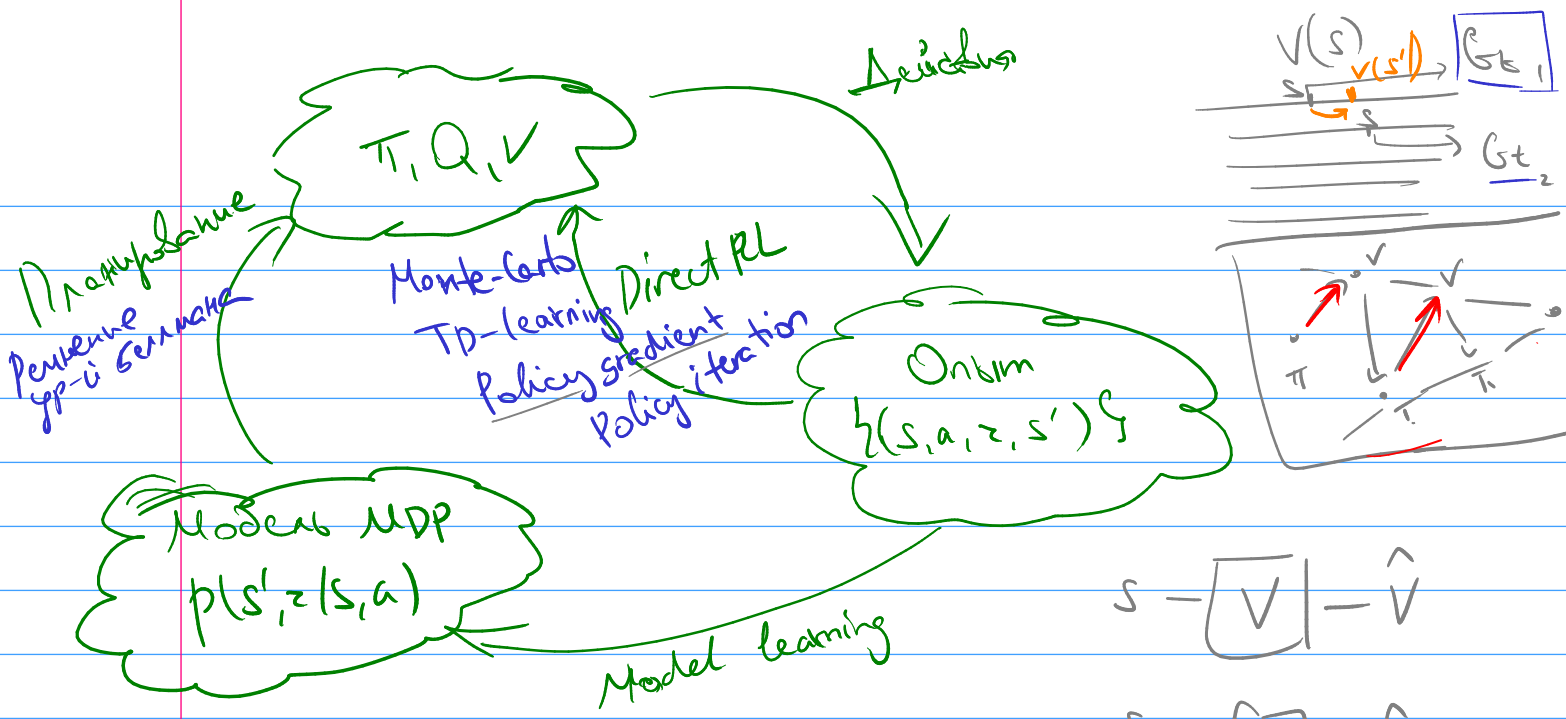






$s \to [V] \to \hat{v}$

$s, a \to [Q] \to \hat{q}$

Policy gradient

$\pi(a|s,\theta)$

$s \to [\pi] \to$ Softmax $\to$ Softmax $\to \pi(a|s)$

$$J(\theta) = V_{\pi_\theta}(s_0) \xrightarrow[\theta]{} \max$$

$$\nabla_\theta J(\theta) = \nabla_\theta \left[ \mathbb{E}_{\pi_\theta} [ G_0 ] \right] = \dots =$$

Policy gradient theorem

$$\nabla_\theta J(\theta) \propto \mathbb{E}_\pi \left[ \cdot G_t \cdot \nabla_\theta [\ln \pi_\theta(A_t|S_t)] \right]$$

