

Title: PSI 2018/2019 - Machine Learning - Lecture 10

Speakers: Bohdan Kulchyt'skyy

Collection: PSI 2018/2019 - Machine Learning (Hayward Sierens)

Date: April 05, 2019 - 9:00 AM

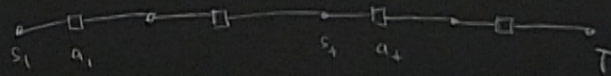
URL: <http://pirsa.org/19040005>

Policy gradient

$$\nabla_{\theta} J(\pi_{\theta}) = \left\langle \sum_{t=1}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \cdot \hat{Q}^{\pi}(s_t, a_t) \right\rangle_{\tau \sim p_{\pi}(\tau)}$$

$$\approx \sum_{t=1}^T \nabla_{\theta} \pi_{\theta}(a_t | s_t) \frac{1}{\pi_{\theta}(a_t | s_t)} \hat{Q}^{\pi}(s_t, a_t)$$

↑
1 trajectory



$$\sum_{t=0}^T r(s_t, a_t)$$

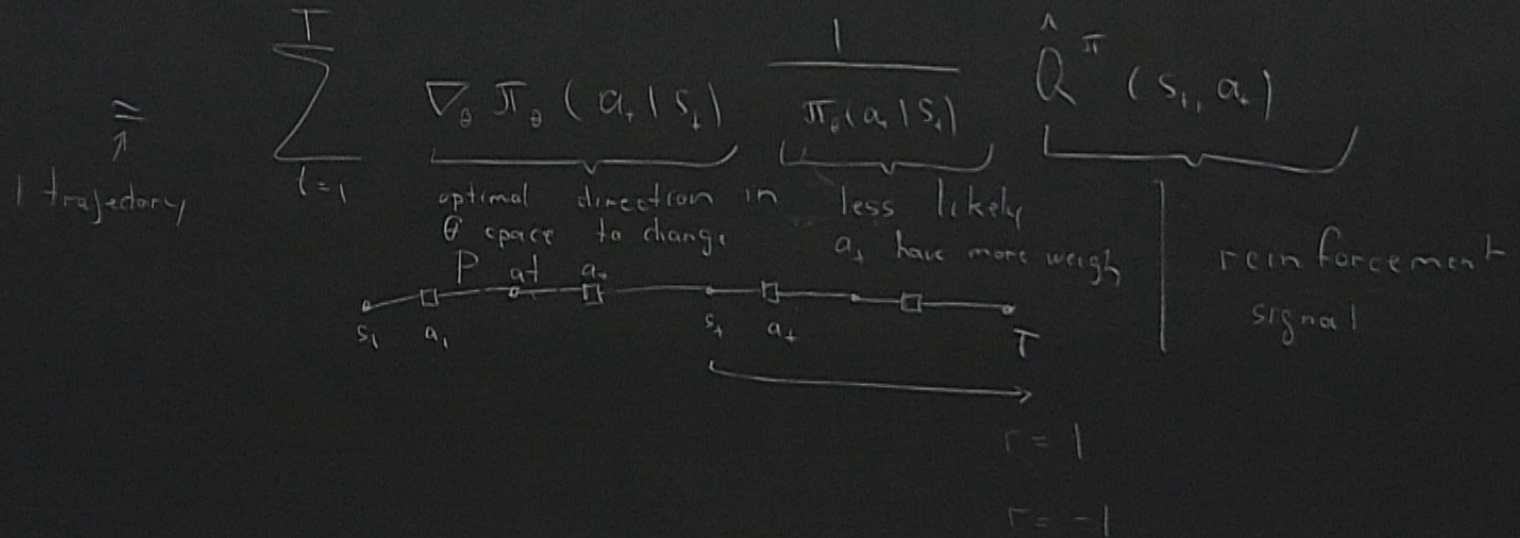
$$a_t \rangle \tau \sim p_\pi(\tau)$$

(s_t, a_t)

$$p_\pi(\tau) = p_1(s_1) \prod_{t=1}^T \underbrace{p(s_{t+1} | s_t, a_t) \pi(a_t | s_t)}$$

Policy gradient

$$\nabla_{\theta} J(\pi_{\theta}) = \left\langle \sum_{t=1}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \cdot \hat{Q}^{\pi}(s_t, a_t) \right\rangle_{\tau \sim p_{\pi}(\tau)} = \sum_{t=1}^T r(s_t, a_t)$$



