On the convergence of Q-learning

Elif Özge Özdamar
elif.ozdamar@helsinki.fi

T-61.6020 Reinforcement Learning - Theory and Applications
February 14, 2006

the covergence of stochastic iterative algorithms

outline

Q-learning algortihm as a stochastic form of DP

Present a proof of convergence for a general class of stochastic processes of which Q-learning is a special case

key works

Watkins (1989) and Watkins and Dayan (1992) proved that Q-learning converges with probability one

Dayan (1992) observed that TD(0) is a special case of Q-learning and therefore converges with probability one

Let

$S = 1, 2, \ldots, N$	be a discrete state space,
$U(i)$	be the discrete set of actions available to the learner when the chain is in state i
$p_{ij}(u)$	probability of making a transition from state i to state j where $u \in U(i)$
$policy\ \mu$	a function from states to actions which the learner defines.
$p_{ij}(\mu(i))$	state transition probabilities. Associated with every policy μ is a markov chain defined by these probabilities.
$c_i(u)$	is instantaneous cost associated with each state i and action μ $c_i(u)$ is a random variable with expected value $\bar{c}_i(u)$
$V_\mu(i)$	is a value function which is the expected sum of discounted future costs

given that the system begins in state i and follows policy μ; Value function *is:*

$$V_\mu(i) = \lim_{N\to\infty} E\{\sum_{t=0}^{N-1} \gamma^t c_{s_t}(\mu(s_t))|s_0 = i\}, \quad 1$$

Where $s_t \in S$ is the state of the markov chain at time t and

Future costs are discounted by a factor γ^t where $\gamma \in (0,1)$

We wish to find a policy that minimizes the value function

$$V^*(i) = \min_\mu V_\mu(i) \quad 2$$

Such a policy is referred to as an *optimal policy* and the corresponding function is referred to as the *optimal value function*. Optimal value function is unique, but an optimal policy need not to be!

The Bellman's equation characterizes the optimal value of the state in terms of optimal values of possible successor states

$$V^*(i) = \min_{u\in U(i)} \{\bar{c}_i(u) + \gamma \sum_{j\in S} p_{ij}(u) V^*(j)\} \quad 3$$

To motivate Belman's equation, suppose that the system is in state i at time t. consider how $V^*(i)$ should be characterized in terms of possible transitions out of state i ??

$$V^*(i) = \min_{u \in U(i)} \{\bar{c}_i(u) + \gamma \sum_{j \in S} p_{ij}(u) V^*(j)\} \quad 3$$

Suppose that action u is selected and the system transitions to state j. The expression

$$c_i(u) + \gamma V^*(j)$$

is the cost of making a transition out of state i plus the discounted cosf of following an optimal policy thereafter

The minimum of the expected value of this expression, over possible choices of actions, is a plausible measure of the optimal cost at i and by Bellman's equation is indeed equal to $V^*(i)$

Solving Bellmans's equation by value iteration

$$V^*(i) = \min_{u \in U(i)} \{\bar{c}_i(u) + \gamma \sum_{j \in S} p_{ij}(u) V^*(j)\} \quad 3$$

Value iteration solves for $V^*(i)$ by setting up a recurrence relation for which Bellman's equation is a fixed point

Denoting the estimate of $V^*(i)$ at the k^{th} iteration as $V^{(k)}(i)$,

we have;

$$V^{(k+1)}(i) = \min_{u \in U(i)} \{\bar{c}_i(u) + \gamma \sum_{j \in S} p_{ij}(u) V^{(k)}(j)\} \quad 4$$

This iteration can be shown to converge to $V^*(i)$ for arbitary initial $V^{(0)}(i)$ (Bertsekas, 87)

The proof is based on showing that the iteration from $V^{(k)}(i)$ to $V^{(k+1)}(i)$ is a contraction mapping. It can be shown that

$$\max_i |V^{(k+1)}(i) - V^*(i)| \leq \gamma \max_i |V^{(k)}(i) - V^*(i)| \quad 5$$

which implies that $V^{(k)}(i)$ converges to $V^*(i)$ and also places an upper bound on the convergence rate

Watkins (89) utilized an alternative notation for expressing Bellman's equation that is particulary convenient for deriving learning algorithms.

