# Q-learning

Santiago Paternain and Miguel Calvo-Fullana
Electrical and Systems Engineering, University of Pennsylvania
{spater,cfullana}@seas.upenn.edu

November 7 —November 14, 2019

# Outline

- Recall: value function for given policy $\pi$

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

- Goal: Obtain optimal policy which maximizes $v_\pi(s)$

$$v_\star(s) = \max_\pi v_\pi(s) \text{ for all } s \in \mathcal{S}$$

- So far we have been considering policy gradient and variants
- And running gradient ascent, we update the parameters of the policy as

$$\theta_{k+1} = \theta_k + \alpha \nabla v_{\pi_\theta}(\theta_k)$$

- We can only establish convergence to a local maximum

$\Rightarrow$ Can we do better? $\Rightarrow$ At least for tabular cases we can

$\Rightarrow$ Q-learning and Sarsa

- Recall the definition of the q-function

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

- If we are able to compute the optimal $q$-function

$$q_\star(s,a) \doteq \max_\pi q_\pi(s,a)$$

- Optimal value function maximizes over the immediate action

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

- This action is easy to select in tabular cases
- If we consider function approximations only for specific cases

- Bellman equation for $q_\star$:

$$q_\star(s,a) = \mathbb{E}_{S_{t+1},R_{t+1}}[R_{t+1} + \gamma \max_{a'} q_\star(S_{t+1},a')|S_t = s, A_t = a]$$

- Q-learning will find a fixed point of this equation
- Let us prove the result
  - $\Rightarrow$ By definition of $q_\star$ and the $q$-function we have that

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

  - $\Rightarrow$ Expectation is linear and $R_{t+1}$ independent of $\pi$ since $A_t = a$

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

# Proof of Bellman equations for $q_\star(s, a)$

- From the previous slide we had that

$$q_\star(s,a) = \mathbb{E}[R_{t+1}|S_t = s, A_t = a] + \gamma \max_\pi \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1}|S_t = s, A_t = a\right]$$

- We will show that

$$\max_\pi \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1}|S_t = s, A_t = a\right] = \mathbb{E}\left[\max_{a'} q_\star(S_{t+1}, a')\,|S_t = s, A_t = a\right]$$

- That being the case we have that

$$q_\star(s,a) = \mathbb{E}[R_{t+1}|S_t = s, A_t = a] + \gamma \mathbb{E}\left[\max_{a'} q_\star(S_{t+1}, a')\,|S_t = s, A_t = a\right]$$

- Regrouping would complete the proof

$$q_\star(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_\star(S_{t+1}, a')\,|S_t = s, A_t = a\right]$$

- It remains to prove that

$$\max_\pi \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1} \mid S_t = s, A_t = a\right] = \mathbb{E}\left[\max_{a'} q_\star(S_{t+1}, a') \,|S_t = s, A_t = a\right]$$

- We will use that expectation w.r.t. $S_{t+1}$ does not depend on $\pi$
- Notice that the left hand by the towering property is

$$\max_\pi \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1} | S_{t+1}\right] \middle| S_t = s, A_t = a\right]$$

$$= \max_\pi \mathbb{E}\left[v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right] = \mathbb{E}\left[\max_\pi v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right]$$

$$= \mathbb{E}\left[v_\star(S_{t+1}) \middle| S_t = s, A_t = a\right] = \mathbb{E}\left[\max_{a'} q_\star(S_{t+1}, a') | S_t = s, A_t = a\right] \quad \square$$

- Define Bellman operator $F(q) : \mathbb{R}^{N_S \times N_A} \to \mathbb{R}^{N_S \times N_A}$ given by

$$F(q)|_{(s,a)} = \mathbb{E}_{S_{t+1},R_{t+1}}[R_{t+1} + \gamma \max_{a'} q(S_{t+1}, a')|S_t = s, A_t = a]$$

- Bellman equation can be written as $q_\star = F(q_\star)$
- Will prove:
  - $\Rightarrow$ $F(q)$ is a contraction: $\|F(q') - F(q)\| \leq \gamma \|q' - q\|$
  - $\Rightarrow$ Optimal $q_\star$ is the unique fixed point of $F(q)$; i.e., $q_\star = F(q_\star)$
  - $\Rightarrow$ The iteration $q_{n+1} = F(q_n)$ converges to $q_\star$ from any initial point $q_0$

- Think of tabular $q$ as a matrix

| q | $s=1$ | $s=2$ | $\dots$ | $s=N_s$ |
|---|---|---|---|---|
| $a=1$ | 3 | 2 | $\dots$ | 4 |
| $a=2$ | 4 | 1 | $\dots$ | 2 |
| $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ |
| $a=N_s$ | 1 | 2 | $\dots$ | 5 |

- Infinite (maximum) norm $\|q\|_\infty := \max_{s\in S, a\in A} |q(s,a)|$
- Claim: Bellman operator is contractive $\|F(q') - F(q)\|_\infty \leq \gamma \|q' - q\|_\infty$

$$F(q)|_{(s,a)} = \mathbb{E}_{S_{t+1},R_{t+1}}[R_{t+1} + \gamma \max_{a'} q(S_{t+1}, a')|S_t = s, A_t = a]$$

- Define $\Delta F(s,a) = |F(q)|_{s,a} - F(q')|_{s,a}|$, for every $(s,a)$ we have

$$\Delta F(s,a) = \left| \mathbb{E}_{S_{t+1},R_{t+1}} \left[ R_{t+1} + \gamma \max_{a'} q(S_{t+1},a') \Big|_{\substack{S_t=s \\ A_t=a}} \right] \right.$$
$$\left. - \mathbb{E}_{S_{t+1},R_{t+1}} \left[ R_{t+1} + \gamma \max_{a'} q'(S_{t+1},a') \Big|_{\substack{S_t=s \\ A_t=a}} \right] \right|$$

