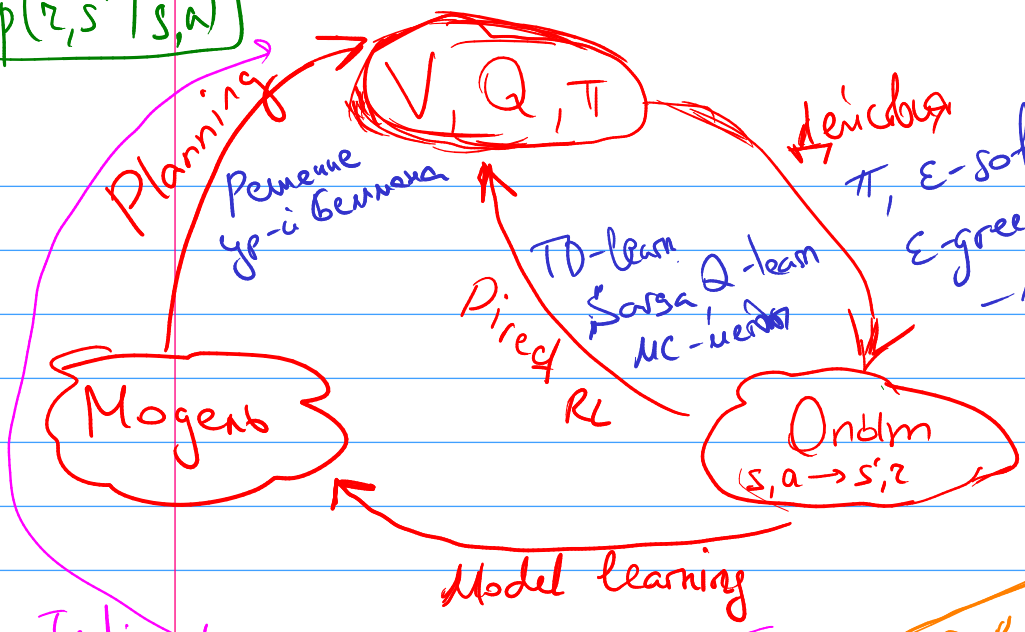
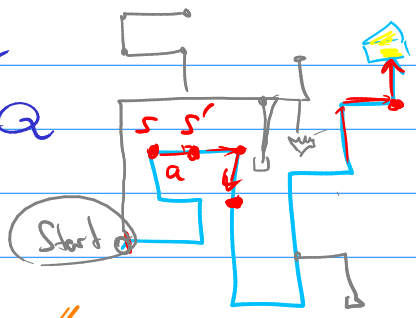
$p(r, s' | s, a)$