Define the function $Q^*(i,u)$ to be the expression appearing inside the "min" operator of Bellman's equation

$$Q^*(i,u) = \bar{c}_i(u) + \gamma \sum_{j \in S} p_{ij}(u) V^*(j) \quad (6)$$

Using this notation Bellman's equation can be written as:

$$V^*(i) = \min_{u \in U(i)} Q^*(i,u) \quad (7)$$

Moreover, value iteration can be expressed in terms of Q functions:

$$Q^{(k+1)}(i,u) = \bar{c}_i(u) + \gamma \sum_{j \in S} p_{ij}(u) V^{(k)}(j), \quad 8$$

Where $V^{(k)}i$ is defined in terms of $Q^{(k)}(i,u)$

$$V^{(k)}(i) = \min_{u \in U(i)} Q^{(k)}(i,u) \quad 9$$

Using Q' s instead of V ' s derives from the fact that the minimization operator appears inside the expectation in EQ.8 whereas it appears outside of the expectation in EQ. 4 This fact plays an important role in the convergence proof.

$$V^{(k+1)}(i) = \min_{u \in U(i)} \{\bar{c}_i(u) + \gamma \sum_{j \in S} p_{ij}(u) V^{(k)}(j)\} \quad 4$$

The Q-learning algorithm is a stochastic form of value iteration.

$$Q^{(k+1)}(i,u) = \bar{c}_i(u) + \gamma \sum_{j \in S} p_{ij}(u) V^{(k)}(j), \quad 8$$

EQ.8 expresses the update of the Q values in term of the Q values of successor states. To perform a step of value iteration requires knowing the expected cost and the transition probabilities. Although such a step cannot be performed without a model, it is possible to estimate the appropriate update.

For an arbitrary V function the quantity $\sum_{j \in S} p_{ij}(u) V(j)$ can be estimated by the quantity $V(j)$, if successor state j is chosen with probability $p_{ij}(u)$. But this is assured by simply following the transitions of the actual Markovian environment, which makes a transition from state i to state j with probability $p_{ij}(u)$.

Thus the sample value of V at the successor state is an unbiased estimate of the sum. Moreover, $c_i(u)$ is an unbiased estimate of $\bar{c}_i(u)$

This reasoning leads to the following algorithm ,

$$Q_{t+1}(s_t, u_t) = (1 - \alpha_t(s_t, u_t))Q_t(s_t, u_t) + \alpha_t(s_t, u_t)[c_{s_t}(u_t) + \gamma V_t(s_{t+1})] \quad 10$$

where

Where $Q_t(i, u)$ and $V_t(i)$ denote the learner's estimates of the Q and V functions at time t respectively

$$V_t(s_{t+1}) = \min_{u \in U(s_{t+1})} Q_t(s_t, u_t) \quad 11$$

The variables $\alpha_t(s_t, u_t)$ are zero except for the state that is being updated at time t

as conclusion

The fact that Q-learning is a stochastic form of value iteration immediately suggests the use of stochastic approximatiin theory, in particular the classical framework of Robbins and Monro(51)

Robbins-Monto theory treats the stochastic convergence of a sequence of unbiased estimates of a regression function, providing conditions under which the sequence converges to a root of the function.Although the stochastic convergence of Q-learning is not an immediate concequence of Robbins Monro theory the theory does provide results that can be adapter to studying the convergence of DP based learning algorithms!!??

Convergence proof of Q learning

Theorem 1 *A random iterative process* $\Delta_{n+1}(x) = (1-\alpha_n(x))\Delta_n(x)+\beta_n(x)F_n(x)$ *converges to zero w.p.1 under the following assumptions:*

1) The state space is finite.

