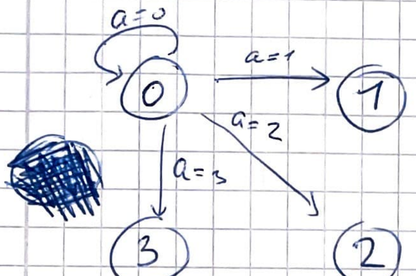
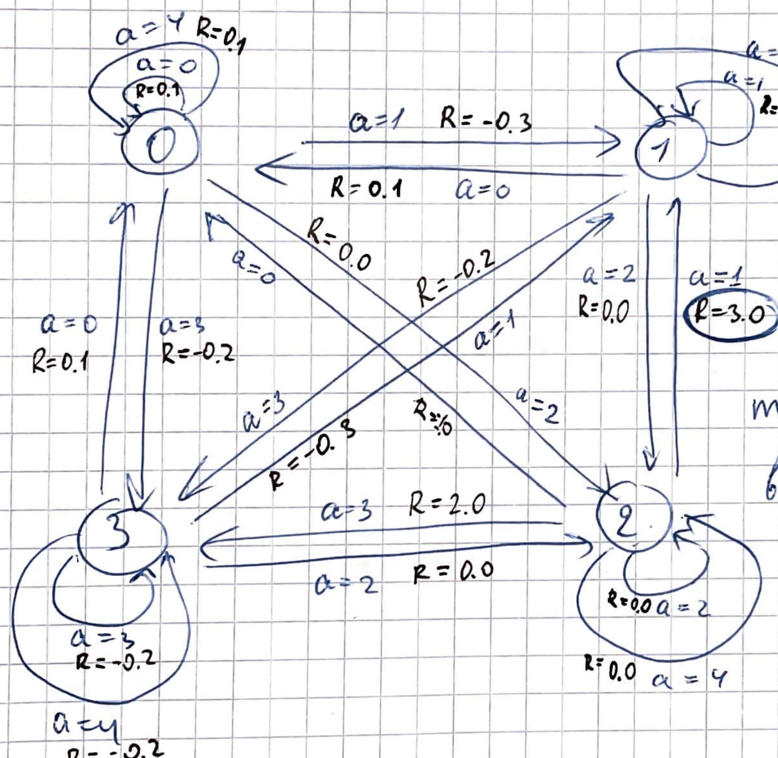
HW2 ① Test Environment

$S = \{0, 1, 2, 3\}$

$A = \{0, 1, 2, 3, 4\}$ act. $0 \le i \le 3 \rightarrow$ goes to $s = i$

act. $4 \rightarrow$ stays in state





$S_0 = 0$

$\max G_0$ — ?

всього можливих траєкторій: $5 \cdot 5 \cdot 5 \cdot 5 \cdot 5 = 5^5$

? траєкторія, що максимізує G_0:

$S_0 = 0,\ a_0 = \ \ ,\ R_0 = 0.1$

$S_1 = 3,\ a_1 = 2,\ R_1 = 0$

$S_2 = 0,\ a_2 = 2,\ R_2 = 0$

$S_3 = ?,\ a_3 = 1,\ R_3 = $

$S_4 = 2,\ a_4 = 1,\ R_4 = 3$

$S_5 = 1$

~~$G_0 = 3,3$. Дана траєкторія єдина із початкових до кінцевих дій, "...". Отже рівний між $S=0$ $a=4$ та $S=0$, $a=0$: 1) усі $R_i \ge 0$, $0 \le i \le 4$ 2) $R_3 = 3$ — тут важлива … що … виконується цей перехід додаючи … G_0. А решту траєкторії будували так, щоб максимізувати~~

$G_0 = 6,1$. Жодна інша траєкторія не матиме більшу G_0, бо ми двічі проходимо $2 \rightarrow 1$, що дає реворд $3+3=6$. Ще 0,1 вдалося досягти на проміжних етапах.

4. Linear Approximation

Represent Q-values as a parametric function

$Q_{\bar{w}}(s,a)$; $\bar{w} \in \mathbb{R}^p$ - parameters of the function (typically weights and biases of a linear function or a nn)

update rule:

$$\bar{w} \leftarrow \bar{w} + \alpha \left(r + \gamma \max_{a' \in A} Q_{\bar{w}}(s', a') - Q_{\bar{w}}(s,a) \right) \nabla_{\bar{w}} Q_{\bar{w}}(s,a)$$

$Q_{\bar{w}}(s,a) = \bar{w}^T \delta(s,a)$ - linear appr., $\bar{w} \in \mathbb{R}^{|S| \cdot |A|}$,