- The expectation of $R_{t+1}$ is independent of the $q$ function

$$\Delta F(s,a) = \gamma \left| \mathbb{E}_{S_{t+1}} \left[ \max_{a'} q(S_{t+1},a') - \max_{a'} q'(S_{t+1},a') \Big|_{\substack{S_t=s \\ A_t=a}} \right] \right|$$

- Using the fact that $|\mathbb{E}X| \leq \mathbb{E}|X|$

$$\Delta F(s,a) \leq \gamma \mathbb{E}_{S_{t+1}} \left[ \left| \max_{a'} q(S_{t+1},a') - \max_{a'} q'(S_{t+1},a') \right| \Big|_{\substack{S_t=s \\ A_t=a}} \right]$$

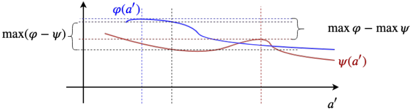
[1] T. Jaakkola, M. I. Jordan, S. P. Singh "On the convergence of Stochastic Iterative Dynamic Programming Algorithms" Neural Computations,, vol. 6, no. 6, pp. 1185-1201, Nov. 1994.

- From the previous slide we have that

$$\Delta F(s,a) \leq \gamma \mathbb{E}_{S_{t+1}} \left[ \left| \max_{a'} q(S_{t+1}, a') - \max_{a'} q'(S_{t+1}, a') \right| \Bigg|_{\substack{S_t = s \\ A_t = a}} \right]$$

- Commutation of maximum and difference operators

$$\left| \max_{a'} \varphi\left(a'\right) - \max_{a'} \psi\left(a'\right) \right| \leq \max_{a'} |\varphi(a') - \psi(a')|$$




$$\begin{aligned} &\varphi(\arg\max \varphi) - \psi(\arg\max \psi) \\ &\leq \varphi(\arg\max \varphi) - \psi(\arg\max \varphi) \\ &\leq \max_{a'} |\varphi(a') - \psi(a')| \end{aligned}$$

- Using the commutation of maximum and difference operators

$$\Delta F(s,a) \leq \gamma \mathbb{E}_{S_{t+1}} \left[ \max_{a'} \left| q(S_{t+1}, a') - q'(S_{t+1}, a') \right| \Bigg|_{\substack{S_t = s \\ A_t = a}} \right]$$

- From the previous slide we have that

$$\Delta F(s,a) \leq \gamma \mathbb{E}_{S_{t+1}}\left[\max_{a'}\left|q(S_{t+1},a') - q'(S_{t+1},a')\right|\Bigg|_{\substack{S_t=s\\A_t=a}}\right]$$

- Maximum and expectation inequality $E_{s'}\left[\varphi(s')\right] \leq \max_{s'}\varphi(s')$

$$\Delta F(s,a) \leq \gamma \max_{a'}\max_{s'}\left|q(s',a') - q'(s',a')\right| = \gamma\|q-q'\|_\infty$$

- Hence by definition of the infinity norm we have that

$$\|F(q) - F(q')\|_\infty = \max_s \max_a \Delta F(s,a) \leq \gamma\|q-q'\|_\infty$$

- We showed that $F(q)$ is contractive $\|F(q) - F(q')\| \leq \gamma \|q - q'\|$
- Optimal $q_\star$ is the unique fixed point of $F(q)$
  - $\Rightarrow$ Assume that $q^\dagger$ is also a fix point $F(q^\dagger) = q^\dagger$

$$\|q^\dagger - q^\star\| = \|F(q^\dagger) - F(q^\star)\| \leq \gamma \|q^\dagger - q^*\| \Rightarrow \|q^\dagger - q^\star\| = 0$$

- Iteration $q_{n+1} = F(q_n)$ converges to $q_\star$ from any initial point $q_0$

$$\|q_{n+1} - q^\star\| = \|F(q_n) - F(q^\star)\| \leq \gamma \|q_n - q^\star\| \leq \gamma^{n+1} \|q_0 - q^\star\| \to 0$$

- We have derived the iterative method $q_{n+1} = F(q_n)$ for obtaining $q^\star$
  - $\Rightarrow$ Requires computing $\mathbb{E}_{S_{t+1}, R_{t+1}}[\cdot]$ at each iteration - unavailable
  - $\Rightarrow$ Idea: use stochastic approximation $\Rightarrow$ q-learning

- Iteration $q_{n+1} = F(q_n)$ converges to $q_\star$ from any initial point $q_0$
- Consider modified version with $\alpha \in (0, 1]$

$$q_{n+1} = q_n + \alpha(F(q_n) - q_n)$$

- Notice that $q_{n+1} = F(q_n)$ is just the same algorithm $\alpha = 1$
- Smaller step-sizes are useful in stochastic versions to reduce noise
- New algorithm also converges to optimal $q^\star$
- Let us prove this claim

- We are analyzing the following algorithm with $\alpha \in (0, 1]$

$$q_{n+1} = q_n + \alpha(F(q_n) - q_n)$$

- Let us look at the difference $\|q_{n+1} - q^\star\|_\infty$

$$\begin{aligned}\|q_{n+1} - q^\star\|_\infty &= \|q_n + \alpha(F(q_n) - q_n) - q_\star\|_\infty \\ &= \|(1-\alpha)(q_n - q_\star) + \alpha(F(q_n) - q^\star)\|_\infty\end{aligned}$$