2) $\sum_n \alpha_n(x) = \infty$, $\sum_n \alpha_n^2(x) < \infty$, $\sum_n \beta_n(x) = \infty$, $\sum_n \beta_n^2(x) < \infty$, *and* $\mathrm{E}\{\beta_n(x)|P_n\} \leq \mathrm{E}\{\alpha_n(x)|P_n\}$ *uniformly w.p.1.*

3) $\| \mathrm{E}\{F_n(x)|P_n\} \|_W < \gamma \| \Delta_n \|_W$, *where* $\gamma \in (0,1)$.

4) $\mathrm{Var}\{F_n(x)|P_n\} \leq C(1+ \| \Delta_n \|_W)^2$, *where* C *is some constant.*

Here $P_n = \{\Delta_n, \Delta_{n-1}, \ldots, F_{n-1}, \ldots, \alpha_{n-1}, \ldots, \beta_{n-1}, \ldots\}$ *stands for the past at step* n. $F_n(x)$, $\alpha_n(x)$ *and* $\beta_n(x)$ *are allowed to depend on the past insofar as the above conditions remain valid. The notation* $\| \cdot \|_W$ *refers to some weighted maximum norm.*

proof

Lemma 1 *A random process*

$$w_{n+1}(x) = (1 - \alpha_n(x))w_n(x) + \beta_n(x)r_n(x).$$

converges to zero with probability one if the following conditions are satisfied:

1) $\sum_n \alpha_n(x) = \infty$, $\sum_n \alpha_n^2(x) < \infty$, $\sum_n \beta_n(x) = \infty$, *and* $\sum_n \beta_n^2(x) < \infty$ *uniformly w.p.1.*

2) $\mathrm{E}\{r_n(x)|P_n\} = 0$ *and* $\mathrm{E}\{r_n^2(x)|P_n\} \leq C$ *w.p.1, where*

$$P_n = \{w_n, w_{n-1}, \ldots, r_{n-1}, r_{n-2}, \ldots, \alpha_{n-1}, \alpha_{n-2}, \ldots, \beta_{n-1}, \beta_{n-2}, \ldots\}$$

All the random variables are allowed to depend on the past P_n.

Proof. Except for the appearance of $\beta_n(x)$ this is a standard result. With the above definitions convergence follows directly from Dvoretzky's extended theorem (Dvoretzky, 1956).

Lemma 2 *Consider a random process* $X_{n+1}(x) = G_n(X_n, x)$, *where*

$$G_n(\beta X_n, x) = \beta G_n(X_n, x)$$

Let us suppose that if we kept $\| X_n \|$ *bounded by scaling, then* X_n *would converge to zero w.p.1. This assumption is sufficient to guarantee that the original process converges to zero w.p.1.*

Proof. Note that the scaling of X_n at any point of the iteration corresponds to having started the process with scaled X_0. Fix some constant C. If during the iteration, $\| X_n \|$ increases above C, then X_n is scaled so that $\| X_n \| = C$. By the assumption then this process must converge w.p.1. To show that the net effect of the corrections must stay finite w.p.1 we note that if $\| X_n \|$ converges then for any $\epsilon > 0$ there exists M_ϵ such that $\| X_n \| < \epsilon < C$ for all $n > M_\epsilon$ with probability at least $1 - \epsilon$. But this implies that the iteration stays below C after M_ϵ and converges to zero without any further corrections. □

Lemma 3 *A stochastic process* $X_{n+1}(x) = (1 - \alpha(x))X_n(x) + \gamma\beta_n(x) \parallel X_n \parallel$ *converges to zero w.p.1 provided*

1) $x \in S$, *where* S *is a finite set.*

2) $\sum_n \alpha_n(x) = \infty$, $\sum_n \alpha_n^2(x) < \infty$, $\sum_n \beta_n(x) = \infty$, $\sum_n \beta_n^2(x) < \infty$, *and* $\mathrm{E}\{\beta_n(x)\} \leq \mathrm{E}\{\alpha_n(x)\}$ *uniformly w.p.1.*

