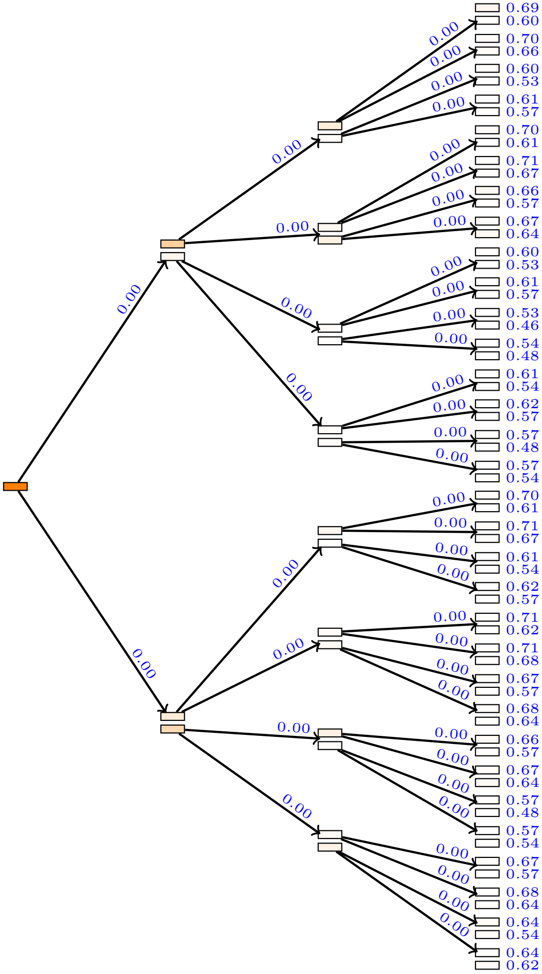
- we can do the same decomposition as in M&D but for belief states:

$$\text{EVB}(b_k, a_k) = \gamma^h P(b_\rho \to b_k, h, \pi_{\text{old}}) \times \sum_a [\pi_{\text{new}}(a \mid b_k) - \pi_{\text{old}}(a \mid b_k)] q_{\pi_{\text{new}}}(a, b_k)$$

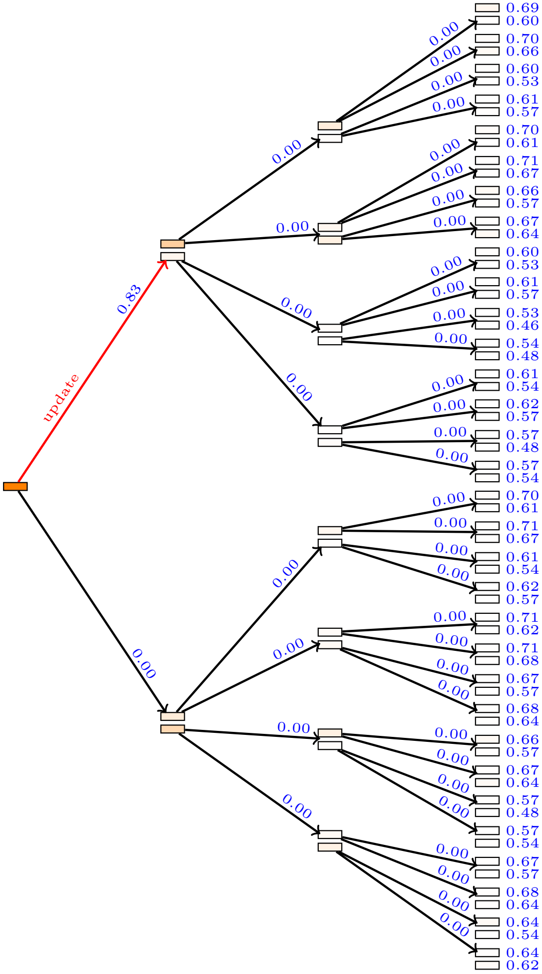
- where b_ρ is the prior belief at the root, and h is the horizon of belief b_k
- note that Need here is not cumulative; this is because each belief state can be visited at most once due to continual learning

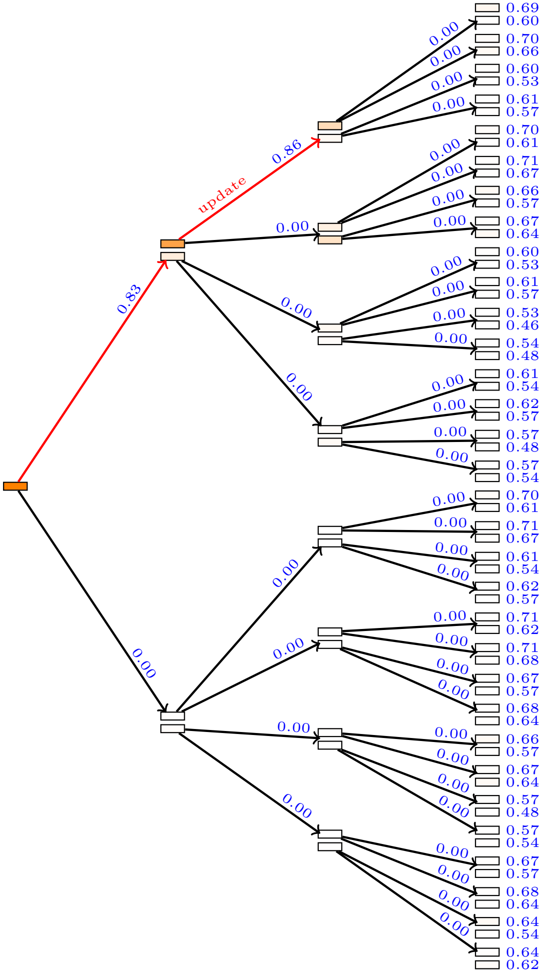
- each rectangle is a belief state
- colour intensity corresponds to Need, $\gamma^h P(b_\rho \rightarrow b_k, h, \pi_{\text{old}})$
- each arrow is an action
- blue numbers are Q-values
- in this tree example, the root belief is $(\alpha_0 = 5, \beta_0 = 1, \alpha_1 = 2, \beta_1 = 4)$
- the tree policy is a softmax with $\beta = 4$