- Using the triangle inequality we have that

$$\begin{aligned}\|q_{n+1} - q^\star\|_\infty &\leq (1-\alpha)\|q_n - q_\star\|_\infty + \alpha\|F(q_n) - q_\star\|_\infty \\ &= (1-\alpha)\|q_n - q_\star\| + \alpha\|F(q_n) - F(q_\star)\|_\infty \\ &\leq (1-\alpha)\|q_n - q_\star\| + \alpha\gamma\|q_n - q_\star\|_\infty\end{aligned}$$

- We have used that $F(q)$ is contractive and $q^\star$ is a fixed point of $F(q)$

$$\|q_{n+1} - q^\star\|_\infty = (1-\alpha+\alpha\gamma)\|q_n - q_\star\|_\infty \quad \leq (1-\alpha+\alpha\gamma)^{n+1}\|q_0 - q_\star\|_\infty$$

- Error converges to zero since $\gamma < 1 \;\Rightarrow\; 1 - \alpha + \alpha\gamma < 1$

- Given a deterministic algorithm

$$q_{t+1} = q_t + \alpha \mathbb{E}_w[\varphi(q_t, w)]$$

- Drop expectation and run

$$q_{t+1} = q_t + \alpha_t \varphi(q_t, w_t)$$

- Then $q_t \to q^\star$ such that $\mathbb{E}_w[\varphi(q_\star, w)] = 0$ for square-summable $\alpha_t$
- For q-learning we had that that

$$q_{t+1} = q_t + \alpha \left( \mathbb{E}_{S_{t+1}, R_{t+1}}[R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(s, a) | S_t = s, A_t = a] \right)$$

- Consider matrix valued function

$$\varphi(q_t, S_{t+1}, R_{t+1})|_{s,a} := R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(s, a)$$

with $w_t = (S_{t+1}, R_{t+1})$ distributed by $p(S_{t+1}, R_{t+1} | S_t = s, A_t = a)$

- By Bellman's equation we have that $q^\star$ is the argument that satisfies

$$\mathbb{E}_{S_{t+1},R_{t+1}}\left[\varphi(q, S_{t+1}, R_{t+1})\right] = 0$$

- Stochastic q-iteration: for all $(s,a) \in N_S \times N_A$

$$q_{t+1}(s,a) = q_t(s,a) + \alpha_t\big(R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(s,a)\big)$$

- Converges to $q^\star$ with probability one
- Need to update all entries of matrix $q$
- For an online implementation $\Rightarrow$ asynchronos stochastic approximation

- Idea: update only one entry of $q$ at a time
  - $\Rightarrow$ select entry $q(a, s)$ and update it according to

$$q_{t+1}(s, a) = q_t(s, a) + \alpha_t(R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(s, a))$$

- All other entries remain unchanged, i.e.,

$$q_{t+1}(\bar{s}, \bar{a}) = \begin{cases} q_t(s, a) + \alpha_t(R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(s, a)) & \text{if } \bar{a} = a, \bar{s} = s \\ q_t(\bar{s}, \bar{a}) & \text{otherwise} \end{cases}$$

- Asynchronous updates allow for an online implementation
  - $\Rightarrow$ Given matrix $q_t$ and pair $(s, a) = (S_t, A_t)$ sampled from policy $b_t(s, a)$

$$q_t, S_t, A_t \xrightarrow{\text{SYSTEM}} R_{t+1}, S_{t+1} \xrightarrow{\text{Q-LEARNING}} q_{t+1}(s, a)$$

**Input:** Behavior policies $b_t(s,a)$, and tabular $q_0(s,a)$
**Initialize:** $q(s,a) = q_0(s,a)$, for all $s \in \mathcal{S},\ a \in \mathcal{A}$
**for** *time* $t = 0, 1, 2, \ldots$ **do**
  Draw $(s,a) = (S_t, A_t) \sim b_t(a,s)$
  Run system one step ahead $\Rightarrow R_{t+1},\ S_{t+1}$
  Update $q(s,a) \leftarrow q(s,a) + \alpha_t(R_{t+1} + \gamma \max_{a'} q(S_{t+1}, a') - q(s,a))$
**end**
**Output:** $q(s,a)$

**Algorithm 1:** Q-LEARNING

- Particular case: $\epsilon-$ greedy Q-learning
  - $\Rightarrow$ Uses $S_t$ as given by system in previous step
  - $\Rightarrow$ Selects $A_t$ by maximizing current $q_t(S_t, a)$
  - $\Rightarrow$ Explores $A_t \sim \mathrm{rand}(\mathcal{A})$ with probability $\epsilon$

**Input:** State $s_0$, probability $\epsilon$, and tabular $q_0(s, a)$
**Initialize:** $S_0 = s0$ and $q(s, a) = q_0(s, a)$, for all $s \in \mathcal{S}, \ a \in \mathcal{A}$
**for** *time* $t = 0, 1, 2, \ldots$ **do**
  Set $S_t$ from previous step
  Draw $A_t = \begin{cases} \arg\max_{a'} q(S_t, a') & \text{w.p. } 1-\epsilon \\ \mathrm{rand}(\mathcal{A}) & \text{w.p. } \epsilon \end{cases}$
  Run system one step ahead $\Rightarrow R_{t+1}, \ S_{t+1}$
  Update $q(S_t, A_t) \leftarrow q(S_t, A_t) + \alpha_t (R_t + \gamma \max_{a'} q(S_{t+1}, a') - q(S_t, A_t))$
**end**
**Output:** $q(s, a)$

**Algorithm 2:** Q-LEARNING ($\epsilon-$ greedy)

- Recall q-learning algorithm

$$q_{t+1}(s,a) = q_t(s,a) + \alpha_t\left(R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(s,a)\right)$$

$\Rightarrow$ with $S_{t+1}, R_{t+1}$ drawn from $p(R_{t+1}, S_{t+1} | S_t = s, A_t = a)$