Proof. Essentially the proof is an application of Lemma 2. To this end, assume that we keep $\parallel X_n \parallel \leq C_1$ by scaling which allows the iterative process to be bounded by

$$|X_{n+1}(x)| \leq (1 - \alpha_n(x))|X_n(x)| + \gamma\beta_n(x)C_1$$

This is linear in $|X_n(x)|$ and can be easily shown to converge w.p.1 to some $X^*(x)$, where $\parallel X^* \parallel \leq \gamma C_1$. Hence, for small enough ϵ, there exists $M_1(\epsilon)$ such that $\parallel X_n \parallel \leq C_1/(1+\epsilon)$ for all $n > M_1(\epsilon)$ with probability at least $p_1(\epsilon)$. With probability $p_1(\epsilon)$ the procedure can be repeated for $C_2 = C_1/(1+\epsilon)$. Continuing in this manner and choosing $p_k(\epsilon)$ so that $\prod_k p_k(\epsilon)$ goes to one as $\epsilon \to 0$ we obtain the w.p.1 convergence of the bounded iteration and Lemma 2 can be applied. □

Theorem1

Proof. By defining $r_n(x) = F_n(x) - \mathrm{E}\{F_n(x)|P_n\}$ we can decompose the iterative process into two parallel processes given by

$$\begin{aligned} \delta_{n+1}(x) &= (1-\alpha_n(x))\delta_n(x) + \beta_n(x)\mathrm{E}\{F_n(x)|P_n\} \\ w_{n+1}(x) &= (1-\alpha_n(x))w_n(x) + \beta_n(x)r_n(x) \end{aligned} \tag{21}$$

where $\Delta_n(x) = \delta_n(x) + w_n(x)$. Dividing the equations by $W(x)$ for each x and denoting $\delta_n'(x) = \delta_n(x)/W(x)$, $w_n'(x) = w_n(x)/W(x)$, and $r_n'(x) = r_n(x)/W(x)$ we can bound the δ_n' process by assumption 3) and rewrite the equation pair as

$$\begin{aligned} |\delta_{n+1}'(x)| &\leq (1-\alpha_n(x))|\delta_n'(x)| + \gamma\beta_n(x) \parallel |\delta'| + w_n' \parallel \\ w_{n+1}'(x) &= (1-\alpha_n(x))w_n'(x) + \gamma\beta_n(x)r_n'(x) \end{aligned}$$

Assume for a moment that the Δ_n process stays bounded. Then the variance of $r_n^{'}(x)$ is bounded by some constant C and thereby $w_n^{'}$ converges to zero w.p.1 according to Lemma 1. Hence, there exists M such that for all $n > M \parallel w_n^{'} \parallel < \epsilon$ with probability at least $1 - \epsilon$. This implies that the $\delta_n^{'}$ process can be further bounded by

$$|\delta_{n+1}^{'}(x)| \leq (1 - \alpha_n(x))|\delta_n^{'}(x)| + \gamma\beta_n(x) \parallel \delta_n^{'} + \epsilon \parallel$$

with probability $> 1 - \epsilon$. If we choose C such that $\gamma(C+1)/C < 1$ then for $\parallel \delta_n^{'} \parallel > C\epsilon$

$$\gamma \parallel \delta_n^{'} + \epsilon \parallel \leq \gamma(C+1)/C \parallel \delta_n^{'} \parallel$$

and the process defined by this upper bound converges to zero w.p.1 by Lemma

3. Thus $\| \delta_n^{'} \|$ converges w.p.1 to some value bounded by $C\epsilon$ which guarantees the w.p.1 convergence of the original process under the boundedness assumption.