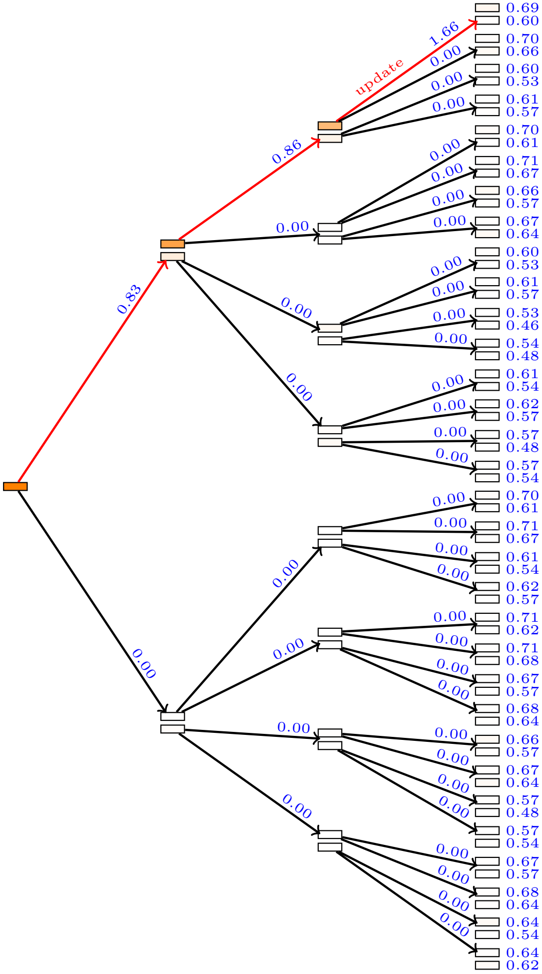


0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.69
0.60
0.70
0.66
0.60
0.53
0.61
0.57
0.70
0.61
0.71
0.67
0.66
0.57
0.67
0.64
0.60
0.53
0.61
0.57
0.53
0.46
0.54
0.48
0.61
0.54
0.62
0.57
0.57
0.48
0.57
0.54
0.70
0.61
0.71
0.67
0.61
0.54
0.62
0.57
0.71
0.62
0.71
0.68
0.67
0.57
0.68
0.64
0.66
0.57
0.67
0.64
0.57
0.48
0.57
0.54
0.67
0.57
0.68
0.64
0.64
0.54
0.64
0.62

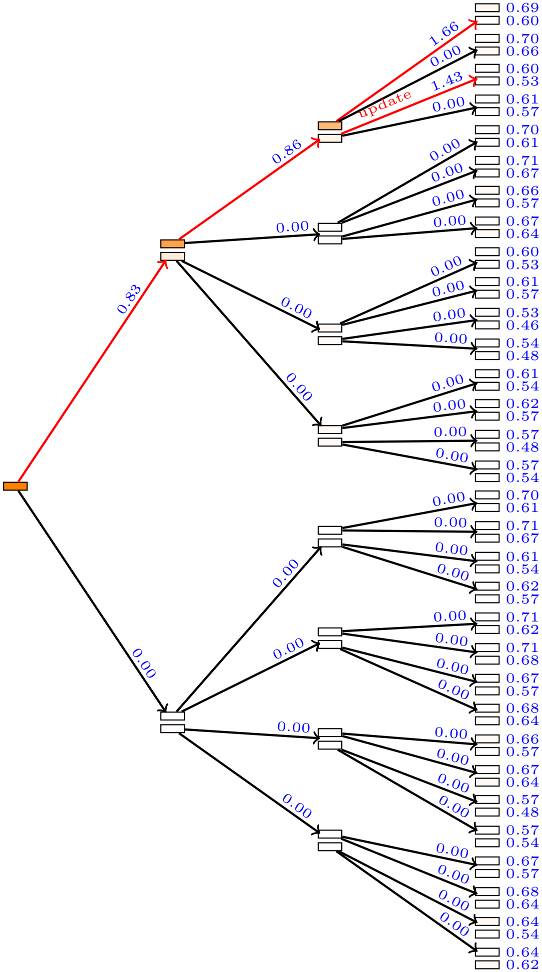

update
0.83
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.69
0.60
0.70
0.66
0.60
0.53
0.61
0.57
0.70
0.61
0.71
0.67
0.66
0.57
0.67
0.64
0.60
0.53
0.61
0.57
0.53
0.46
0.54
0.48
0.61
0.54
0.62
0.57
0.57
0.48
0.57
0.54
0.70
0.61
0.71
0.67
0.61
0.54
0.62
0.57
0.71
0.62
0.71
0.68
0.67
0.57
0.68
0.64
0.66
0.57
0.67
0.64
0.57
0.48
0.57
0.54
0.67
0.57
0.68
0.64
0.64
0.54
0.64
0.62