- Derived as stochastic algorithm for the Bellman operator

$$F(q)|_{(s,a)} = \mathbb{E}_{S_{t+1},R_{t+1}}\left[R_{t+1} + \gamma \max_{a'} q(S_{t+1}, a') | S_t = s, A_t = a\right]$$

$\Rightarrow$ Write q-learning in terms of $F(q)$ $\Rightarrow$ Add and subtract $F(q)$

$$q_{t+1}(s,a) = q_t(s,a) + \alpha_t\left(F(q_t)|_{(s,a)} + w_t(s,a) - q_t(s,a)\right)$$

$\Rightarrow$ where the unbiased noise term $w_t$ is defined by

$$w_t(s,a) = R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - F(q_t)|_{(s,a)}$$

- Consider reshaping $q$ into vector $\bar{q} \in \mathbb{R}^{N_S N_A}$
- Rewrite $q - learning$ as

$$\bar{q}_{t+1} = \bar{q}_t + D_t(\bar{F}(\bar{q}_t) + \bar{w}_t - \bar{q}_t)$$

$\Rightarrow$ Bellman's $\bar{F}(\bar{q}_t)$ and noise $\bar{w}_t$ are reshaped versions of $F(q_t)$ and $w_t$

$\Rightarrow$ $D_t$ diagonal with diagonal $\bar{\alpha}_t$ $\quad \rightarrow \quad$ $D_t = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{\alpha}_t(i) & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

$\Rightarrow$ Only one nonzero entry of $\bar{\alpha}_t$ per $t$ $\Rightarrow$ entry of $\bar{q}$ to be updated

Theorem (Tsitsiklis'94)

*Let* $D_n = diag(\bar{\alpha}_t)$ *and* $\bar{q}_t$ *be defined by*

$$\bar{q}_{t+1} = \bar{q}_t + D_n(F(\bar{q}_t) + \bar{w}_n - \bar{q}_t)$$

*Under the following assumptions*

as1) $\bar{\alpha}_t(i) \geq 0 \;:\; \sum_t \bar{\alpha}_t(i) = \infty$ *and* $\sum_t \bar{\alpha}_t^2(i) < \infty$ *(as) for all* $i$

as2) $\mathbb{E}[\bar{w}_t | S_t, A_t] = 0$ *and* $\mathbb{E}[\bar{w}_t^2 | S_t, A_t] \leq A + B \max_i |\bar{q}_t(i)|^2$ *for all* $i, t$

as3) $\|\bar{F}(\bar{q}) - \bar{F}(\bar{q}')\| \leq \gamma \|\bar{q} - \bar{q}'\|$ *with* $\bar{F}(\bar{q}_\star) = \bar{q}_\star$,

*then for any initial point* $\bar{q}_0$, $\bar{q}_n \to \bar{q}_\star$ *(as)* .

- Subtract $\bar{q}_\star$ from both sides of the update

$$\bar{q}_{t+1} - \bar{q}_\star = \bar{q}_t - \bar{q}_\star + D_t(F(\bar{q}_t) - (\bar{q}_t - \bar{q}_\star) - \bar{q}_\star + \bar{w}_t)$$

- Define error $\tilde{q}_t = \bar{q}_t - \bar{q}_\star$

$$\tilde{q}_{t+1} = \tilde{q}_t + D_t(F(\tilde{q}_t + \bar{q}_\star) - \tilde{q}_t - \bar{q}_\star + \bar{w}_t)$$

- Define $\tilde{F}(\tilde{q}_t) = \bar{F}(\tilde{q}_t + \bar{q}_\star) - \bar{q}_\star$, and $\tilde{w}_t = \bar{w}_t$

$$\tilde{q}_{t+1} = \tilde{q}_t + D_t(\tilde{F}(\tilde{q}_t) - \tilde{q}_t + \tilde{w}_t)$$

- Fixed point at the origin $\tilde{F}(0) = \bar{F}(\bar{q}_\star) - \bar{q}_\star = 0$
- Contraction $\|\tilde{F}(\tilde{q}) - \tilde{F}(\tilde{q}')\| = \|\bar{F}(\tilde{q} + \bar{q}_\star) - \bar{F}(\tilde{q}' + \bar{q}_\star)\|$
  $\leq \gamma\|(\tilde{q} + \bar{q}_\star) - (\tilde{q}' + \bar{q}_\star)\| = \gamma\|\tilde{q} - \tilde{q}'\|$
- Assumptions as1)-as5) are satisfied by $\tilde{q}_t$ and $\tilde{F}(\cdot)$ with $\tilde{q}_\star = 0$
- Proving convergence of q-learning amounts to show that $\tilde{q}_t \to 0$

- For notation brevity will drop the tildes, keeping $F(0) = 0$ and

$$q_{t+1} = q_t + D_t(F(q_t) + w_t - q_t)$$

Lemma (Tsitsiklis'94)

*Under assumptions as1)-as3) the sequence $q_t$ is bounded by a time invariant random variable $M_0$ with probability one.*

- Proof: Let be $\epsilon = 1/\gamma - 1$ and $m_t = \max_{\tau \leq t}\{\|q_\tau\|\}$
  - $\Rightarrow$ Define $G_t$ such that $G_0 = m_0$ and

$$G_{t+1} = \begin{cases} G_t & \text{if } m_{t+1} \leq (1+\epsilon)G_t \\ m_{t+1} & \text{otherwise} \end{cases}$$

  - $\Rightarrow$ Can select infinite $t_0$ such that $G_{t_0} = m_{t_0}$ otherwise $m_t$ is bounded
  - $\Rightarrow$ Introduce $W_i(t_0, t)$ initialized at $W_i(t_0, t_0) = 0$ and defined by

$$W_i(t_0, t+1) = (1 - \alpha_t(i))W_i(t_0, t) + \alpha_t(i)w_t(i)/G_t$$

  - $\Rightarrow$ We can show that $W_i(t_0, t) \to 0$ as $t_0 \to \infty$ with $t \geq t_0$
  - $\Rightarrow$ $\exists$ $t_0$ such that $|W_i(t_0, t)| \leq \epsilon$ $\forall i$ and $t \geq t_0$