$\pi$: — - ——

REINFORCE

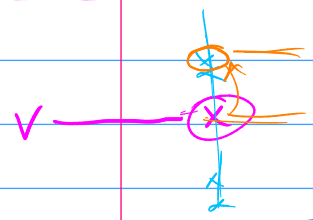
1) $s, a \to [NN] \to \hat{Q}(s,a)$

2) $s \to [NN] \to \hat{Q}(s,a)$

$$\theta = \theta + \alpha \left( r + \gamma \max_{a'} Q_{\theta}(s',a') - Q_{\theta}(s,a) \right) \cdot \nabla_{\theta} Q$$

Winner's curse

$$\hat{Q}(s',a') \approx Q(s',a')$$

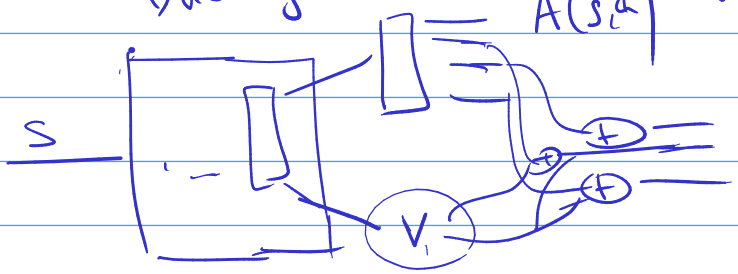


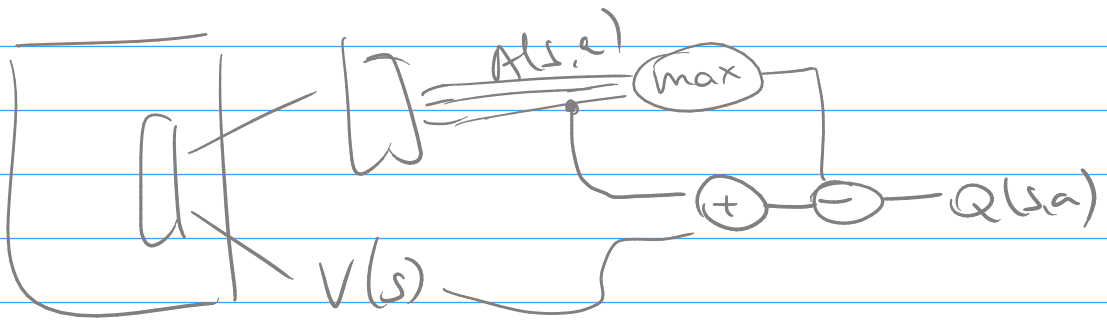

$$V_i \sim \mathcal{N}(V_i \mid V, \sigma_i^2)$$

---

Dueling DQN

3)

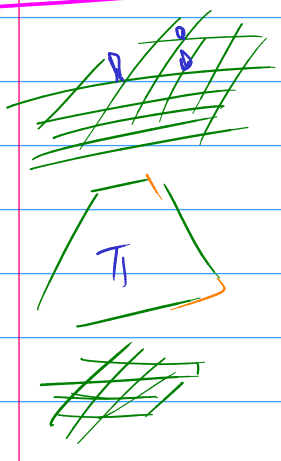


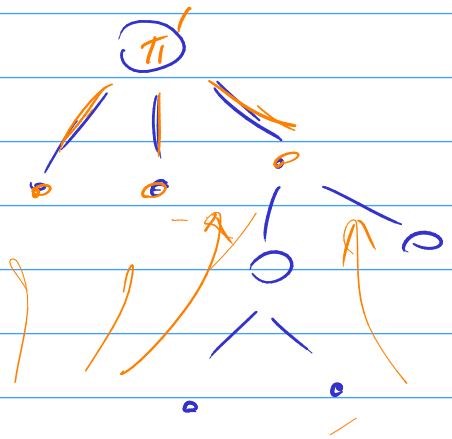



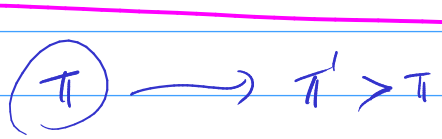

$$Q(s,a) = V(s) + \left( A(s,a) - \max_{a'} A(s,a') \right)$$

$$- \left( \frac{1}{|A|} \sum_{a'} A(s,a') \right)$$

---



$$p_t \approx \pi(a_{\cdot\cdot} \mid \text{history})$$

$$v_t \approx V(s \mid \text{history})$$

$$r_t \approx R_{t+1}$$