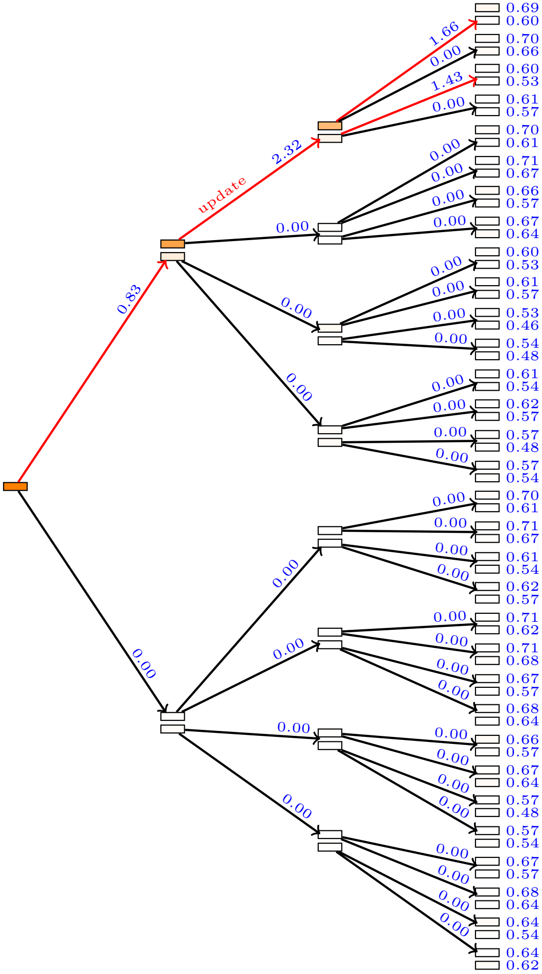


update
0.83
0.86
0.00
0.69
0.60
0.70
0.66
0.60
0.53
0.61
0.57
0.70
0.61
0.71
0.67
0.66
0.57
0.67
0.64
0.60
0.53
0.61
0.57
0.53
0.46
0.54
0.48
0.61
0.54
0.62
0.57
0.57
0.48
0.57
0.54
0.70
0.61
0.71
0.67
0.61
0.54
0.62
0.57
0.71
0.62
0.71
0.68
0.67
0.57
0.68
0.64
0.66
0.57
0.67
0.64
0.57
0.48
0.57
0.54
0.67
0.57
0.68
0.64
0.64
0.54
0.64
0.62

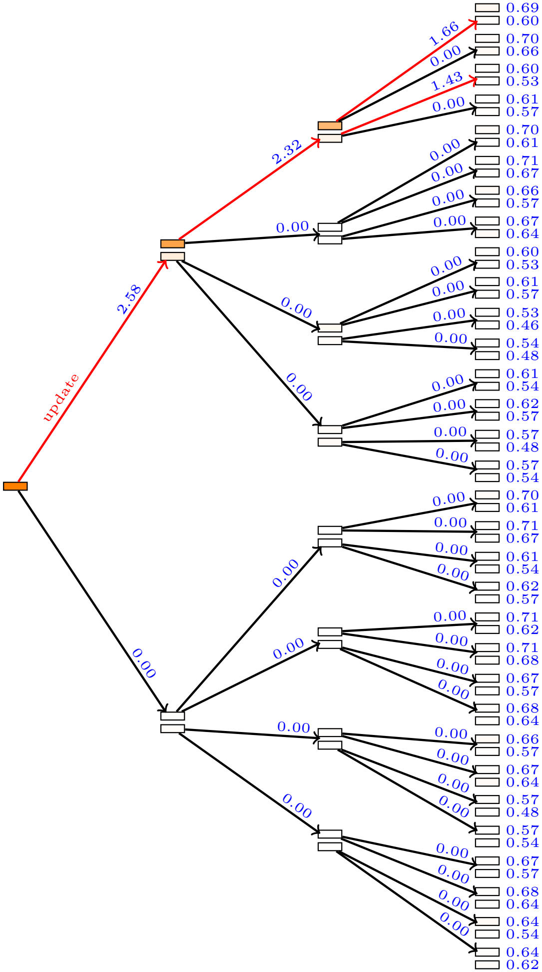


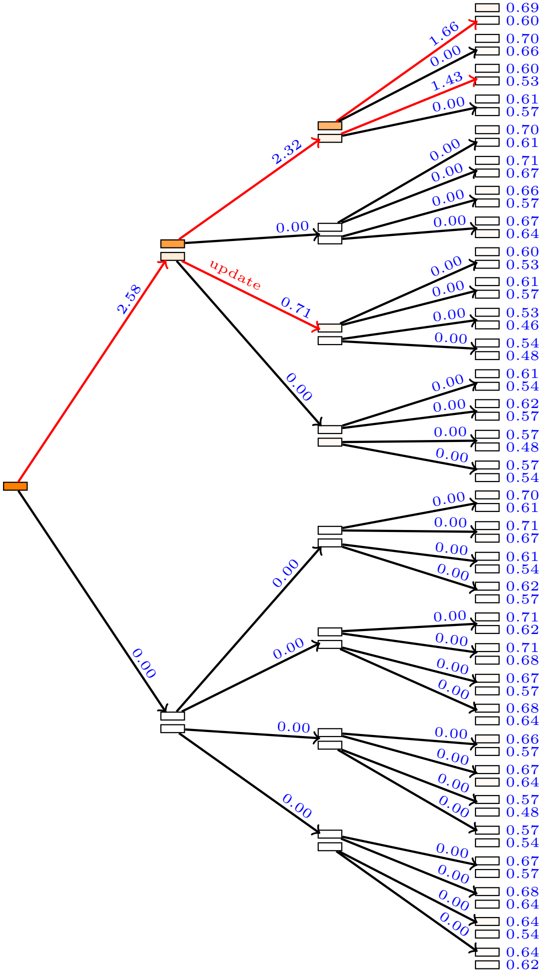
0.69
0.60
1.66
update
0.70
0.00
0.66
0.60
0.00
0.53
0.61
0.00
0.57
0.70
0.86
0.00
0.61
0.71
0.00
0.67
0.66
0.00
0.57
0.00
0.67
0.00
0.64
0.60
0.00
0.53
0.61
0.00
0.57
0.83
0.00
0.53
0.00
0.46
0.00
0.54
0.48
0.00
0.61
0.00
0.54
0.62
0.00
0.57
0.57
0.00
0.48
0.00
0.57
0.54
0.70
0.00
0.61
0.71
0.00
0.67
0.61
0.00
0.54
0.00
0.62
0.00
0.57
0.71
0.00
0.62
0.00
0.71
0.00
0.68
0.00
0.67
0.00
0.57
0.68
0.64
0.00
0.66
0.00
0.57
0.00
0.67
0.64
0.00
0.57
0.00
0.48
0.57
0.54
0.00
0.67
0.00
0.57
0.00
0.68
0.64
0.00
0.64
0.00
0.54
0.64
0.62