- We want to prove that for all $t \geq t_0$ $G_t = G_{t_0}$ and

$$-G_{t_0}(1+\epsilon) < -G_{t_0} + G_{t_0} W_i(t_0, t) \leq q_t(i) \leq G_{t_0} + G_{t_0} W_i(t_0, t) < G_{t_0}(1+\epsilon)$$

- True for $t = t_0$ since $\|q_{t_0}\| \leq m_{t_0} = G_{t_0}$ and $W_i(t_0, t_0) = 0$
- Assume that it holds for time $t$ let us show for time $t+1$

$$\begin{aligned} q_{t+1}(i) &= (1-\alpha_t(i))q_t(i) + \alpha_t(i)(F_i(q_t) + w_t(i)) \\ &\leq (1-\alpha_t(i))(G_{t_0} + G_{t_0} W_i(t_0, t)) + \alpha_t(i)\left(\gamma\|q_t\| + G_t w_t(i)/G_t\right) \end{aligned}$$

- Where we have use that $F(q_t)$ is a contraction and that $G_t = G_{t_0}$

$$q_{t+1}(i) \leq (1-\alpha_t(i))(G_{t_0} + G_{t_0} W_i(t_0, t)) + \alpha_t(i)\left(\gamma(1+\epsilon)G_{t_0} + G_{t_0} w_t(i)/G_t\right)$$

- Using that $\epsilon = 1/\gamma - 1$ it follows that

$$q_{t+1}(i) \leq G_{t_0} + G_{t_0}\left((1-\alpha_t(i))W_i(t_0, t) + \alpha_t(i) w_t(i)/G_t\right) = G_{t_0}(1 + W_i(t_0, t+1))$$

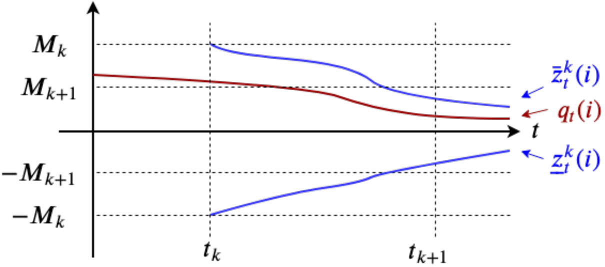
- Since $|W_i(t_0, t+1)| \leq \epsilon$ (We showed it in the previous slide) it follows

$$\|q_{t+1}\| < G_{t_0}(1+\epsilon) = G_t(1+\epsilon) \Rightarrow G_{t+1} = G_t = G_{t_0} \quad \square$$

- Idea: construct bounding sequences $\bar{z}_t^k(i)$ and $\underline{z}_t^k(i)$ such that

$$\underline{z}_t^k(t) \leq q_t(i) \leq \bar{z}_t^k(i)$$

- Initialize $\bar{z}_0^0(i) = M_0$ and $\underline{z}_0^0(i) = -M_0$ with $M_0$ from the Lemma
- Select $\epsilon > 0$ such that $\gamma(1+\epsilon) < 1$ and $M_{k+1} = M_k\gamma(1+\epsilon)$
- Prove that eventually $\|\bar{z}_t^k(i)\| \leq M_{k+1}$ (same for $\underline{z}_t^k(i)$)
- Conclude that $q_t(i) \to 0$



# Bounding sequences

- Define the bounding sequences by iteration starting at $t_k$

$$\bar{z}^k_{t+1}(i) = \bar{z}^k_t(i) + \alpha_t(i)\left(\gamma M_k + w_t(i) - \bar{z}^k_t(i)\right), \quad \bar{z}^k_{t_k}(i) = M_k$$
$$\underline{z}^k_{t+1}(i) = \underline{z}^k_t(i) + \alpha_t(i)\left(-\gamma M_k + w_t(i) - \underline{z}^k_t(i)\right), \quad \underline{z}^k_{t_k}(i) = -M_k$$

- Will prove

$$\underline{z}^k_t(t) \leq q_t(i) \leq \bar{z}^k_t(i)$$

- Inductive hypothesis $\|q_t\| \leq M_k$, for $t \geq t_k$
  - $\Rightarrow$ The previous Lemma established that this is true for $k = 0$
  - $\Rightarrow$ So we just need to prove the inductive step

- Let us prove $q_{t+1}(i) \leq \overline{z}^k_{t+1}(i)$

$$q_{t+1}(i) = q_t(i) + \alpha_t(i)\left(F_i(q_t) + w_t(i) - q_t(i)\right)$$

- Add and subtract $\overline{z}^k_{t+1}(i) = \overline{z}^k_t(i) + \alpha_t(i)\left(\gamma M_k + w_t(i) - \overline{z}^k_t(i)\right)$

$$\begin{aligned} q_{t+1}(i) = {} & q_t(i) + \alpha_t(i)\left(F_i(q_t) + w_t(i) - q_t(i)\right) \\ & + \overline{z}^k_{t+1}(i) - \overline{z}^k_t(i) - \alpha_t(i)\left(\gamma M_k + w_t(i) - \overline{z}^k_t(i)\right) \end{aligned}$$

- Rearrange the terms and note that $w_t(i)$ cancels

$$q_{t+1}(i) = \overline{z}^k_{t+1}(i) + (1 - \alpha_t(i))\left(q_t(i) - \overline{z}^k_t(i)\right) + \alpha_t(i)\left(F_i(q_t) - \gamma M_k\right)$$

