Reinforcement Learning

STS/H – *Harmonic step sizes.* Show that the step sizes

$$\alpha_n := \frac{1}{an+b}, \qquad a, b \in \mathbb{R},$$

(where $a, b \in \mathbb{R}$ are chosen such that $an+b \neq 0$) satisfy the convergence conditions

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \qquad \sum_{n=1}^{\infty} \alpha_n^2 < \infty.$$

$$a_n = \frac{1}{an+b} = \frac{1}{a} \cdot \underbrace{\frac{1}{n+\frac{b}{a}}}_{=: \tilde{a}_n}$$

da $\frac{1}{a}$ aus den Summen herausgezogen werden kann, reicht es $a = 1$ zu betrachten.

$\Rightarrow$ OBdA $a = 1$, $b \in \mathbb{R}$

$$\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} \frac{1}{n+b}$$

$$= \sum_{n=0}^{\infty} \frac{1}{n + \lfloor b \rfloor + (b - \lfloor b \rfloor)}$$

$b \geq 0$: $= \sum_{n=\lfloor b \rfloor}^{\infty} \frac{1}{n + (b - \lfloor b \rfloor)}$

$\geq \sum_{n=\lfloor b \rfloor}^{\infty} \frac{1}{n+1}$

$= \sum_{n=\lfloor b \rfloor + 1}^{\infty} \frac{1}{n} = \infty$

$b < 0$: $= \sum_{\lfloor b \rfloor}^{0} (\dots) + \underbrace{\sum_{n=0}^{\infty} (\dots)}_{\geq \sum \frac{1}{n}}$

□

$$\sum_{n=0}^{\infty} a_n^2 = \sum_{n=\lfloor b \rfloor}^{\infty} \frac{1}{(n+(b-\lfloor b \rfloor))^2} \quad \leftarrow \text{if } b \in -\mathbb{N} \text{ division by 0 happens}$$

$\Rightarrow b \in \mathbb{R} \setminus (-\mathbb{N})$

$b < 0$: $= \underbrace{\sum_{n=\lfloor b \rfloor}^{0} (\dots)}_{< \infty} + \underbrace{\sum_{n=0}^{\infty} (\dots)}_{\leq \sum \frac{1}{n^2} < \infty}$

$b \geq 0$: $\leq \sum_{n=\lfloor b \rfloor}^{\infty} \frac{1}{n^2} < \infty$

□

STS/U – *Unbiased step sizes.* We use the iteration

$$\begin{aligned} &Q_1 \in \mathbb{R}, \\ &Q_{n+1} := Q_n + \alpha_n(R_n - Q_n), \qquad n \geq 1, \end{aligned}$$

to estimate Q_n using R_n, where

$$\alpha_n := \frac{\alpha}{\beta_n}, \qquad \alpha \in (0,1), \quad n \geq 1,$$

and

$$\begin{aligned} &\beta_0 := 0, \\ &\beta_n := \beta_{n-1} + \alpha(1 - \beta_{n-1}), \qquad n \geq 1. \end{aligned}$$

Show that the iteration for Q_n above yields an exponential recency-weighted average *without initial bias* (i.e., the Q_n do not depend on the initial value Q_1).

$Q_2 = Q_1 + \alpha_1(R_1 - Q_1)$ // $\alpha_1 = \frac{\alpha}{\beta_1} = 1$

$= Q_1 + 1 \cdot (R_1 - Q_1)$

$= R_1 \Rightarrow Q_n$ is independent of $Q_1 \; \forall n \geq 2$

$n = 1$ $\quad Q_1 = Q_1$

$n = 2$ $\quad Q_2 = Q_1 + \alpha_1(R_1 - Q_1) = Q_1(1-\alpha_1) + R_1\alpha_1$

$n = 3$ $\quad Q_3 = Q_2 + \alpha_2(R_2 - Q_2)$

$= Q_1(1-\alpha_1) + R_1\alpha_1 + R_2\alpha_2 - Q_1\alpha_2(1-\alpha_1) - R_1\alpha_1\alpha_2$