0.83
0.00
0.86
0.00
0.00
0.00
0.00
0.00
0.00
0.00
1.66
0.00
1.43
update
0.00
0.69
0.60
0.70
0.66
0.60
0.53
0.61
0.57
0.70
0.61
0.71
0.67
0.66
0.57
0.67
0.64
0.60
0.53
0.61
0.57
0.53
0.46
0.54
0.48
0.61
0.54
0.62
0.57
0.57
0.48
0.57
0.54
0.70
0.61
0.71
0.67
0.61
0.54
0.62
0.57
0.71
0.62
0.71
0.68
0.67
0.57
0.68
0.64
0.66
0.57
0.67
0.64
0.57
0.48
0.57
0.54
0.67
0.57
0.68
0.64
0.64
0.54
0.64
0.62

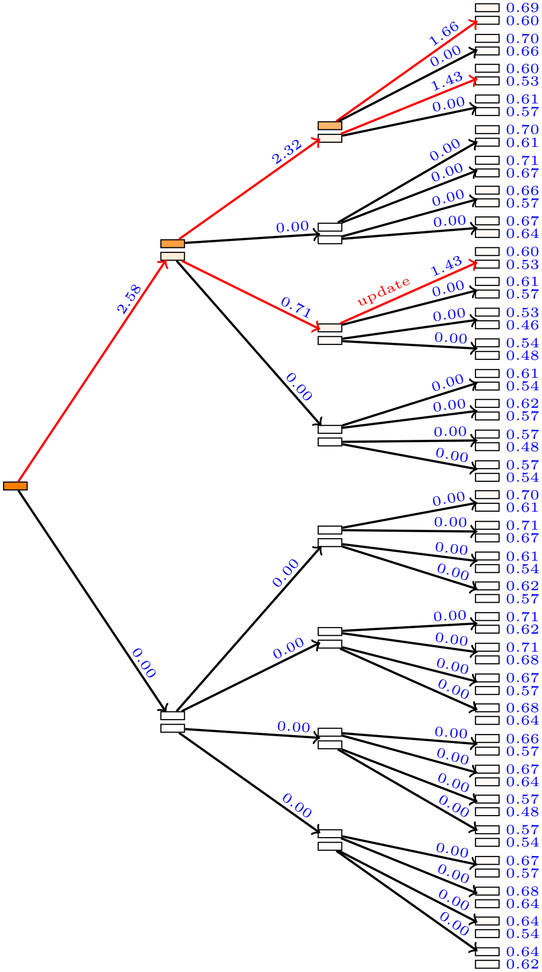


0.69
0.60
1.66
0.70
0.00
0.66
0.60
1.43
0.53
0.61
0.00
0.57
2.32
0.70
0.61
update
0.00
0.71
0.00
0.67
0.66
0.00
0.57
0.00
0.67
0.00
0.64
0.60
0.00
0.53
0.61
0.00
0.57
0.83
0.00
0.53
0.00
0.46
0.00
0.54
0.48
0.61
0.00
0.54
0.00
0.62
0.00
0.57
0.57
0.00
0.48
0.00
0.57
0.54
0.00
0.70
0.61
0.00
0.71
0.67
0.00
0.61
0.00
0.54
0.00
0.62
0.57
0.00
0.71
0.62
0.00
0.71
0.00
0.68
0.00
0.67
0.57
0.00
0.68
0.64
0.00
0.00
0.66
0.57
0.00
0.67
0.64
0.00
0.57
0.00
0.48
0.57
0.54
0.00
0.00
0.67
0.57
0.00
0.68
0.64
0.00
0.64
0.00
0.54
0.64
0.62

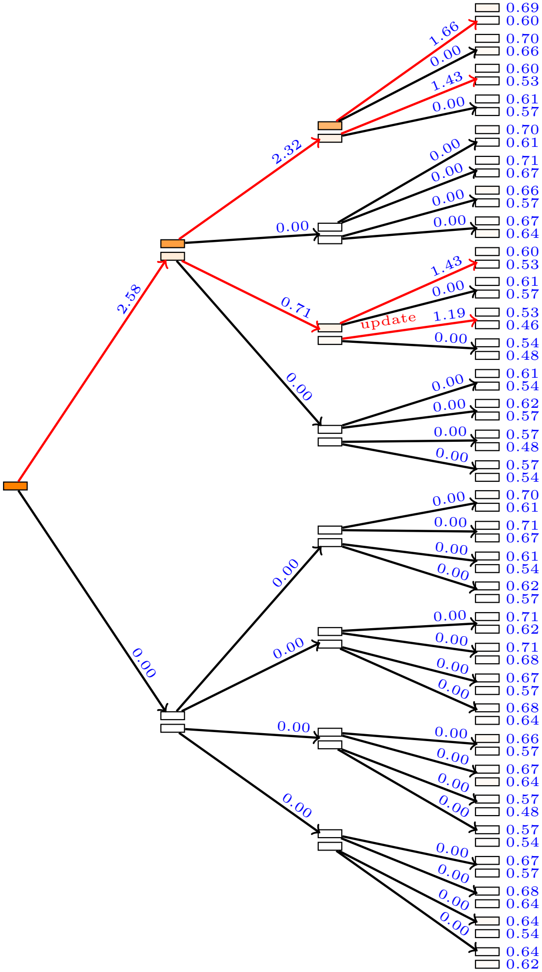

update
2.58
0.00
2.32
0.00
0.00
0.00
0.00
0.00
0.00
0.00
1.66
0.00
1.43
0.00
0.69
0.60
0.70
0.66
0.60
0.53
0.61
0.57
0.70
0.61
0.71
0.67
0.66
0.57
0.67
0.64
0.60
0.53
0.61
0.57
0.53
0.46
0.54
0.48
0.61
0.54
0.62
0.57
0.57
0.48
0.57
0.54
0.70
0.61
0.71
0.67
0.61
0.54
0.62
0.57
0.71
0.62
0.71
0.68
0.67
0.57
0.68
0.64
0.66
0.57
0.67
0.64
0.57
0.48
0.57
0.54
0.67
0.57
0.68
0.64
0.64
0.54
0.64
0.62

