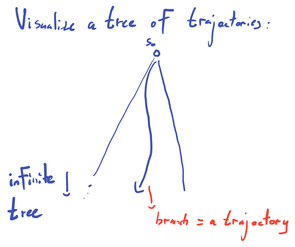
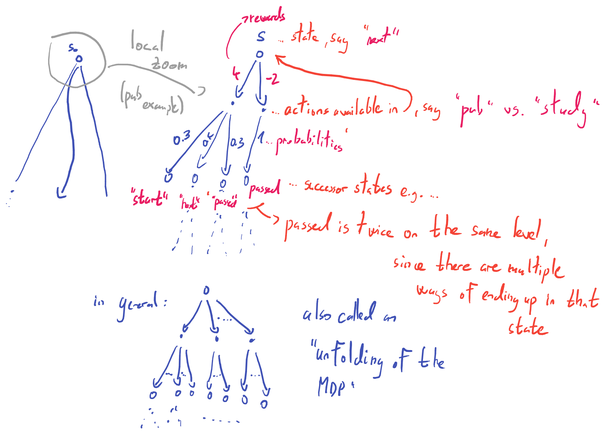
Visualizing MDP computations

Visualize a tree of trajectories:





Fixing a policy = assigning a probability to each edge outgoing from "o" (state)

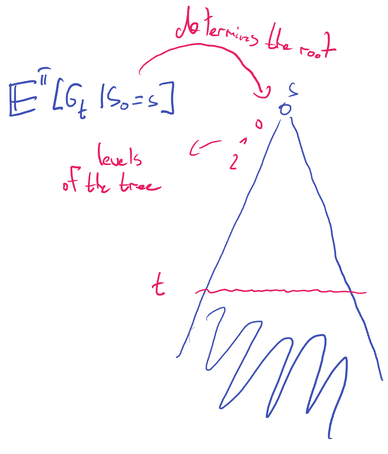
Then a probability of getting to a certain node of the tree
(= of producing a certain prefix of a trajectory)
= the product of all probabilities on the path from the root to that node

Payoffs:

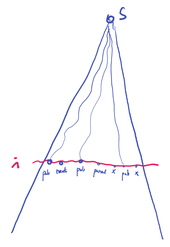
$r^\pi(s)$: imagine an unfolding rooted in s:

$r^\pi(s)$ = essentially the sum of all rewards appearing in the unfolding with the following 2 caveats:

- reward appearing at level i is multiplied by γ^i before being added to the sum
- each reward in the sum is multiplied by the probability of getting to the edge it labels



$\mathbb{E}^\pi[G_t \mid S_0 = s]$ is basically the sum of reward labels appearing below the level t (all of the tree, with the same caveats as above. The only difference is that rewards at level t+i are multiplied by γ^i instead of γ^{t+i}.



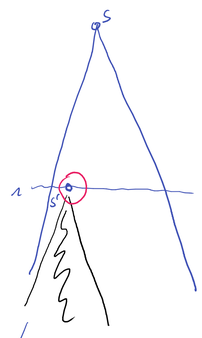
determines probabilities on edges outgoing from 'o'-nodes

$\mathbb{P}^\pi[S_i = \text{pub} \mid S_0 = s]$

... the probability of all paths from the root to an occurence of a node labelled 'pub' on level i

for each such path, multiply all the probabilities along the path

$\mathbb{E}^\pi[G_i \mid S_i = s', S_0 = s]$



↳ Like in $\mathbb{E}^\pi[G_i \mid S_0 = s]$, we are summing the rewards in this sub-tree, using the same discounting rules as for $\mathbb{E}^\pi[G_i \mid S_0 = s]$.

However, there is a difference: $[S_i = s']$ tells us that we assume we were in s' at step i. We are not interested in the probability of $S_i = s'$ happening, since we assumed it happened.

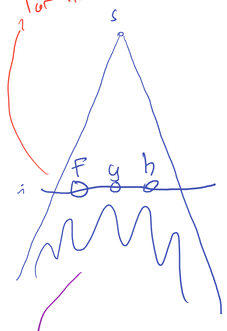
Hence, each reward is weighted by the probability of getting from the o-highlighted occurence of s' to the edge corresponding to that reward.

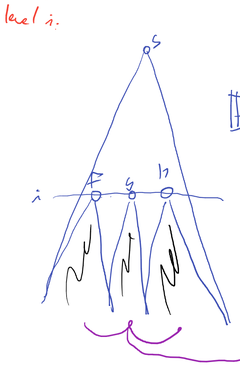
Note that this is the same as considering the sub-tree rooted in the o-highlighted node as a stand-alone unfolding.

Hence, $\mathbb{E}^\pi[G_i \mid S_i = s', S_0 = s] = \mathbb{E}^\pi[G_0 \mid S_0 = s']$

$= \mathbb{E}^\pi[G_i \mid S_i = s']$, since we really do not care what happened before level i.

For illustration, let there be 3 states at level i: states F, g, h





$$\mathbb{E}^\pi[G_i \mid S_0 = s] = \sum_{s'} \mathbb{P}^\pi[S_i = s' \mid S_0 = s] \cdot \mathbb{E}^\pi[G_i \mid S_i = s']$$

Here, each reward is weighted by prob. of the whole path from root s to the reward.

, but

Here, the rewards are only weighted by probabilities of paths from the sub-tree root to the reward edge (see above). Hence, to get equality with the left hand side, each $\mathbb{E}^\pi[G_i \mid S_i = s']$ must be multiplied by the probability of getting from the root s to the respective s' on level i.

$$\mathbb{E}^\pi[G_i \mid S_0 = s] = \mathbb{P}^\pi[S_i = F \mid S_0 = s] \cdot \mathbb{E}^\pi[G_i \mid S_i = F] + \mathbb{P}^\pi[S_i = g \mid S_0 = s] \cdot \mathbb{E}^\pi[G_i \mid S_i = g] + \mathbb{P}^\pi[S_i = h \mid S_0 = s] \cdot \mathbb{E}^\pi[G_i \mid S_i = h]$$