$\delta : S \times A \to \mathbb{R}^{|S| \cdot |A|}$

$$[\delta(s,a)]_{(s',a')} = \begin{cases} 1, & s' = s, a' = a \\ 0, & \text{otherwise.} \end{cases}$$

$$Q_{\bar{w}}(s,a) = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix} \begin{bmatrix} \delta_1 & \ldots & \delta \end{bmatrix} = \sum_{i=1}^{|S| \cdot |A|} w_i \underbrace{\delta_i(s,a)}_{\{1, 0}} ;$$

$$\nabla_{\bar{w}} Q_{\bar{w}}(s,a) = \left(\frac{\partial Q_{\bar{w}}}{\partial w_1}, \ldots \right) = (\delta_1(s,a), \ldots) = \delta(s,a).$$

Show: (1) and (2) are equal when $Q_{\bar{w}}(s,a) = \bar{w}^T \delta(s,a)$

(1): $Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a' \in A} Q(s',a') - Q(s,a) \right)$

$Q_{\bar{w}}(s,a) = \bar{w}^T \delta(s,a)$

$$\bar{w}^T \delta(s,a) \leftarrow w^T \delta(s,a) + \alpha \left(r + \gamma \max_{a' \in A} Q_{\bar{w}}(s',a') - Q_{\bar{w}}(s,a) \right)$$

$\forall \tilde{s}, \tilde{a} \in S \times A$:

$$w_{\tilde{s}, \tilde{a}} \leftarrow w_{\tilde{s}, \tilde{a}} + \alpha \left(r + \gamma \max_{a' \in A} Q_{\bar{w}}(s',a') - Q_{\bar{w}}(\tilde{s}, \tilde{a}) \right) + $$
$$+ 0 \cdot \alpha \left(r + \gamma \max_{a' \in A} Q_{\bar{w}}(s',a') - Q_{\bar{w}}(\tilde{s}, \tilde{a}) \right) + 0 \cdot \ldots$$

будь-які інші набори

$$\Updownarrow$$

$$\bar{w} \leftarrow \bar{w} + \alpha \left(r + \gamma \max_{a' \in A} Q_{\bar{w}}(s',a') - Q_{\bar{w}}(s,a) \right) \cdot \underbrace{\delta(s,a)}_{= \nabla_{\bar{w}} Q_{\bar{w}}(s,a)}.$$

⑧ Distributions induced by a policy

infinite-horizon MDP $\mathcal{M} = \langle S, A, R, P, \gamma \rangle$

Stohastic policies $\pi : S \mapsto \Delta(A)$

$\pi(a|s)$ – prob. of taking action a in a state s.

$$\forall s : \sum_a \pi(a|s) = 1.$$

assume MDP has a single fixed start. state $s_0 \in S$

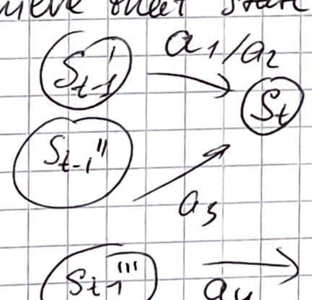
a) the probability of sampling a trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$ from running π in $\mathcal{M}$.

$$V^{\pi}(s_0) = \mathbb{E}_{\tau \sim p^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 \right]$$

$$p^{\pi}(\tau) = p(s_0, a_0) \cdot p(s_1, a_0) \cdot \ldots = \pi(a_0 | s_0) \cdot \pi(a_1 | s_0) \times \ldots = \prod_{i=0}^{\infty} \pi(a_i | s_i)$$

b) $p^{\pi}(s_t = s)$ = P{being in state s at timestep t while following the policy π}.

If the actor is in state s at time t, this means that at time $t-1$ he took an action from previous step to achieve that state. So,



$$p^{\pi}(s_t = s) = \sum_{\pi(s_{t-1}, a') = s_t} \pi(a' | s_{t-1}).$$

? c) $d^{\pi}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t p^{\pi}(s_t = s)$

$d^{\pi}(s|a) = d^{\pi}(s) \pi(a|s)$

$f : S \times A \to \mathbb{R}$. Prove:

$$\mathbb{E}_{\tau \sim p^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) \right] = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}} \left[\mathbb{E}_{a \sim \pi(s)} [f(s,a)] \right]$$

?

– $f(s,a) = 1$, $\forall (s,a) \in S \times A$.

$$\underbrace{\mathbb{E}_{\tau \sim p^{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \right]}_{= \frac{1}{1-\gamma}} = \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d^{\pi}} \left[\mathbb{E}_{a \sim \pi(s)} [1] \right]}_{= 1}$$

$\frac{1}{1-\gamma}$ – не залежить від π

$$\frac{1}{1-\gamma} = \frac{1}{1-\gamma}.$$

BRUNNEN