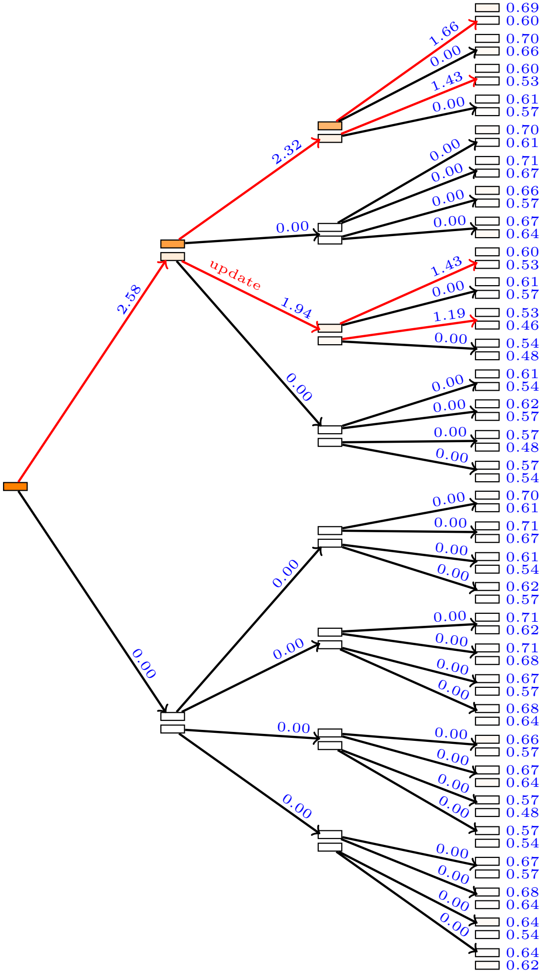


0.69
0.60
1.66
0.70
0.00
0.66
0.60
1.43
0.53
0.61
0.00
0.57
0.70
2.32
0.00
0.61
0.71
0.00
0.67
0.66
0.00
0.57
0.67
0.00
0.00
0.64
0.60
0.00
0.53
0.61
update
0.00
0.57
0.71
0.53
0.00
0.46
2.58
0.54
0.00
0.48
0.61
0.00
0.00
0.54
0.62
0.00
0.57
0.57
0.00
0.48
0.57
0.00
0.54
0.70
0.00
0.61
0.71
0.00
0.67
0.61
0.00
0.54
0.00
0.62
0.00
0.57
0.71
0.00
0.62
0.71
0.00
0.00
0.68
0.67
0.00
0.57
0.00
0.68
0.00
0.64
0.66
0.00
0.57
0.67
0.00
0.64
0.00
0.57
0.00
0.48
0.00
0.57
0.54
0.67
0.00
0.57
0.68
0.00
0.64
0.00
0.64
0.54
0.00
0.64
0.62



0.69
0.60
1.66
0.70
0.00
0.66
0.60
1.43
0.53
0.61
0.00
0.57
2.32
0.70
0.00
0.61
0.71
0.00
0.67
0.66
0.00
0.57
0.67
0.00
0.00
0.64
0.60
1.43
0.53
update
0.61
0.00
0.57
0.71
0.53
0.00
0.46
0.54
0.00
0.48
2.58
0.00
0.61
0.00
0.54
0.62
0.00
0.57
0.57
0.00
0.48
0.57
0.00
0.54
0.70
0.00
0.61
0.71
0.00
0.67
0.00
0.61
0.00
0.54
0.62
0.00
0.57
0.71
0.00
0.62
0.71
0.00
0.00
0.68
0.67
0.00
0.57
0.00
0.68
0.00
0.64
0.66
0.00
0.00
0.57
0.67
0.00
0.64
0.57
0.00
0.48
0.57
0.00
0.54
0.00
0.67
0.00
0.57
0.68
0.00
0.64
0.64
0.00
0.54
0.64
0.00
0.62

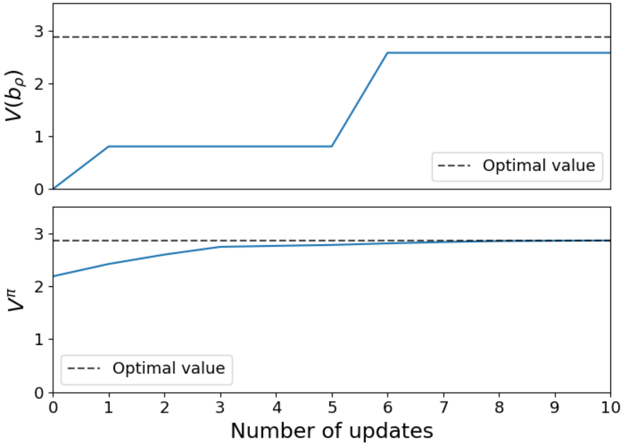

2.58
0.00
2.32
0.00
0.71
0.00
0.00
0.00
0.00
0.00
1.66
0.00
1.43
0.00
0.00
0.00
0.00
0.00
1.43
0.00
update
1.19
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.69
0.60
0.70
0.66
0.60
0.53
0.61
0.57
0.70
0.61
0.71
0.67
0.66
0.57
0.67
0.64
0.60
0.53
0.61
0.57
0.53
0.46
0.54
0.48
0.61
0.54
0.62
0.57
0.57
0.48
0.57
0.54
0.70
0.61
0.71
0.67
0.61
0.54
0.62
0.57
0.71
0.62
0.71
0.68
0.67
0.57
0.68
0.64
0.66
0.57
0.67
0.64
0.57
0.48
0.57
0.54
0.67
0.57
0.68
0.64
0.64
0.54
0.64
0.62

