Error Bounds for Constant Step-Size \(Q\)-learning

C. L. Beck\textsuperscript{a}, R. Srikant\textsuperscript{b}

\textsuperscript{a}Department of Industrial and Enterprise Systems Engineering and Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; beck3@illinois.edu
\textsuperscript{b}Department of Electrical and Computer Engineering and Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; rsrikant@illinois.edu

Abstract

We provide a bound on the first moment of the error in the \(Q\)-function estimate resulting from fixed step-size algorithms applied to finite state-space, discounted reward Markov decision problems. Motivated by Tsitsiklis’ proof for the decreasing step-size case, we decompose the \(Q\)-learning update equations into a dynamical system driven by a noise sequence and another dynamical system whose state variable is the \(Q\)-learning error, i.e., the difference between the true \(Q\)-function and the estimate. A natural persistence of excitation condition allows us to sample the system periodically and derive a simple scalar difference equation from which the convergence properties and bounds on the error of the \(Q\)-learning algorithm can be derived.

1. Introduction

\(Q\)-learning is a reinforcement learning technique which was introduced by Watkins [11]. In traditional stochastic control problems, it is assumed that the system parameters, i.e., the parameters of the Markov chain to be controlled, are known and the goal is to choose the control as a function of the state to maximize some objective. When the system parameters are unknown, there are two options: one can first attempt to learn the parameters through system identification techniques and then control the system, or learn the parameters as the system is being controlled. \(Q\)-learning falls into the latter category, and has been shown to perform very well in many practical situations. From a machine learning perspective, \(Q\)-learning can be viewed as a technique to select the best action given the current state by learning a function which measures the quality of taking a particular action at a given state.
The basic idea behind $Q$-learning is to learn a variant of the well-known value function, known as the $Q$-function, in dynamic programming. However, when this is done in an online fashion, i.e., by observing the random evolution of the system which is being controlled, the randomness in the system has to be taken into account. In such situations, stochastic approximation suggests that the estimates of the unknown quantities should be updated slowly. The rate of update is controlled by a parameter called the step-size. In the most commonly studied forms of learning, the step-size slowly (but not too slowly) tends to zero as time goes to infinity. However, many practical implementations of $Q$-learning use a fixed step-size to account for very slow variations in the parameters of the Markov chain. In this paper, we study the fixed step-size case for the traditional model with a finite state space.

The first convergence proof for $Q$-learning was given by Watkins and Dayan in [12], and has been studied fairly extensively since. Tsitsiklis studied the convergence properties of the $Q$-learning algorithm for the case of a decreasing step-size, by making connections to stochastic approximation algorithms [10]. Related results were obtained at the same time by Jaakola, Jordan and Singh in [7]. Tsitsiklis’ results were extended in [1, 5] where the relationship between the rate at which the step-size decreases and the rate of convergence in $Q$-learning is studied.

Stochastic approximation itself has a long and rich history of convergence results, mainly via the study of related ODEs (see [8, 2] and the references therein). Large deviations and central limit theorem arguments provide estimates of the probability of deviation from the given ODE. These estimates can be used to understand the tradeoff between convergence rate and steady-state error for fixed step-size $Q$-learning algorithms. A central limit theorem based approach was used by Levy et al [9], to study $Q$-learning in a non-stationary environment. Borkar and Meyn introduced a fluid model approach to prove stability, which generalized the applicability of $Q$-learning convergence proofs to more general settings than had been previously established [3].

In this paper, we consider the classic and most widely-used model for $Q$-learning in practice, that is the discounted reward, infinite-horizon problem, where the system is a discrete, finite state-space Markov chain with a finite number of available control actions in each state. We assume the $Q$-learning step-size is constant. Surprisingly, for this seemingly simple problem, there does not appear to be any error bounds under the assumption of a sufficiently small but fixed step-size.

Here we provide a straightforward bound on the first moment of the error in the constant step-size $Q$-learning problem, using elementary tools. We first derive the error bound for the case of synchronous $Q$-learning in Section 2, followed
by the error bound for the asynchronous case in Section 3. In each section, we use the Tsitsiklis method of decomposing the $Q$-learning iteration into a convergent stochastic iteration and a contraction mapping iteration, perturbed by a small noise. However, the proof we provide is different beyond this decomposition due to the fact that we consider the case of a fixed step-size. Interestingly, the expression for the bound on the first moment on the error has similar forms in both the synchronous and asynchronous cases; see (9) for the synchronous case and (20)-(21) for the asynchronous case.

2. Synchronous $Q$-Learning

In this paper, as noted earlier, we only consider the case of finite-state Markov chains where the objective is to maximize an infinite-horizon discounted cost. In each state, we assume that the set of allowable control actions is finite and the reward obtained when taking a particular action in a particular state is deterministic. If the parameters of the Markov chain are known, then one can use Dynamic Programming (DP) to compute the optimal policy. $Q$-learning is used in the case where the system parameters are unknown. In particular, $Q$-learning does not estimate the system parameters directly (as in system identification) but rather estimates the DP value function directly.