- Using that $F_i(q_t) \leq \gamma \|q_t\| \leq M_k$ and that $q_t(i) \leq \overline{z}^k_t(i)$

$$q_{t+1}(i) \leq \overline{z}^k_{t+1}(i)$$

- The proof for $\underline{z}^k_{t+1}(i) \leq q_{t+1}(i)$ is analogous

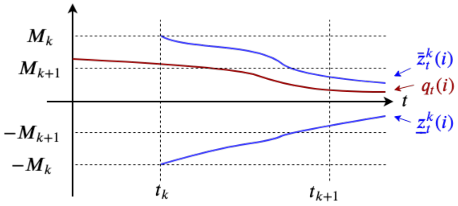
- In next slide we prove that $\bar{z}_t^k(i) \to \gamma M_k$
- As a consequence we can establish that for any $\epsilon > 0$

$$\exists \bar{t}_{k+1} : \forall t \geq \bar{t}_{k+1} \Rightarrow \bar{z}_t^k(i) \leq (1+\epsilon)\gamma M_k = M_{k+1}$$

- Correspondingly $\exists \underline{t}_{k+1} : \forall t \geq \underline{t}_{k+1} \Rightarrow \underline{z}_t^k(i) \geq -M_{k+1}$
- In particular, let's select $\epsilon$ such that $(1+\epsilon)\gamma < 1$
- Recall

$$\underline{z}_t^k(i) \leq q_t(i) \leq \bar{z}_t^k(i)$$

- Define $t_{k+1} = \max\{\bar{t}_{k+1}, \underline{t}_{k+1}\}$ so that $|q_t(i)| \leq M_{k+1}$ for $t \geq t_{k+1}$




- Since $M_k = M_0\gamma^k(1+\epsilon)^k \to 0 \Rightarrow |q_t(i)| \to 0$ $\square$

- It remains to prove that $z_t \to \gamma M$ ($i$ and $k$ dropped)

$$z_{t+1} = z_t + \alpha_t \left( \gamma M + w_t - z_t \right), \quad z(0) = M$$

- Define sequence $S_t = (z_t - \gamma M)^2 + C \sum_{t=\tau}^{\infty} \alpha_\tau^2$ with $C \in \mathbb{R}$ arbitrary
- Want to prove that $S_t$ is a super martingale

$$S_{t+1} = (z_{t+1} - \gamma M)^2 + C \sum_{\tau=t+1}^{\infty} \alpha_\tau^2$$

$$= (z_t + \alpha_t(\gamma M + w_t - z_t) - \gamma M)^2 + C \sum_{\tau=t+1}^{\infty} \alpha_\tau^2$$

- Define $y_t = z_t - \gamma M$

$$S_{t+1} = (y_t + \alpha_t(w_t - y_t))^2 + C \sum_{\tau=t+1}^{\infty} \alpha_\tau^2$$

$$= (1 - \alpha_t)^2 y_t^2 + \alpha_t^2 w_t^2 + 2(1 - \alpha_t y_t) w_t \alpha_t + C \sum_{\tau=t+1}^{\infty} \alpha_\tau^2$$

- From the previous slide we have that

$$S_{t+1} \leq (1-\alpha_t)^2 y_t^2 + \alpha_t^2 w_t^2 + 2(1-\alpha_t y_t) w_t \alpha_t + C \sum_{\tau=t+1}^{\infty} \alpha_\tau^2$$

- Recall Assumption 2 $\mathbb{E}[w_t]$ and $\mathbb{E}[w_t^2] \leq A + B \max_i |q_t(i)|$

$$\mathbb{E}[S_{t+1}|\mathcal{F}_t] = (1-\alpha_t)^2 y_t^2 + \alpha_t^2 \mathbb{E}[w_t^2|\mathcal{F}_t] + C \sum_{\tau=t+1}^{\infty} \alpha_\tau^2$$
$$\leq (1-\alpha_t)^2 y_t^2 + C\alpha_t^2 + C \sum_{\tau=t+1}^{\infty} \alpha_\tau^2$$

- Where we can chose $C = A + BM_0$ with $M_0$ the bound of the Lemma

$$\mathbb{E}[S_{t+1}|\mathcal{F}_t] \leq (1-\alpha_t)^2 y_t^2 + C \sum_{\tau=t}^{\infty} \alpha_\tau^2 < S_t$$

- We established that $S_t$ is a super martingale $\Rightarrow S_t \to S$ with $\mathbb{E}[S] < \mathbb{E}[S_0]$

- Will show $S = 0 \Rightarrow \; y_t = S_t - C\sum_{\tau=t}^{\infty} \alpha_t^2 \to 0$
- If $\alpha_t < 1$ then we have that $\mathbb{E}[S_{t+1}|\mathcal{F}_t] \leq S_t - 2\alpha_t y_t^2$

$$\mathbb{E}[S_{t+1}|\mathcal{F}_{t-1}] \leq \mathbb{E}\left[S_t - 2\alpha_t y_t^2|\mathcal{F}_{t-1}\right] \leq S_{t-1} - 2\alpha_{t-1}y_{t-1}^2 - 2\mathbb{E}\left[\alpha_t y_t^2|\mathcal{F}_{t-1}\right]$$

$\Rightarrow$ Recursively we have that $\mathbb{E}[S_{t+1}] \leq \mathbb{E}\left[S_0\right] - 2\sum_{\tau=0}^{t}\mathbb{E}\left[\alpha_\tau y_\tau^2\right]$

$\Rightarrow$ Rearranging terms

$$2\sum_{\tau=0}^{t}\mathbb{E}\left[\alpha_\tau y_\tau^2\right] \leq \mathbb{E}[S_0] - \mathbb{E}[S_{t+1}] \to \mathbb{E}[S_0] - \mathbb{E}[S] \leq \infty$$