$\Rightarrow$
↑
1 trajectory

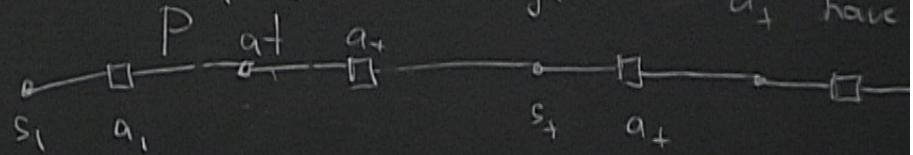
$$\sum_{t=1}^T$$

$$\nabla_{\theta} \pi_{\theta}(a_t | s_t)$$

optimal direction in θ space to change

$$\frac{1}{\pi_{\theta}(a_t | s_t)}$$

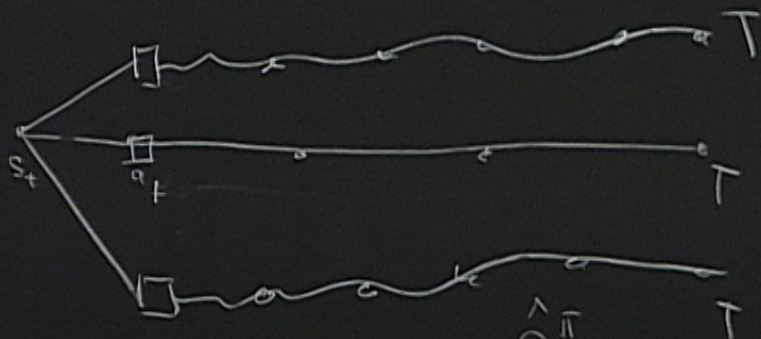
less a_t have



- if positive, + , a_t is encouraged by gradient

- — , a_t is discouraged

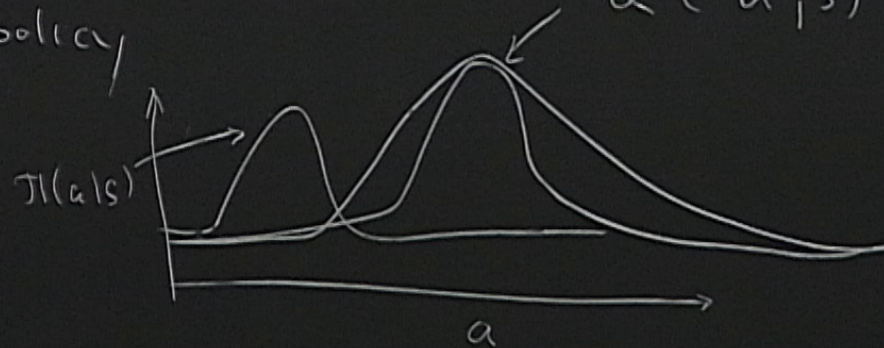
$$- \hat{Q}^{\pi}(s_t, a_t)$$



$$\begin{aligned} Q^{\pi}(s_k, a) &= \mathbb{E} \left[\hat{Q}^{\pi}(s_t, a_t) \mid s_t, a_t = a \right]_{\tau \sim p(\tau)} \\ &= \mathbb{E} \left[\sum_{t'=t}^T r(s_{t'}, a_{t'}) \mid s_t, a_t = a \right]_{\tau \sim p(\tau)} \end{aligned}$$

$$\pi'(a|s) = \begin{cases} \arg\max_a Q^\pi(s, a) & \text{if } s = s_+ \\ \pi(a|s) & \text{if } s \neq s_+ \end{cases}$$