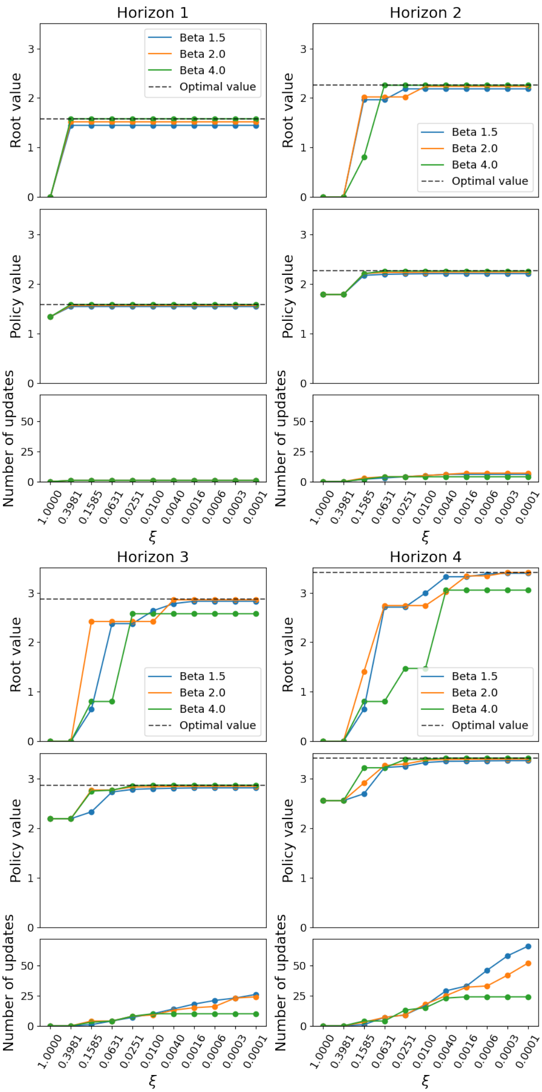


0.69
0.60
1.66
0.70
0.00
0.66
0.60
1.43
0.53
0.61
0.00
0.57
0.70
2.32
0.61
0.00
0.71
0.00
0.67
0.66
0.00
0.57
0.67
0.00
0.00
0.64
0.60
1.43
0.53
update
0.61
0.00
0.57
2.58
1.94
0.53
1.19
0.46
0.00
0.54
0.48
0.61
0.00
0.54
0.00
0.62
0.00
0.57
0.00
0.57
0.48
0.00
0.57
0.54
0.00
0.70
0.61
0.00
0.71
0.67
0.00
0.61
0.00
0.54
0.00
0.62
0.57
0.00
0.71
0.62
0.00
0.00
0.71
0.68
0.00
0.67
0.00
0.57
0.68
0.64
0.00
0.00
0.66
0.57
0.00
0.67
0.64
0.00
0.57
0.00
0.48
0.00
0.57
0.54
0.00
0.67
0.57
0.00
0.68
0.64
0.00
0.64
0.00
0.54
0.64
0.62

replay in Bayesian bandits

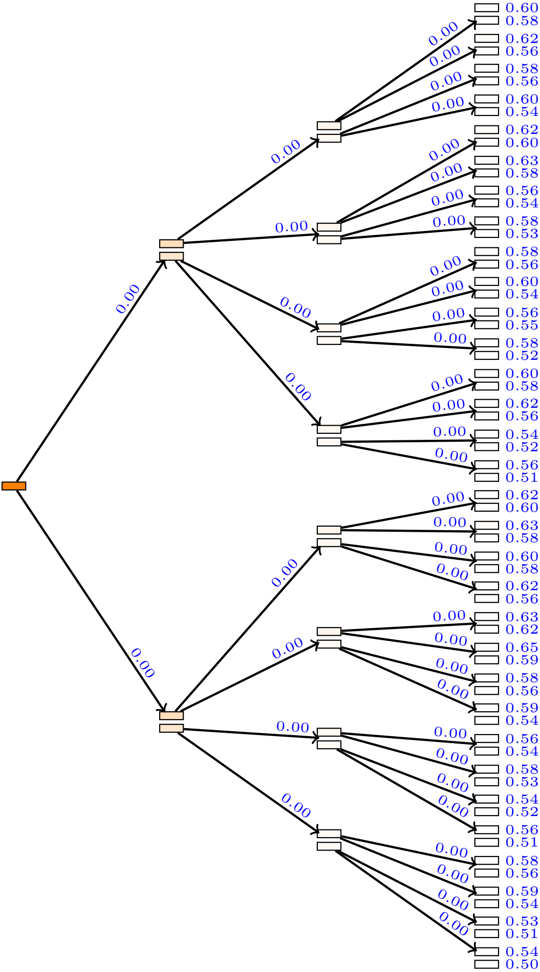
- the effects of each consequitive update on the root value, $V(b_\rho)$, as well as the value of the new (updated) policy evaluated in the tree, V^π

