② Temporal difference

Issue: What happens if trajectories are too long ?

Typical update rule:

$$\text{NEW} = \text{OLD} + \alpha(\text{CURRENT} - \text{OLD})$$

Example:

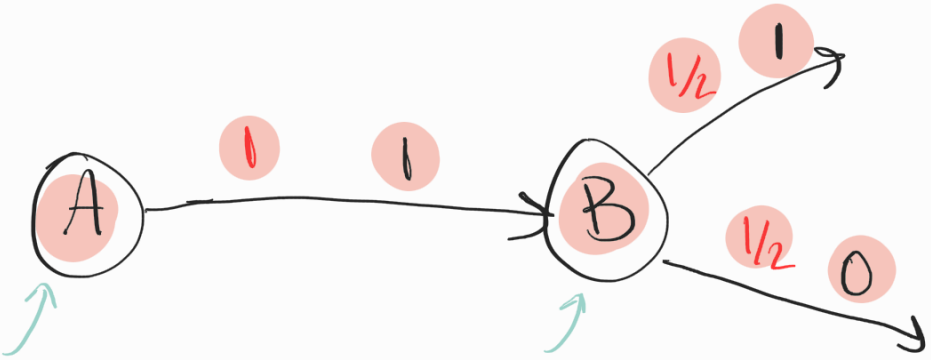
discount γ

A, B states

4 samples:

- A, 1, B, 1
- B, 0
- B, 1
- B, 0

MDP:



$$val(B) = 0.5$$

$$val(A) = 1 \cdot (1 + \gamma \cdot val(B))$$

$$= 1 + \gamma \cdot 0.5$$

temporal difference answer

Bootstrapping: estimating using estimates

Temporal difference update

Sample trajectories:

For each s_0, a_0, r_0, s_1:

$$\widehat{val}(s_0) = \widehat{val}(s_0) + \alpha \left(r_0 + \gamma \widehat{val}(s_1) - \widehat{val}(s_0) \right)$$

$$\text{NEW} = \text{OLD} + \alpha (\text{CURRENT} - \text{OLD})$$

Key advantage:

we update at each step (!)

Theorem: TD converge to $val(s)$

$\widehat{val}(s) \longrightarrow val(s)$

Recap: two approaches for evaluating $\widehat{val}_\sigma$ for σ a policy:

Monte Carlo: requires full trajectories

Temporal Difference: step by step.