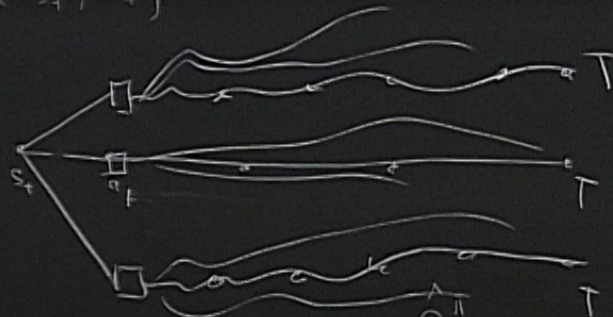
↑
improved policy



$\tau \sim p(\tau)$

$\tau \sim p(\tau)$

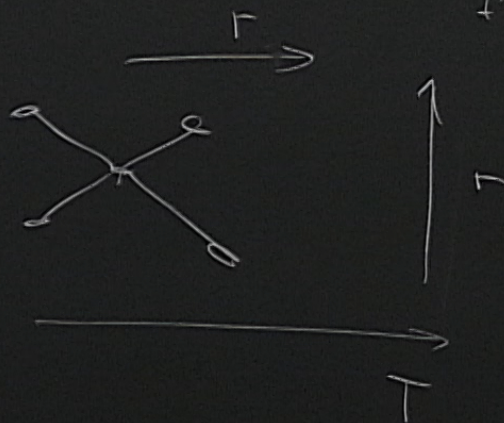
$$- \hat{Q}^{\pi}(s_t, a_t)$$



$$\begin{aligned} Q^{\pi}(s_t, a) &= \langle \hat{Q}^{\pi}(s_t, a_t) \mid s_t, a_t = a \rangle_{\tau \sim p(\tau)} \\ &= \langle \sum_{t'=t}^T r(s_{t'}, a_{t'}) \mid s_t, a_t = a \rangle_{\tau \sim p(\tau)} \end{aligned}$$

Small disgression

"episodic" $Q^\pi(s, a) = \mathbb{E} \left[\sum_{t'=t}^T r(s_{t'}, a_{t'}) \mid s_t, a_t = a \right] \tau \sim p$



disagression

$$Q^{\pi}(s, a) = \mathbb{E} \left[\sum_{t'=t}^{T \rightarrow \infty} \gamma^{t'-t} r(s_{t'}, a_{t'}) \mid s_t, a_t = a \right] \tau \sim p(\tau) < \infty$$

$$Q^{\pi}(s, a) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}) \mid s_t, a_t \right] \tau \sim p(\tau)$$

discount factor $\gamma \in [0, 1)$

$\gamma \rightarrow 0$ prioritize immediate reward

$\gamma \rightarrow 1$ cumulative reward

$\gamma = 0.99$

Small disgression

"episodic" $Q^\pi(s, a) = \mathbb{E} \left[\sum_{t'=t}^{\tau} r(s_{t'}, a_{t'}) \mid s_t, a_t = a \right] \tau \sim p(\tau)$

"continuity" tasks

$$Q^\pi(s, a) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}) \mid s_t, a_t \right] \tau \sim p$$

discount factor $\gamma \in [0, 1)$

$\gamma \rightarrow 0$ prioritize immediate reward

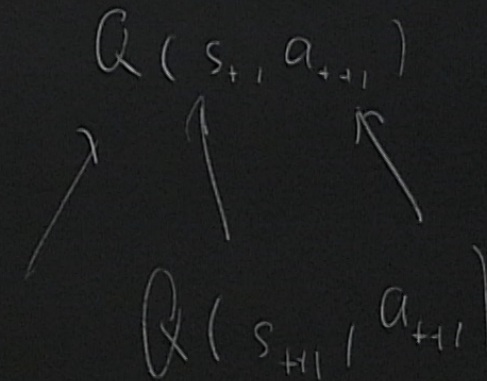
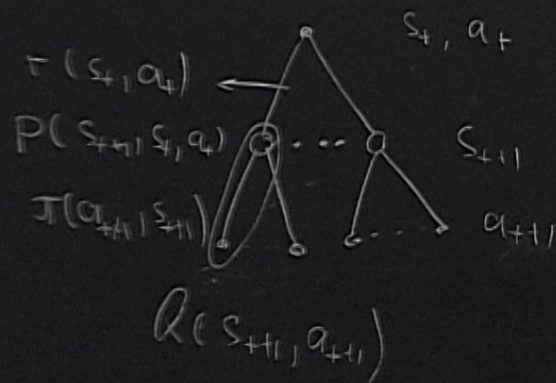
$\gamma \rightarrow 1$ cumulative reward

$$Q^{\pi}(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \sum_{a_{t+1}} \pi(a_{t+1} | s_{t+1}) Q^{\pi}(s_{t+1}, a_{t+1})$$