$\Rightarrow$ By (as1) $\sum_t \alpha_t = \infty$ w.p.1 $\Rightarrow \liminf_t E\left[y_t^2\right] \to 0$

$\Rightarrow$ Fatou's $\Rightarrow E[\liminf_t y_t^2] \leq \liminf_t E[y_t^2] = 0 \;\Rightarrow\; \liminf_t y_t^2 = 0$

$\Rightarrow$ Subsequence $y_{t_j}^2 \to 0 \;\Rightarrow S_{t_j} = y_{t_j} + C\sum_{\tau=t_j}^{\infty}\alpha_\tau^2 \to 0 \;\Rightarrow\; S = 0$ $\square$

# Outline

# Off-policy v.s. On-policy

- So far we have studied $q$-learning as an Off-policy algorithm

$$q_{t+1}(s,a) = q_t(s,a) + \alpha_t(R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(s,a))$$

- Where the exploration can be done with any behavior policy

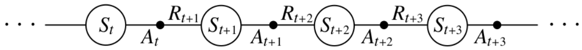
  $\Rightarrow$ Once we have reached $S_{t+1}$ any selection of the action guarantees convergence to $q^\star$

  $\Rightarrow$ As long as we explore enough

  $\Rightarrow$ In the end we want to always select the action according to

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} q_\star(S_t, a)$$

- We can also do on Policy control $\Rightarrow$ SARSA



- The policy being learned about is called the *target* policy $\pi$
- The policy used to generate behavior is called the *behavior* policy $b$
- In Q-learning the update is for the greedy *target* policy
- In SARSA, the target and behavior policy are the same
  $\Rightarrow$ typically the $\varepsilon$-greedy policy
- By making $\varepsilon \to 0$, SARSA approximates the optimal policy $q^\star$

**Input:** State $S_0$, probability $\epsilon \in (0,1)$, and tabular $q_0(s,a)$
**Initialize:** $q(s,s) = q_0(s,a)$, for all $s \in \mathcal{S},\ a \in \mathcal{A}$
Draw $A_0$ randomly according to policy.
**for** *time* $t = 0, 1, 2, \ldots$ **do**

Run system one step ahead $(S_t, A_t) \Rightarrow R_{t+1},\ S_{t+1}$

Draw $A_{t+1} = \begin{cases} \arg\max_{a'} q(S_{t+1}, a') & \text{w.p. } 1-\epsilon \\ \mathrm{rand}(\mathcal{A}) & \text{w.p. } \epsilon \end{cases}$

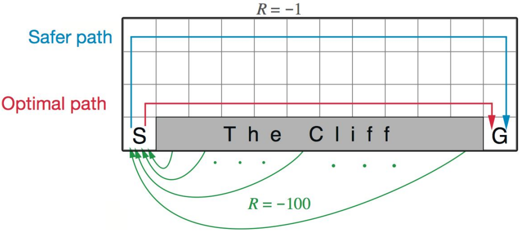
Update $q(s_t, A_t) \leftarrow q(S_t, A_t) + \alpha_t(R_{t+1} + \gamma q(S_{t+1}, A_{t+1}) - q(S_t, A_t))$

**end**
**Output:** $q(s,a)$

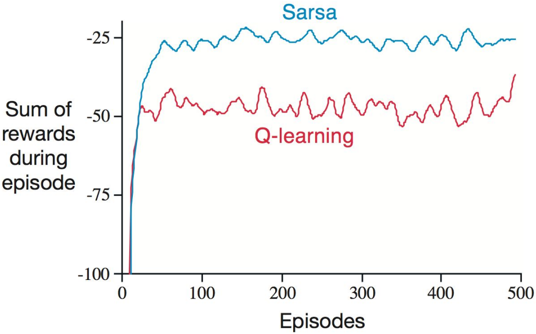
**Algorithm 3:** SARSA

- Recall that $q$-learning udpates using

$$q(s_t, A_t) \leftarrow q(S_t, A_t) + \alpha_t(R_{t+1} + \gamma \max_{a \in \mathcal{A}} q(S_{t+1}, a) - q(S_t, A_t))$$



$R = -1$
Safer path
Optimal path
S
The Cliff
G
$R = -100$

Sarsa
Q-learning
Sum of rewards during episode
-25
-50
-75
-100
0
100
200
300
400
500
Episodes

Sketch of proof for one-step tabular SARSA. (Recall Q-learning assumptions)

Theorem

*Let $q_n$ be defined by*

$$q_{n+1} = q_n + A_n(F(q_n) + w_n - q_n)$$

*with $A_n = diag(\alpha_n)$.*

as1) *Stepsize* $\sum_t \alpha_t(i) = \infty$ *and* $\sum_t \alpha_t^2(i) < \infty$ *(as) for all* $i$ *(GLIE)*

as2) $E[w_t|S_t, A_t] = 0$ $E[w_t^2|S_t, A_t] \leq A + B \max |q_t(s,a)|^2$ *for all* $s, a$

as3) *"Asymptotic contraction"* $\|F(q) - q_\star\| \leq \gamma \|q - q_\star\| + c_n$, *with* $c_n \to 0$

*then for any initial point* $q_0$, $q_n \to q_\star$ *(as) .*

*Proof in Singh et al. "Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms", 2000.*

- In Q-learning, given $s, a$,

$$F^Q(q_t) = \mathbb{E}_{R_t, S_{t+1}}\left[R_t + \gamma \max_{a'} q_t(S_{t+1}, a')\right]$$