$= Q_1(1-\alpha_1)(1-\alpha_2) + R_1\alpha_1(1-\alpha_2) + R_2\alpha_2$

$n = N$ $\quad Q_N = Q_1 \prod_{i=1}^{N-1}(1-\alpha_i) + \sum_{i=1}^{N-1} R_i\alpha_i \prod_{k=i+1}^{N-1}(1-\alpha_k)$

IA: $n=1$ $\quad Q_1 = Q_1 + 0 = Q_1$ ✓

IS: $n \to n+1$ // IV: $Q_n = Q_1 \prod_{i=1}^{n-1}(1-\alpha_i) + \sum_{i=1}^{n-1} R_i\alpha_i \prod_{k=i+1}^{n-1}(1-\alpha_k)$

$Q_{n+1} = Q_n + \alpha_n(R_n - Q_n)$

$= Q_n(1-\alpha_n) + R_n\alpha_n$

$$= Q_1 \underbrace{\left(\prod_{i=1}^{n-1}(1-\alpha_i)\right)(1-\alpha_n)}_{=\prod_{i=1}^{n}(1-\alpha_i)} + \underbrace{\left[\sum_{i=1}^{n-1} R_i \alpha_i \prod_{k=i+1}^{n-1}(1-\alpha_k)\right](1-\alpha_n)}_{=\sum_{i=1}^{n-1} R_i \alpha_i \prod_{k=i+1}^{n}(1-\alpha_k)} + R_n \alpha_n$$

$$= \sum_{i=1}^{n} R_i \alpha_i \prod_{k=i+1}^{n}(1-\alpha_k)$$

$$= Q_1 \prod_{i=1}^{n}(1-\alpha_i) + \sum_{i=1}^{n} R_i \alpha_i \prod_{k=i+1}^{n}(1-\alpha_k)$$

We need to show that $\prod_{i=1}^{n}(1-\alpha_i) \xrightarrow{n\to\infty} 0$

$$\beta_n = 1 - \sum \binom{n}{k} (-1)^{n-k} \alpha^k = 1-(1-\alpha)^n$$

IA: $n=1$ $\quad \beta_1 = 1-(1-\alpha) = \alpha$ ✓

IS: $n \to n+1$ $\quad \beta_{n+1} = \beta_n + \alpha(1-\beta_n)$

$$= \beta_n(1-\alpha) + \alpha$$

$$\overset{IV}{=} (1-(1-\alpha)^{n+1}$$

$$= 1-(1-\alpha)^{n+1}$$ ✓

$$\Rightarrow (1-\alpha_n) = 1 - \frac{\alpha}{1-(1-\alpha)^n} = \frac{1-(1-\alpha)^n-\alpha}{1-(1-\alpha)^n}$$

$$= (1-\alpha)\frac{1-(1-\alpha)^{n-1}}{1-(1-\alpha)^n}$$

$$= (1-\alpha)\frac{\beta_{n-1}}{\beta_n}$$

$\beta_n = 1-(1-\alpha)^n \xrightarrow{n\to\infty} 1 \Rightarrow \frac{\beta_{n-1}}{\beta_n} \to 1$

$\xrightarrow{n\to\infty} (1-\alpha)$

$$\Rightarrow \prod_{i=1}^{n}(1-\alpha_i) \xrightarrow{n\to\infty} 0$$

□

AS – *ϵ-greedy action selection.* Suppose $|\mathcal{A}| = 2$ and $\epsilon = 1/2$. When using ϵ-greedy action selection, what is the probability that the greedy action is selected?

$$\mathbb{P}(\text{greedy action picked}) = \mathbb{P}(\text{greedy choice}) + \mathbb{P}(\text{non-greedy choice AND greedy action picked})$$
$$= 0{,}5 + 0{,}5 \cdot \frac{1}{|\mathcal{A}|} = 0{,}75$$

ENV/EX – *Think of application.* Think of a (preferably creative) application of reinforcement learning. Specify the states, actions, and rewards as well as what is needed to satisfy the Markov property.