Bellman equation

$|S| \times |A|$

backup
diagram



T_γ — linear map, operator

$Q_0 \sim$ random values

$$Q_{k+1} = T_\gamma Q_k$$

$$T_\gamma(r, p, \pi, \gamma)$$

$$TQ_1 \longrightarrow TQ_2$$

$$Q_1 \longrightarrow Q_2$$

$$0 \leq \gamma < 1$$

T_γ is contractive

$$\gamma \|Q_1 - Q_2\|_\infty$$

$$\geq \|T_\gamma(Q_1 - TQ_2)\|_\infty$$

$$T_\gamma(r, p, \pi, \gamma)$$

$$\lim_{n \rightarrow \infty} T_\gamma^n Q_0 \rightarrow Q^*$$

$$T_\gamma Q_* = Q_* = Q^\pi$$

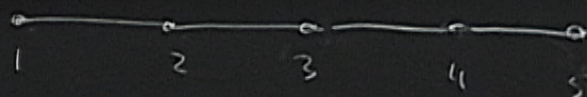
$$\begin{array}{c} \xrightarrow{TQ_2} \\ \xrightarrow{Q_2} \end{array}$$

$$-TQ_2 \parallel_\infty$$

E_+

$$p(s_{t+1} | s_t, a_t) = 1$$

$$r = 1$$



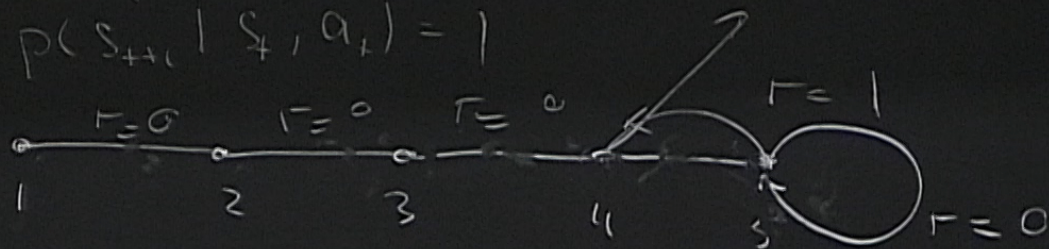
$$Q_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$Q_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}$$

$$Q_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -Y \\ 0 \end{pmatrix}$$

E_+

$$p(s_{t+1} | s_t, a_t) = 1$$



$$Q_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$Q_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$Q_2$$