- In SARSA, given $s, a, \pi$

$$\begin{aligned} F(q_t) &= \mathbb{E}\left[R_t + \gamma q_t(S_{t+1}, A_{t+1})\right] \\ &= \mathbb{E}\left[R_t + \gamma \max_{a'} q_t(S_{t+1}, a') + \gamma q_t(S_{t+1}, A_{t+1}) - \gamma \max_{a'} q_t(S_{t+1}, a')\right] \\ &= F^Q(q_t) + C_t \end{aligned}$$

with $C_t = \mathbb{E}\left[\gamma q_t(S_{t+1}, A_{t+1}) \ - \gamma \max_{a'} q_t(S_{t+1}, a')\right]$

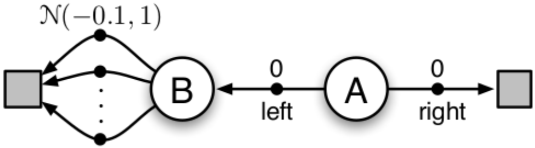
- Then,

$$\begin{aligned} \|F(q_t) - q_\star\| = \|F^Q(q_t) + C_t - q_\star\| &\leq \|F^Q(q_t) - q_\star\| + \|C_t\| \\ &\leq \gamma\|q_t - q_\star\| + \|C_t\| \end{aligned}$$

- Finally, because the behaviour policy $\pi$ used in SARSA tends to the greedy policy (GLIE assumption), we have that $\|C_t\| \to 0$.

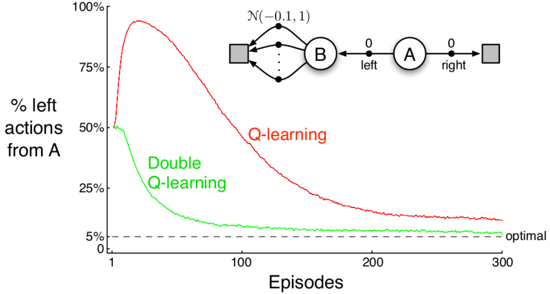
- Off-Policy gives us the optimal solution if the system is an MDP
- On-Policy is more robust to modeling error
- Off-Policy might lead to high variance estimates due to maximization
- Variance is easier to reduce the variance of the estimate On-Policy

- Consider the following MDP




- Always start on `A` $\Rightarrow q(\mathtt{A}, \mathtt{right}) = 0$
- The estimate of $q(\mathtt{A}, \mathtt{left})$ may be positive if we received positive rewards
- But from the MDP is clear that the optimal action is choosing `right`
- With $0.1 = \epsilon$-greedy `left` should be selected only 5% of the time

- With $\epsilon = 0.1$ `left` should be selected only 5% of the time




- The selection of the maximum bias the policy to choose the action `left`
- What is this double $Q$-learning that gets rid of this bias?

$\Rightarrow$ It keeps two estimates of $Q$ and with probability 0.5 updates

$$q_1(S_t, A_t) = q_1(S_t, A_t) + \alpha(R_{t+1} + \gamma q_2(S_{t+1}, \operatorname*{argmax}_a q_1(S_{t+1}, a)) - q_1(S_t, A_t))$$

- As we did before define the *n*-step return as

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} \gamma^{n} + q_{t+n-1}(S_{t+n}, A_{t+n})$$

- Then it is easy to extend SARSA to *n* step by updating

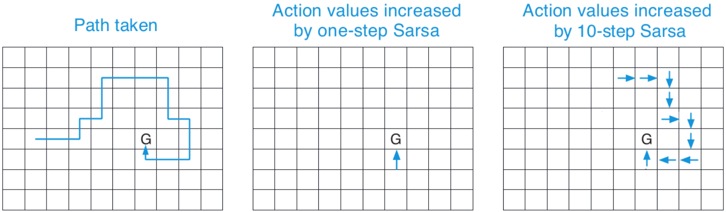
$$q_{t+n}(S_t, A_t) = q_{t+n-1}(S_t, A_t) + \alpha(G_{t:t+n} - q_{t+n-1}(S_t, A_t))$$

- It converges because the update is based on the *n*-step Bellman equation

  $\Rightarrow$ It is an assymptotic contraction

**Input:** Policy $\pi$ to be $\epsilon$-greedy, step-size $\alpha$, positive integer $n$
**Initialize:** $q(s,a) = 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$
**for** *episode* $k = 0, 1, 2, \ldots$ **do**
  Initialize and store $S_0 \neq \texttt{terminal}$ and $A_0 \sim \pi(\cdot|S_0)$ , $T = \infty$ **for** *each step of the episode* $t = 0, 1, \ldots, T$ **do**
    **if** $t < T$ **then**
      Take action $A_t$, observe $R_{t+1}$ and state $S_{t+1}$
      **if** $S_{t+1}$ *is terminal* **then**
        $T = t + 1$
      **end**
      **else**
        Select and store action $A_{t+1} \sim \pi(\cdot|S_{t+1})$
      **end**
      $\tau = t - n + 1$ ▷ (time whose state's estimate is being updated)
    **end**
    **if** $\tau \geq 0$ **then**
      $G = \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$ ▷ (Compute return)
      $q(S_\tau, A_\tau) =$
        $q(S_\tau, A_\tau) + \alpha \left[G + \gamma^n q(S_{\tau+n}, A_{\tau+n}) \mathbb{1}(\tau + n < T) - q(S_\tau, A_\tau)\right]$
    **end**
  **end**
**end**

- All transitions have reward zero except reaching the goal
- Initialization is $q(s,a) = 0$ for all $s, a$




- $n$-step SARSA can update more entries of the $q$ function

$$q(S_\tau, A_\tau) = q(S_\tau, A_\tau) + \alpha \left[G_{\tau:\tau+n} + \gamma^n q(S_{\tau+n}, A_{\tau+n}) - q(S_\tau, A_\tau)\right]$$