A program that design new proteins. The state would be the current sequence of amino acids; The actions would correspond to the possible Amino acids that can be added in the next step; The reward would correspond to the chemical properties of the protein (for example, one could look for antibacterial properties; these could be determined by experiment, from known protens or by simulations. // Alternative Football manager: state: position and state of all players; action: for every player next action; rewards: goals - enemy goals

ENV/COUNTEREX – *Goal-directed learning task that is not an MDP.* Try to find a goal-directed learning task that cannot be represented by a Markov decision process.

A MDP requires, that a description exists such that the current state contains all information of a system. This is not fullfilled for systems where:

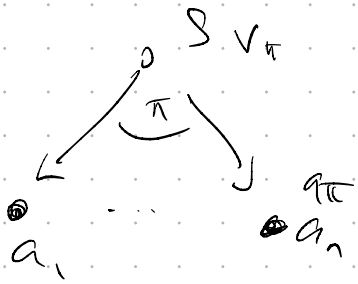
1) The state cannot completely measured because it is infeasible
   (For example a weather prediction method would require precise measurements of temp, humidity etc. at every point in space)
2) The state cannot be measured because it is physically not possible (e.g. a cosmic ray prediction system. EM travels with c as does information)

$$v_\pi(s) = \mathbb{E}_\pi(G_t \mid S_t = s)$$
$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right], \quad \forall s \in \mathcal{S}$$

$$q_\pi(s,a) = \mathbb{E}_\pi(G_t \mid S_t = s, A_t = a)$$
$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right]$$

MDP/V – *Equation for v_π.* Give an equation for v_π in terms of q_π and π.

$$\boxed{\sum_a \pi(a|s)\, q_\pi(s,a)}$$



$$\mathbb{E}(X \mid Y, Z) = \sum_x x\, \mathbb{P}(X \mid Y, Z)$$

$$\left(\mathbb{P}(X,Y) = \mathbb{P}(X|Y)\,\mathbb{P}(Y) \right.$$
$$\left. \Rightarrow \mathbb{P}(X,Y|Z) = \mathbb{P}(X|Y,Z)\,\mathbb{P}(Y|Z) \Rightarrow \mathbb{P}(X|Y,Z) = \frac{\mathbb{P}(X,Y|Z)}{\mathbb{P}(Y|Z)} \right.$$
$$= \frac{1}{\mathbb{P}(Y)} \sum_x x\, \mathbb{P}(X,Y|Z)$$

$$\Rightarrow \sum_Y \mathbb{E}(X|Y,Z)\,\mathbb{P}(Y|Z) = \mathbb{E}(X|Z)$$

In our case $X \hat{=} G_t$; $Y \hat{=} A_t$; $Z \hat{=} S_t$ and the expectation is with respect to π.

$$\Rightarrow v_\pi(s) = \mathbb{E}_\pi(G_t \mid S_t = s)$$
$$= \sum_a \pi(a|s)\, \underbrace{\mathbb{E}_\pi(G_t \mid S_t = s, A_t = a)}_{= q(s,a)}$$
$$\| \quad \mathbb{P}(A_t = a \mid S_t = s)$$

MDP/Q – *Equation for q_π.* Give an equation for q_π in terms of v_π and the four-argument p.

$$p(s', r \mid s, a) = \mathbb{P}(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a) =: p_{sa}(s', a)$$

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi(G_t \mid S_t = s, A_t = a) \\ &= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_\pi(R_{t+k+1} \mid S_t = s, A_t = a) \\ &= \mathbb{E}(R_{t+1} \mid S_t = s, A_t = a) + \gamma \mathbb{E}_\pi(G_{t+1} \mid S_t = s, A_t = a) \\ &= \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a) + \gamma \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} p(s', r \mid s, a) \mathbb{E}_\pi(G_{t+1} \mid S_{t+1} = s') \\ &= \sum_{r, s'} p(s', r \mid s, a)[r + \gamma v_\pi(s')] \end{aligned}$$

MDP/RET – *Change of return.* In episodic tasks and in continuing tasks, how does the return G_t change if a constant c is added to all rewards R_t?