$\gamma(P, 4, 7)$

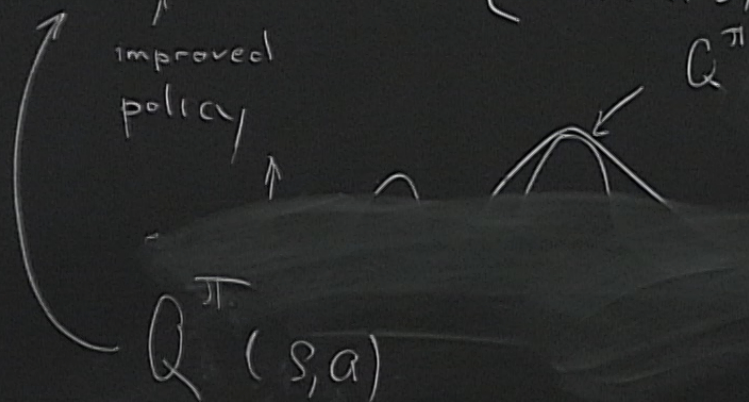
TQ_2

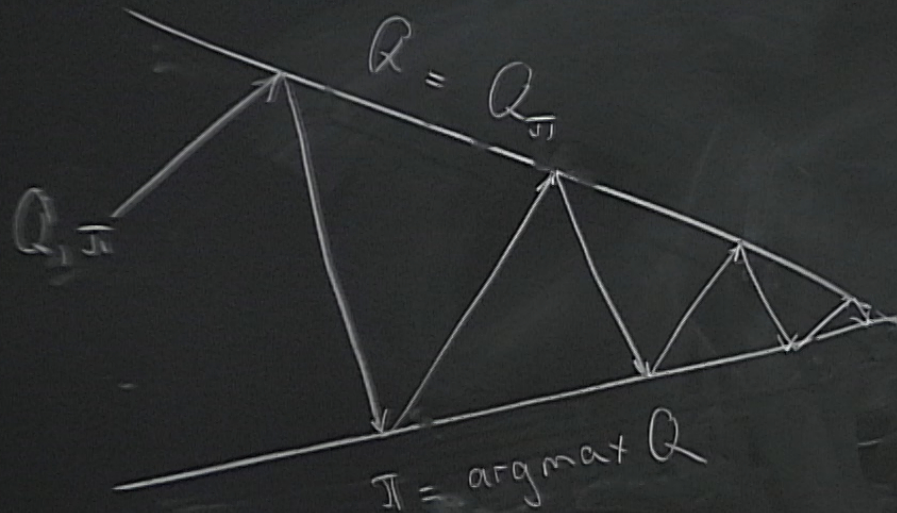
Q_2

TQ_1

$$\lim_{n \rightarrow \infty} T_y^n Q_0 \rightarrow Q^*$$
$$T_y Q_* = Q_* = Q^\pi$$

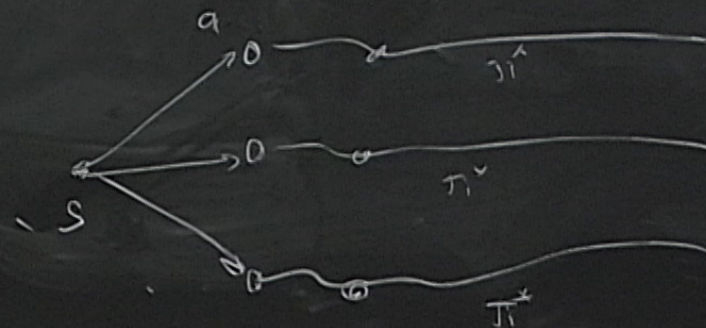
$$\pi'(a|s) = \begin{cases} \operatorname{argmax}_a Q^\pi(s, a) & \text{if } s = s_t \\ \pi(a|s) & \text{if } s \neq s_t \end{cases}$$





π^* Q^*

$$\pi^*(a|s) = \operatorname{argmax}_a Q^*(s, a)$$



$$\bar{Q}(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \sum_{a_{t+1}} \pi(a_{t+1} | s_{t+1}) \bar{Q}^{\pi}(s_{t+1}, a_{t+1})$$

$$\bar{Q}^*(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \max_a \bar{Q}^*(s_{t+1}, a_{t+1})$$

Bellman optimality equations

$$= r(s_{t+1}, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \max_a Q^*(s_{t+1}, a_{t+1})$$

optimality equations

$$Q^*(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \max_a Q^*(s_{t+1}, a_{t+1})$$

$$\sim r(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} | s_t, a_t) \max_a Q^*(s_{t+1}, a_{t+1})$$

Dellman optimality equations

$$\begin{aligned}
 Q_k^*(s_t, a_t) &= r(s_t, a_t) + \gamma \max_a Q_k^*(s_{t+1}, a_{t+1}) \\
 &\sim r(s_t, a_t) + \gamma \max_a Q^*(s_{t+1}, a_{t+1}) \\
 s_{t+1} &\sim p(s_{t+1} | s_t, a_t)
 \end{aligned}$$

$$\begin{aligned}
 & \text{temporal difference} \\
 & \left[r(s_t, a_t) + \gamma \max_a Q_{\pi^*}^v(s_{t+1}, a) - Q_{\pi^*}^v(s_t, a_t) \right]
 \end{aligned}$$

$$+ \gamma \left[\underbrace{r(s_t, a_t) + \gamma \max_a Q_k^\pi(s_{t+1}, a)}_{\text{temporal difference}} - \underbrace{Q_k^\pi(s_t, a_t)}_{Q^\theta} \right]$$

$$Q_{k+1} = \theta_k - \gamma \sum_i x_i \| Q_\theta(s_i, a_i) - y_i \|^2$$

$$Q_{k+1}^{\pi^*}(s_+, a_+) \leftarrow (1-\eta) Q_k^{\pi^*}(s_+, a_+) + \eta \left[\underbrace{r(s_+, a_+)}_{\text{learning rate}} \right]$$

Q-learning

$$Q_{k+1} =$$

learning rate

$$Q_{\pi}^{\pi} \leftarrow (1-\eta) Q_{\pi}^{\pi}(s_+, a_+) + \eta \left[\underbrace{r(s_+, a_+) + \gamma \max_a Q_{\pi}^{\pi}(s_+, a)}_{\text{temporal}} \right]$$

(s_+, a_+, r_+, s_{++})

learning

$Q_{\pi+1} = \theta_{\pi} - \eta \sum_i$

