Q-Learning

John Stachurski

Jan 2023

Overview

- Q-factors
- Fixed point iteration
- Stochastic approximation
- Q-learning as stochastic approximation

Q-factors

Consider an MDP with Bellman equation

$$v^*(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \sum_{x'} v^*(x') P(x,a,x') \right\}$$

The corresponding **Q-factor** is the right-hand side

$$q^*(x,a) = r(x,a) + \beta \sum_{x'} v^*(x') P(x,a,x')$$

Hence

$$v^*(x) = \max_{a \in \Gamma(x)} q^*(x,a)$$

Combining the last two equations gives

$$q^*(x,a) = r(x,a) + \beta \sum_{x'} \max_{a' \in \Gamma(x')} q^*(x',a') P(x,a,x')$$

We can use this to solve for q^* and the obtain v^* via

$$v^*(x) = \max_{a \in \Gamma(x)} q^*(x,a)$$

To repeat,

$$q^*(x,a) = r(x,a) + \beta \sum_{x'} \max_{a' \in \Gamma(x')} q^*(x',a') P(x,a,x')$$

Hence q^* is the fixed point of

$$(Sq)(x,a) = r(x,a) + \beta \sum_{x'} \max_{a' \in \Gamma(x')} q(x',a') P(x,a,x')$$

Remarks

- unlike the Bellman equation for v^*, the expectation is <u>outside</u> the max
- this helps with stochastic approximation

Note that

$$|(Sf)(x,a) - (Sg)(x,a)|$$

$$\leqslant \beta \left| \sum_{x'} \max_{a' \in \Gamma(x')} f(x',a') - \sum_{x'} \max_{a' \in \Gamma(x')} g(x',a') \right| P(x,a,x')$$

$$= \beta \sum_{x'} \max_{a' \in \Gamma(x')} \left| f(x',a') - g(x',a') \right| P(x,a,x')$$

$$\therefore \quad |(Sf)(x,a) - (Sg)(x,a)| \leqslant \beta \|f - g\|_\infty$$

$$\therefore \quad \|Sf - Sg\| \leqslant \beta \|f - g\|_\infty$$

Fixed point iteration

Let

- $T : \Theta \to \Theta$ be a contraction map of modulus β
- Θ be a closed subset of $\mathbb{R}^n$

We know that $T^k\theta \to \bar{\theta}$ as $k \to \infty$ where $\bar{\theta}$ is the unique fixed point

Alternatively, we can iterate on the damped sequence

$$\theta_{k+1} = (1-\alpha)\theta_k + \alpha T\theta_k$$

$$= \theta_k + \alpha(T\theta_k - \theta_k)$$

- $\alpha \in (0, 1)$

To see that the damped sequence converges, let

$$F\theta = \theta + \alpha(T\theta - \theta)$$

Then

$$F\bar{\theta} = \bar{\theta} + \alpha(T\bar{\theta} - \bar{\theta}) = \bar{\theta}$$

and

$$\|F\theta - F\theta'\| \leqslant (1-\alpha)\|\theta - \theta'\| + \alpha\|T\theta - T\theta'\| \leqslant (1-\alpha+\alpha\beta)\|\theta - \theta'\|$$

Note

$$1 - \alpha + \alpha\beta < 1 \iff \beta < 1$$

Stochastic Approximation

Suppose T is a map with fixed point $\bar{\theta} = T\bar{\theta}$

We can only evaluate T with noise:

$$\text{input } \theta \text{ and receive } T\theta + W$$

- (W_k) is a random (vector-valued) sequence
- We cannot observe W_k, only $T\theta + W_k$

Robbins–Monro algorithm to compute the fixed point $\bar{\theta}$:

$$\theta_{k+1} = \theta_k + \alpha_k [T\theta_k + W_k - \theta_k]$$

- (α_k) is a sequence in $(0, 1)$

Side-by-side comparison:

Here's damped iteration:

$$\theta_{k+1} = (1 - \alpha)\theta_k + \alpha T \theta_k$$

This is Robbins–Monro

$$\theta_{k+1} = \theta_k + \alpha_k [T \theta_k + W_k - \theta_k]$$

By our earlier analysis, $\theta_k \to \bar{\theta}$ if $W_k \equiv 0$ and $\alpha_k \equiv \alpha$

More generally, [Tsi94] proves that if:

- T is an order-preserving contraction map with fixed point $\bar{\theta}$
- $\mathbb{E}[W_{k+1} \mid \mathscr{F}_k] = 0$ for all $k \geqslant 0$
- $\sum_{k \geqslant 0} \alpha_k = \infty$ and $\sum_{k \geqslant 0} \alpha_k^2 < \infty$
- some other technical assumptions,

then

$$\theta_k \to \bar{\theta} \text{ with probability one}$$

Q-Learning

The Q-learning algorithm [Wat89] proposes to learn the Q-factor of an MDP via

$$q_{k+1}(x,a) = q_k(x,a) + \alpha_k \left[r(x,a) + \beta \max_{a' \in \Gamma(X')} q_k(X',a') - q_k(x,a) \right]$$

where $X' \sim P(x,a,\cdot)$

Thm. Under some assumptions,

$$\mathbb{P}\left\{ \lim_{k \to \infty} q_k = q^* \right\} = 1$$

Proved by [Tsi94], [WD92]

We sketch the proof of [Tsi94]

Let

$$W_k := \beta \max_{a' \in \Gamma(X')} q_k(X', a') - \beta \, \mathbb{E} \max_{a' \in \Gamma(X')} q_k(X', a')$$

and recall that

$$(Sq)(x, a) = r(x, a) + \beta \, \mathbb{E} \max_{a' \in \Gamma(X')} q(X', a')$$

Alternatively,

$$(Sq)(x, a) = r(x, a) + \beta \max_{a' \in \Gamma(X')} q(X', a') - W_k$$

In summary,

$$q_{k+1} = q_k + \alpha_k \left[r + \beta \max_{a' \in \Gamma(X')} q_k(X', a') - q_k \right]$$

and

$$Sq_k + W_k = r + \beta \max_{a' \in \Gamma(X')} q_k(X', a')$$

$$\therefore \quad q_{k+1} = q_k + \alpha_k \left[Sq_k + W_k - q_k \right]$$

To repeat,

$$q_{k+1} = q_k + \alpha_k \left[S q_k + W_k - q_k \right]$$

with

$$\mathbb{E} W_k = \mathbb{E} \left[\beta \max_{a' \in \Gamma(X')} q_k(X', a') - \beta \, \mathbb{E} \max_{a' \in \Gamma(X')} q_k(X', a') \right] = 0$$

This is the Robbins–Monro algorithm applied to computing the fixed point of S

- The fixed point of S is the Q-factor q^*

Hence, under certain assumptions, $q_k \to q^*$ with probability one

Online Q-Learning

We analyzed the Q-learning routine

$$q_{k+1}(x,a) = q_k(x,a) + \alpha_k \left[r(x,a) + \beta \max_{a' \in \Gamma(X')} q_k(X',a') - q_k(x,a) \right]$$

where $X' \sim P(x,a,\cdot)$

This is an example of **offline** learning

- update q_{k+1} at every (x,a)

An alternative is **online** learning

- update along a sequence

Let (X_t, A_t) be a state-action sequence

- $X_{t+1} \sim P(X_t, A_t, \cdot)$ for all $t \geqslant 0$
- $R_t := r(X_t, A_t)$ for all $t \geqslant 0$

Update via

$$q_{t+1}(X_t, A_t) =$$

$$q_t(X_t, A_t) + \alpha_t \left[R_t + \beta \max_{a' \in \Gamma(X_{t+1})} q_t(X_{t+1}, a') - q_t(X_t, A_t) \right]$$

- Can learn online without knowing r or P

Q-Learning for Optimal Stopping

Consider an optimal stopping problem with Bellman equation

$$v^*(x) = \max_a \left\{ ae(x) + (1-a) \left[c(x) + \beta \sum_{x'} v^*(x') P(x, x') \right] \right\}$$

- $a \in \{0, 1\}$ stands for "reject", "accept"

Let

$$q^*(x, a) = ae(x) + (1-a) \left[c(x) + \beta \sum_{x'} v^*(x') P(x, x') \right]$$

Now rearrange the Bellman equation and elminate v^* to get

$$q^*(x,a) = ae(x) + (1-a)\left[c(x) + \beta \sum_{x'} \max_{a'} q^*(x',a')P(x,x')\right]$$

Alternatively, $q^* = Sq^*$ where

$$(Sq)(x,a) = ae(x) + (1-a)\left[c(x) + \beta \sum_{x'} \max_{a'} q(x',a')P(x,x')\right]$$

Finally, apply Q-learning with this version of S

References I

- John N Tsitsiklis, *Asynchronous stochastic approximation and q-learning*, Machine learning **16** (1994), no. 3, 185–202.
- Christopher John Watkins, *Learning from delayed rewards*, Tech. report, King's College, Cambridge United Kingdom, 1989.
- Christopher JCH Watkins and Peter Dayan, *Q-learning*, Machine learning **8** (1992), no. 3, 279–292.