Stochastic Approximation and 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

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

Therefore

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

Hence

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

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,

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

Hence Q^* is the fixed point of

$$(SQ)(x, a) = r(x, a) + \beta \sum_{x'} \max_{a' \in \Gamma(x')} Q(x', a')$$

Fixed point iteration

Let

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

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

Alternatively, we can iterate on the damped sequence

$$\begin{aligned} u_{k+1} &= (1-\alpha)u_k + \alpha T u_k \\ &= u_k + \alpha(T u_k - u_k) \end{aligned}$$

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

To see that the damped sequence converges, let

$$Fu = u + \alpha(Tu - u)$$

Then

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

and

$$\|Fu - Fv\| \leqslant (1 - \alpha)\|u - v\| + \alpha\|Tu - Tv\| \leqslant (1 - \alpha + \alpha\beta)\|u - v\|$$

Note

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

Stochastic Approximation (Simplified)

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 sequence with common distribution ϕ
- We cannot observe W_k, only $Y_k = T\theta + W_k$

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 $(0, 1)$

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

Q-Learning

The Q-learning algorithm 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 $X' \sim P(x,a,\cdot)$

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

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

We have

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

and

$$\begin{aligned}
r + \beta & \max_{a' \in \Gamma(X')} Q_k(X', a') \\
& = SQ_k - \beta \, \mathbb{E} \max_{a' \in \Gamma(X')} Q(X', a') + \beta \max_{a' \in \Gamma(X')} Q_k(X', a') \\
& = SQ_k + W_k
\end{aligned}$$

$$\therefore \quad Q_{k+1} = Q_k + \alpha_k \left[SQ_k + W_k - Q_k \right]$$

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, $Q_k \to Q^*$

References I

John N Tsitsiklis, *Asynchronous stochastic approximation and q-learning*, Machine learning **16** (1994), no. 3, 185–202.