Before we describe the $Q$-learning algorithm, we first define the system model. The control system is described by a finite state-space discrete-time Markov chain, where we denote the state-space by $S$. The state at time $k$ is denoted by $x_k$. In each state, there is a set of allowable controls; without loss of generality, we assume that the set of allowable control actions is the same in each state and is denoted by $A$. The control action taken at time $k$ is denoted by $u_k$. Control action $u$ taken when the system is in state $x$ results in a deterministic reward $r(x,u)$. Next we describe the dynamics of the system. When the system is in state $i$ and control action $u$ is applied, the system moves to a new state $j$ with probability $p_{ij}(u)$. In other words, $p_{ij}(u) = \text{Prob}(x_{k+1} = j|x_k = i, u_k = u)$. Our objective is to find a control policy to maximize

$$E\left(\sum_{k=0}^{\infty} \alpha^k r(x_k, u_k)\right),$$

where $\alpha \in (0, 1)$ is a discount factor.

If the state transition probabilities $p_{ij}(u)$ were known, we could solve for the
optimal control using the Bellman equation:

$$V(i) = \max_{u \in \mathcal{A}} \{ r(i,u) + \alpha E \{ V(x_{k+1}) | x_k = i, u_k = u \} \}$$

which can also be written as

$$V(i) = \max_{u \in \mathcal{A}} \left\{ r(i,u) + \alpha \sum_j V(j)p_{ij}(u) \right\}.$$

In general, the state transition probabilities are unknown and therefore, $Q$-learning attempts to infer the following function:

$$Q(i,u) := r(i,u) + \alpha \sum_j V(j)p_{ij}(u).$$

Note that the $Q$-function has the following simple interpretation: if $i$ is the current state, then the $Q$-function is the expected infinite-horizon discounted reward if the control action $u$ is taken at the current time step and the optimal policy is used from the next step onwards. The $Q$-function is related to the value function $V$ as follows:

$$V(i) = \max_{u \in \mathcal{A}} Q(i,u),$$

Using this relationship, the $Q$-function can be rewritten as

$$Q(i,u) = r(i,u) + \alpha \sum_j p_{ij}(u) \left[ \max_{v \in \mathcal{A}} Q(j,v) \right]. \quad (1)$$

Defining the operator $T$ as

$$(TQ)(i,u) = r(i,u) + \alpha \sum_j p_{ij}(u) \left[ \max_{v \in \mathcal{A}} Q(j,v) \right],$$

we note the $Q$-function is the fixed point of the operator, i.e., $Q = TQ$, where $Q$ is a $|\mathcal{S} \times \mathcal{A}|$ vector of $Q$-values, with each element of the vector corresponding to a state-action pair.

We first recall the following well-known contraction property of the operator $T$ [10].

**Fact 1.** $T$ is a contraction mapping, i.e.,

$$\|TQ_1 - TQ_2\|_\infty \leq \alpha \|Q_1 - Q_2\|_\infty$$

for any two vectors $Q_1$ and $Q_2$. 

4
Since \( p_{i,j}(u) \) is unknown, the above contraction property of \( T \) cannot be used to compute \( Q \). Alternatively, \( Q \)-learning uses stochastic approximation, i.e.,

\[
\hat{Q}_{k+1}(i,u) = (1 - \varepsilon) \hat{Q}_k(i,u) + \varepsilon \left[ r(i,u) + \alpha \max_v \hat{Q}_k(j,v) \right],
\]

where \( \varepsilon \in (0,1) \) is a step-size parameter and \( j \) is the (random) next state reached by applying the control \( u \) in state \( i \). We assume that the initial values \( \hat{Q}_0(i,u) \) are deterministic constants. Equation (2) is called the synchronous \( Q \)-learning algorithm. In this section, our goal is to understand the convergence properties of (2).

Define the error in approximating the \( Q \)-function by

\[
\tilde{Q}_k(i,u) := \hat{Q}_k(i,u) - Q(i,u).
\]

Using (1) and (2), the error dynamics can be written as

\[
\tilde{Q}_{k+1}(i,u) = (1 - \varepsilon) \tilde{Q}_k(i,u) + \varepsilon \left[ r(i,u) + \alpha \max_v \hat{Q}_k(j,v) - \alpha \sum_j p_{ij}(u) \max_v Q_k(j,v) \right]
\]

\[
= (1 - \varepsilon) \tilde{Q}_k(i,u) + \varepsilon \left[ \alpha \max_v \hat{Q}_k(j,v) - \alpha \sum_j p_{ij}(u) \max_v Q_k(j,v) \right]
\]

\[
= (1 - \varepsilon) \tilde{Q}_k(i,u) + \varepsilon \alpha \sum_j p_{ij}(u) \left( \max_v \hat{Q}_k(j,v) - \max_v Q_k(j,v) \right)
\]

\[
+ \alpha \varepsilon w_k(i,u)
\]

\[
= (1 - \varepsilon) \tilde{Q}_k(i,u) + \varepsilon \left( (T \hat{Q})(i,u) - (T Q)(i,u) \right) + \alpha \varepsilon w_k(i,u) \quad (3)
\]

where \( w_k(i,u) := \left[ \max_v \hat{Q}_k(j,v) - \sum_j p_{ij}(u) \max_v \hat{Q}_k(j,v) \right] \).