By assumption (4) $r_n^{'}(x)$ can be written as $(1+ \| \delta_n + w_n \|)s_n(x)$, where $\mathrm{E}\{s_n^2(x)|P_n\} \leq C$. Let us now decompose w_n as $u_n + v_n$ with

$$u_{n+1}(x) = (1 - \alpha_n(x))u_n(x) + \gamma\beta_n(x) \| \delta_n^{'} + u_n + v_n \| s_n(x)$$

and v_n converges to zero w.p.1 by Lemma 1. Again by choosing C such that $\gamma(C+1)/C < 1$ we can bound the $\delta_n^{'}$ and u_n processes for $\| \delta_n^{'} + u_n \| > C\epsilon$. The pair $(\delta_n^{'}, u_n)$ is then a scale invariant process whose bounded version was proven earlier to converge to zero w.p.1 and therefore by Lemma 2 it too converges to zero w.p.1. This proves the w.p.1 convergence of the triple $\delta_n^{'}$, u_n, and v_n bounding the original process. □

Theorem 2 *The Q-learning algorithm given by*

$$Q_{t+1}(s_t, u_t) = (1 - \alpha_t(s_t, u_t))Q_t(s_t, u_t) + \alpha_t(s_t, u_t)[c_{s_t}(u_t) + \gamma V_t(s_{t+1})]$$

converges to the optimal $Q^*(s, u)$ *values if*

1) The state and action spaces are finite.

2) $\sum_t \alpha_t(s, u) = \infty$ *and* $\sum_t \alpha_t^2(s, u) < \infty$ *uniformly w.p.1.*

3) $\mathrm{Var}\{c_s(u)\}$ *is bounded.*

3) If $\gamma = 1$ *all policies lead to a cost free terminal state w.p.1.*

proof

Proof. By subtracting $Q^*(s,u)$ from both sides of the learning rule and by defining $\Delta_t(s,u) = Q_t(s,u) - Q^*(s,u)$ together with

$$F_t(s,u) = c_s(u) + \gamma V_t(s_{next}) - Q^*(s,u) \tag{12}$$

the Q-learning algorithm can be seen to have the form of the process in theorem 1 with $\beta_t(s,u) = \alpha_t(s,u)$.

To verify that $F_t(s,u)$ has the required properties we begin by showing that it is a contraction mapping with respect to some maximum norm. This is done by relating F_t to the DP value iteration operator for the same Markov chain. More specifically,

$$\begin{aligned}
\max_u |\mathrm{E}\{F_t(i,u)\}| &= \gamma \max_u |\sum_j p_{ij}(u)[V_t(j) - V^*(j)]| \\
&\leq \gamma \max_u \sum_j P_{ij}(u) \max_v |Q_t(j,v) - Q^*(j,v)| \\
&= \gamma \max_u \sum_j P_{ij}(u) V^{\Delta}(j) = T(V^{\Delta})(i)
\end{aligned}$$

where T is the DP value iteration operator for the case where the costs associated with each state are zero. If $\gamma < 1$ the contraction property of T and thus

of F_t can be seen directly from the above formulas. When the future costs are not discounted ($\gamma = 1$) but the chain is absorbing and all policies lead to the terminal state w.p.1 there still exists a weighted maximum norm with respect to which T is a contraction mapping (see e.g. Bertsekas & Tsitsiklis, 1989).

The variance of $F_t(s, u)$ given the past is within the bounds of theorem 1 as it depends on $Q_t(s, u)$ at most linearly and the variance of $c_s(u)$ is bounded.

Note that the proof covers both the on-line and batch versions. □

Resources

These slides mostly rely on the technical report 'On the convergence of stochastic iterative dynamic programming algorithms' by Jaakkola T., Jordan M., and Singh S. which is one of the resources of the book Neuro-Dynamic Programming, by Bertsekas D. and Tsitsiklis J.

For more theoretical work;

Neuro-Dynamic Programming, by Bertsekas D. and Tsitsiklis J., Athena Scientific, 1996

Eyal Even-Dar and , Yishay Mansour, Learning Rates for Q-learning, Journal of Machine Learning Research 5 (2003) 1-25