$V(s)$, $Q(s,a)$

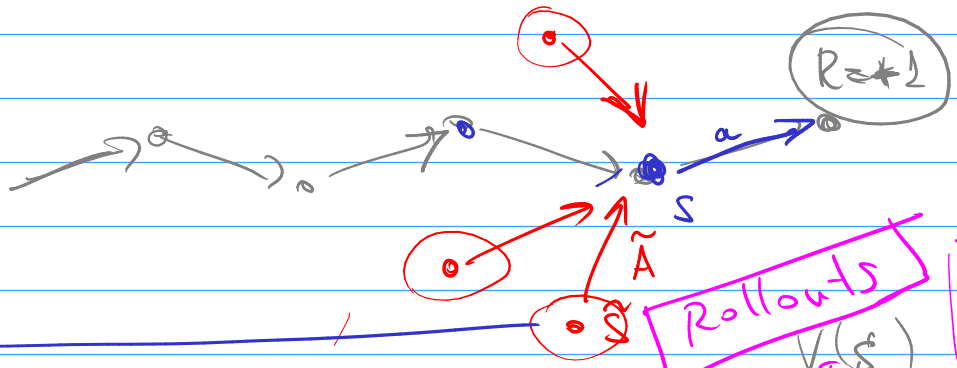




$(s, a) \to (s', r)$

$$Q(S_t, A_t) := Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

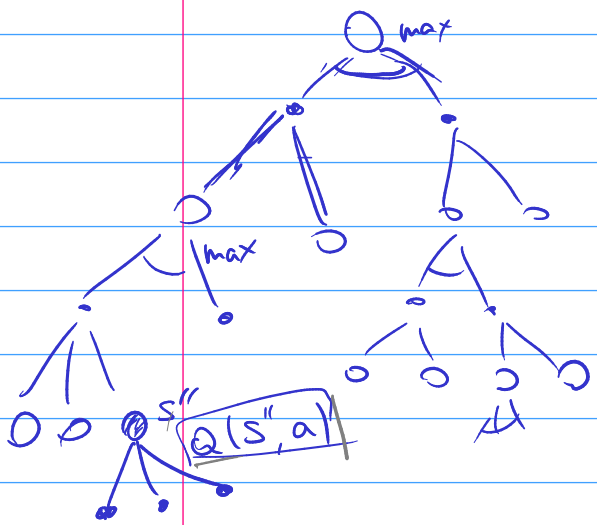
Dyna

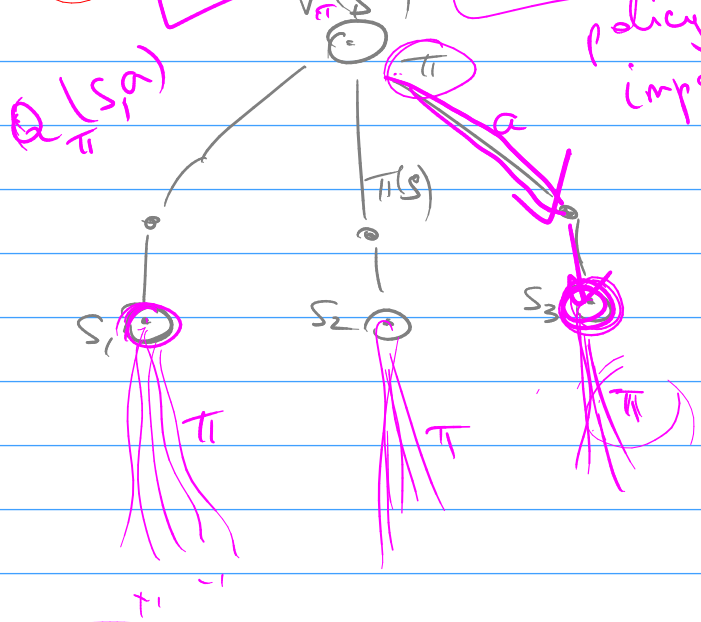


Rollouts

$Q_\pi(s,a) \geq V_\pi(s)$

policy improvement





MCTS - Monte-Carlo Tree Search

n-step methods

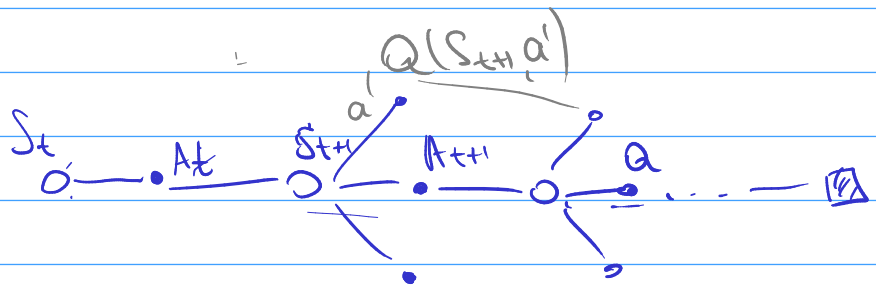
$$Q(S_t, A_t) \leftarrow Q(S_t, a) + \alpha\left(\text{TARGET} - Q(S_t, a)\right)$$

MC: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{T-t-1} R_T$

TD: $G_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$

- t
- $t+1$
- $t+2$
- $\vdots$
- $t+n$

2-step TD: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2}, A_{t+2})$



$$G_t = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q(S_{t+1}, a) + \gamma \pi(A_{t+1}|S_{t+1}) \left(R_{t+2} + \gamma \sum_{a \neq A_{t+2}} \pi(a|S_{t+2}) Q(S_{t+2}, a) + \gamma \pi(A_{t+2}|S_{t+2})(\dots)\right)$$