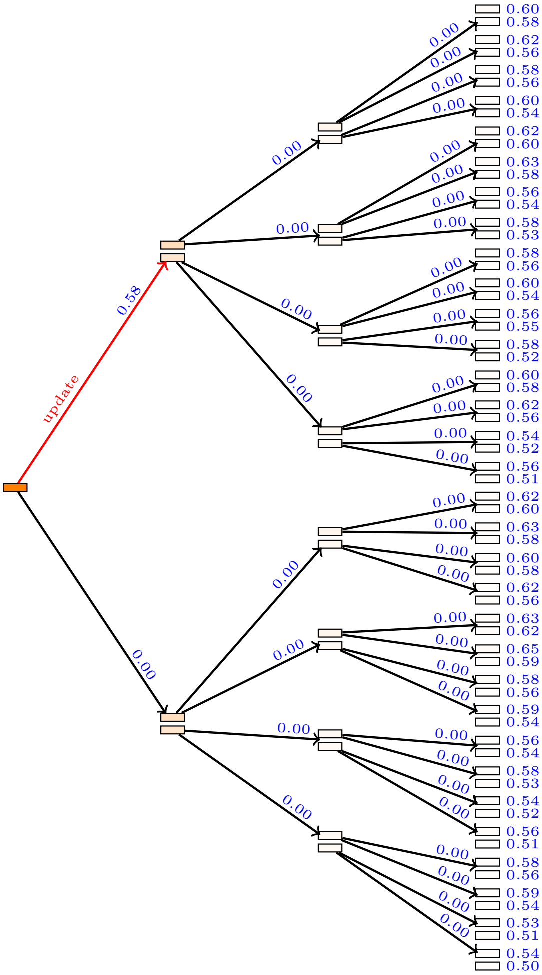



Horizon 1
Horizon 2
Horizon 3
Horizon 4
Beta 1.5
Beta 2.0
Beta 4.0
Optimal value
Root value
Policy value
Number of updates
1.0000
0.3981
0.1585
0.0631
0.0251
0.0100
0.0040
0.0016
0.0006
0.0003
0.0001
ξ

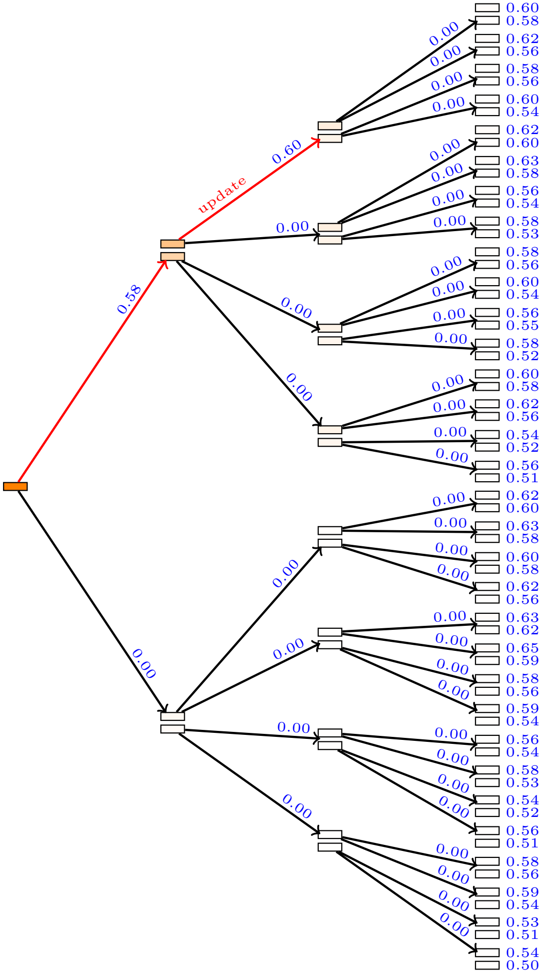
replay in Bayesian bandits

- one more example
- this time for the prior $(\alpha_0 = 14, \beta_0 = 10, \alpha_1 = 4, \beta_1 = 3)$

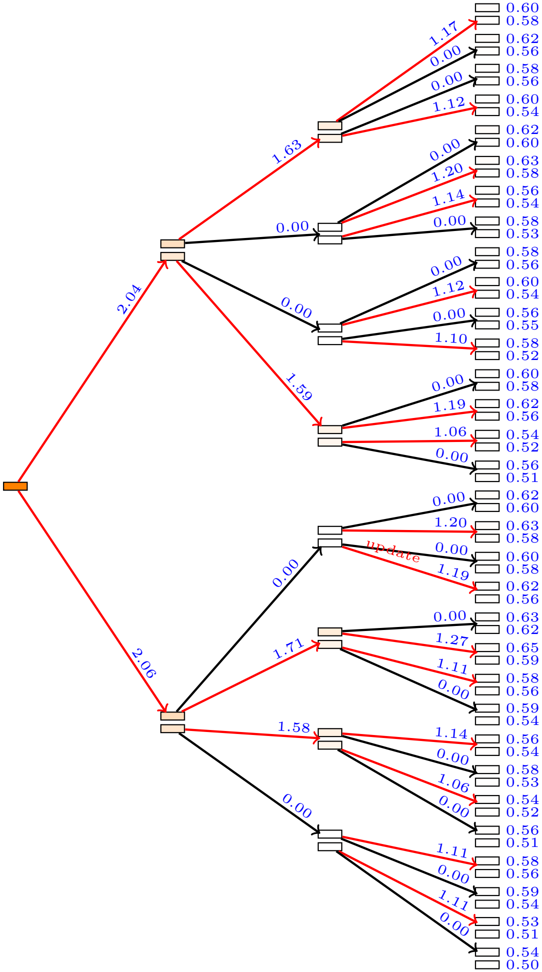

0.60
0.58
0.62
0.56
0.58
0.56
0.60
0.54
0.62
0.60
0.63
0.58
0.56
0.54
0.58
0.53
0.58
0.56
0.60
0.54
0.56
0.55
0.58
0.52
0.60
0.58
0.62
0.56
0.54
0.52
0.56
0.51
0.62
0.60
0.63
0.58
0.60
0.58
0.62
0.56
0.63
0.62
0.65
0.59
0.58
0.56
0.59
0.54
0.56
0.54
0.58
0.53
0.54
0.52
0.56
0.51
0.58
0.56
0.59
0.54
0.53
0.51
0.54
0.50