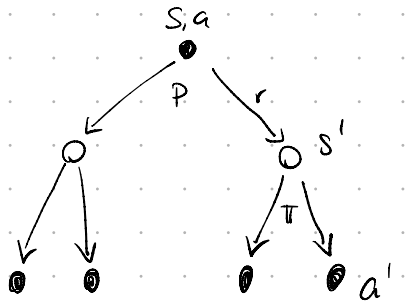
$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$\tilde{R}_t = R_t + c$$

$$\begin{aligned} \Rightarrow \tilde{G}_t &= \sum_{k=0}^{\infty} \gamma^k \tilde{R}_{t+k+1} \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \underbrace{\sum_{k=0}^{\infty} \gamma^k c}_{= \frac{c}{1-\gamma}} \end{aligned}$$

MDP/BELLMAN/QPI – *Bellman equation for q_π.* Analogous to the derivation of the Bellman equation for v_π, derive the Bellman equation for q_π.

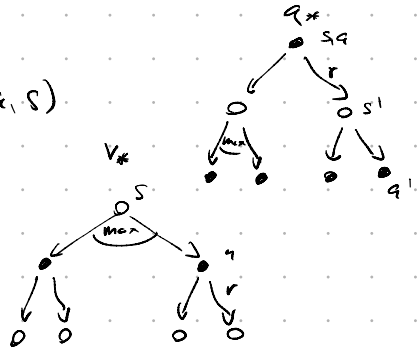
$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi(G_t \mid S_t = s, A_t = a) \\
&= \mathbb{E}_\pi(R_{t+1} + G_{t+1} \mid S_t = s, A_t = a) \\
&= \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)\, r + \gamma \sum_{r,s'} p(s', r \mid s, a) \sum_{a'} \pi(a' \mid s')\, \mathbb{E}_\pi(G_{t+1} \mid S_{t+1} = s', A_{t+1} = a') \\
&= \sum_{r,s'} p(s', r \mid s, a) \left[r + \gamma \sum_{a'} \pi(a' \mid s')\, q_\pi(s', a') \right]
\end{aligned}
$$



MDP/VSTAR – *Equation for v_*.* Give an equation for v_* in terms of q_*.

$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a)$$

$$
\begin{aligned}
v_*(s) = \max_\pi v_\pi(s) &= \max_\pi \sum_a \pi(a \mid s)\, q_\pi(a, s) \\
&= \max_{a \in \mathcal{A}} q_*(a, s)
\end{aligned}
$$



MDP/QSTAR – *Equation for q_*.* Give an equation for q_* in terms of v_* and the four-argument p.

// $p(s', r \mid s, a) = \mathbb{P}(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a)$

$$
\begin{aligned}
q_*(s,a) &= \sum_{k=0}^{\infty} \gamma^k\, \mathbb{E}_{\pi_*}(R_{t+k+1} \mid S_t = s, A_t = a) \\
&= \ldots \quad \text{// see MDP/Q} \\
&= \sum_{r,s'} p(s', r \mid s, a) \left[r + \gamma v_{\pi_*}(s') \right]
\end{aligned}
$$

MDP/PISTAR/VSTAR – *Equation for π_*.* Give an equation for π_* in terms of q_*.

$$\pi_*(a|s) = \begin{cases} 1, & a = \arg\max_a q_{\pi_*}(a,s) \\ 0, & \text{else} \end{cases}$$

↑ There might be multiple values where $q_{\pi_*}(a,s)$ is max → pick one

MDP/PISTAR/QSTAR – *Equation for π_*.* Give an equation for π_* in terms of v_* and the four-argument p.

Combine the last two answers