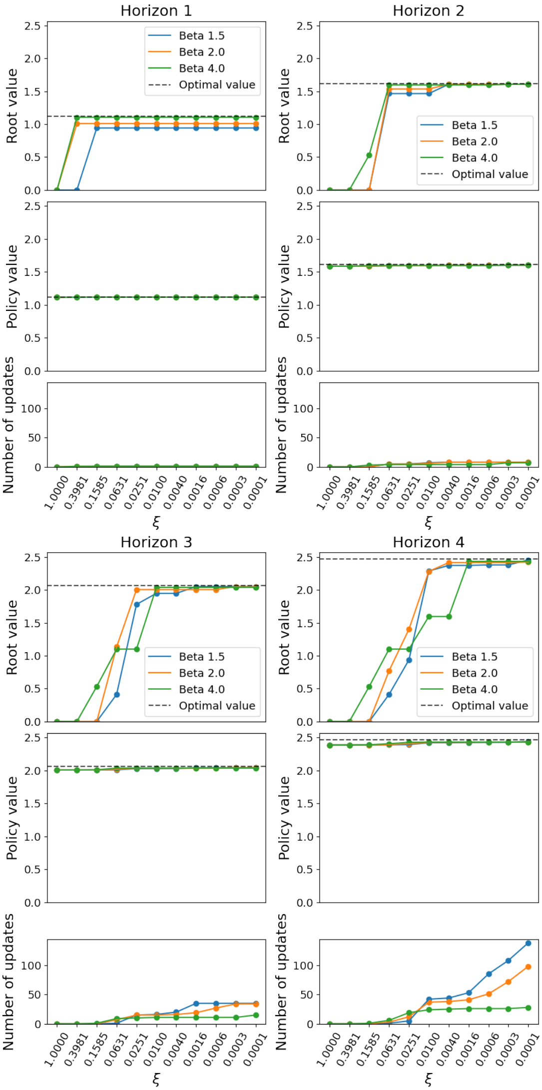
update
0.58
0.00
0.60
0.58
0.62
0.56
0.58
0.56
0.60
0.54
0.62
0.60
0.63
0.58
0.56
0.54
0.58
0.53
0.58
0.56
0.60
0.54
0.56
0.55
0.58
0.52
0.60
0.58
0.62
0.56
0.54
0.52
0.56
0.51
0.62
0.60
0.63
0.58
0.60
0.58
0.62
0.56
0.63
0.62
0.65
0.59
0.58
0.56
0.59
0.54
0.56
0.54
0.58
0.53
0.54
0.52
0.56
0.51
0.58
0.56
0.59
0.54
0.53
0.51
0.54
0.50



0.58
update
0.60
0.00
0.60
0.58
0.62
0.56
0.58
0.56
0.60
0.54
0.62
0.60
0.63
0.58
0.56
0.54
0.58
0.53
0.58
0.56
0.60
0.54
0.56
0.55
0.58
0.52
0.60
0.58
0.62
0.56
0.54
0.52
0.56
0.51
0.62
0.60
0.63
0.58
0.60
0.58
0.62
0.56
0.63
0.62
0.65
0.59
0.58
0.56
0.59
0.54
0.56
0.54
0.58
0.53
0.54
0.52
0.56
0.51
0.58
0.56
0.59
0.54
0.53
0.51
0.54
0.50



2.04
2.06
1.63
0.00
0.00
1.59
0.00
1.71
1.58
0.00
1.17
0.00
0.00
1.12
0.00
1.20
1.14
0.00
0.00
1.12
0.00
1.10
0.00
1.19
1.06
0.00
0.00
1.20
update
0.00
1.19
0.00
1.27
1.11
0.00
1.14
0.00
1.06
0.00
1.11
0.00
1.11
0.00
0.60
0.58
0.62
0.56
0.58
0.56
0.60
0.54
0.62
0.60
0.63
0.58
0.56
0.54
0.58
0.53
0.58
0.56
0.60
0.54
0.56
0.55
0.58
0.52
0.60
0.58
0.62
0.56
0.54
0.52
0.56
0.51
0.62
0.60
0.63
0.58
0.60
0.58
0.62
0.56
0.63
0.62
0.65
0.59
0.58
0.56
0.59
0.54
0.56
0.54
0.58
0.53
0.54
0.52
0.56
0.51
0.58
0.56
0.59
0.54
0.53
0.51
0.54
0.50


Horizon 1
Horizon 2
Horizon 3
Horizon 4
Root value
Policy value
Number of updates
Beta 1.5
Beta 2.0
Beta 4.0
Optimal value
1.0000
0.3981
0.1585
0.0631
0.0251
0.0100
0.0040
0.0016
0.0006
0.0003
0.0001
ξ

replay in Bayesian bandits

- planning in belief space can be optimised
- free generalisation for subsequent beliefs
- note the use of a softmax policy
- we are ultimately interested in exploration in DYNA-like systems which have an MF policy; MB system needs to convince the MF policy that something is worth exploring
- otherwise, the Need term $\gamma^h P(b_\rho \to b_k, h, \pi_{\text{old}})$ would be zero for those beliefs which the current MF policy doesn't expect to visit
- similar issues in MCTS – e.g., UCB

replay in BAMPDs

- in BAMDPs, we transition through belief states as well as physical states. Jointly, these are referred to as information states $z = \langle s, b \rangle$
- the prioritisation in BAMDPs thus takes the following form:

$$\text{EVB}(z_k, a_k) = \sum_{z' \in \mathcal{Z}} \sum_{i=0}^{\infty} \gamma^i P(z \to z', i, \pi_{\text{old}}) \times \sum_a [\pi_{\text{new}}(a \mid z') - \pi_{\text{old}}(a \mid z')] q_{\pi_{\text{new}}}(z', a)$$

- note that each $z = \langle s, b \rangle$ can still be visited at most once
- we know, however, that although the belief changes continuosly, the agent should still expect to visit the same physical state over and over again
- moreover, the summation over $\mathcal{Z}$ allows us to account for generalisation – that is, how updates at single information states (revealing a piece of information) affect policy at other beliefs (not quite there yet)