Due to our system modelling assumptions, the error is bounded (as a function of the initial guess for \( Q \)) as stated in the following fact; for a proof of this result, see [6].

**Fact 2.** There exists \( \tilde{Q}_{max} \) such that

\[
\| \tilde{Q}_k \|_{\infty} \leq \tilde{Q}_{max}, \forall k.
\]

We now provide a bound on the expected value of the difference between \( \hat{Q}_k \) and \( Q \) as a function of the number of time-steps \( k \). From this bound, we can characterize the amount of time that it takes for the probability of error to be small. As mentioned earlier, to do this, we decompose the \( Q \)-learning dynamics into two parts: one is a dynamical system driven by the noise sequence \( w \) and the other is the dynamics of the error sequence.
2.1. A Noise-Driven System

First we define the sequence $z$ as

$$z_{k+1}(i,u) = (1 - \varepsilon)z_k(i,u) + \alpha\varepsilon w_k(i,u), \; z_0(i,u) = 0. \quad (4)$$

Note that $z$ is a dynamical system driven by the noise sequence $w$.

We now consider the convergence properties of $z_k$. From (4),

$$E(z_{k+1}(i,u)) = (1 - \varepsilon)E(z_k(i,u)) + \alpha\varepsilon E(w_k(i,u)); \; E(z_0(i,u)) = 0.$$ 

Let $\eta(i,u,k)$ be the random next state reached by applying the control action $u$ to the system at time $k$ when the state is $i$. Note that $z_k$ is fully determined by the past history $\Xi_k$, where

$$\Xi_k := \{ \eta(i,u,m), i \in \mathcal{S}, u \in \mathcal{U}, m \leq k - 1 \},$$

and also note that $E(w_k(i,u) | \Xi_k) = 0$. Thus, we have

$$E(w_k(i,u)) = 0 \Rightarrow E(z_k(i,u)) = 0 \; \forall k.$$ 

Further,

$$E(z_k(i,u) w_k(i,u) | \Xi_k) = z_k(i,u) E(w_k(i,u) | \Xi_k) = 0,$$

and thus $E(z_k(i,u) w_k(i,u)) = 0$.

Now consider the expected value of $z_k^2(i,u)$. Using (4) and (5), we have

$$E(z_{k+1}^2(i,u)) = (1 - \varepsilon)^2 E(z_k^2(i,u)) + \varepsilon^2 \alpha^2 E(w_k^2(i,u)); \; E(z_0^2(i,u)) = 0$$

$$\Rightarrow E(\|z_{k+1}\|^2) = (1 - \varepsilon)^2 E(\|z_k\|^2) + \varepsilon^2 \alpha^2 E(\|w_k\|^2); \; E(\|z_0\|^2) = 0.$$ 

From the definition for $w_k(i,u)$ we have $|w_k(i,u)| \leq 2Q_{\max}$, thus

$$\|w_k\|^2 \leq |\mathcal{S} \times \mathcal{U}| 4Q_{\max}^2 =: W_{\max}^2,$$

where $|\mathcal{S}|$ denotes the cardinality of the set of states indexed by $i$, and $|\mathcal{U}|$ denotes the cardinality of the set of possible controls $u$, giving us

$$E(\|z_{k+1}\|^2) \leq (1 - \varepsilon)^2 E(\|z_k\|^2) + \varepsilon^2 \alpha^2 W_{\max}^2; \; E(\|z_0\|^2) = 0,$$

or, equivalently,

$$E(\|z_k\|^2) \leq \sum_{l=0}^{k-1} ((1 - \varepsilon)^2)^l \varepsilon^2 \alpha^2 W_{\max}^2.$$

Thus,

$$E(\|z_k\|^2) \leq \alpha^2 \varepsilon^2 W_{\max}^2 \left( \frac{1 - (\varepsilon(2 - \varepsilon))^k}{\varepsilon(2 - \varepsilon)} \right) \leq \frac{\alpha^2 \varepsilon W_{\max}^2}{(2 - \varepsilon)}. \quad (6)$$

Next, we will use (6) to obtain a bound on the error in synchronous $Q$-learning.
2.2. Error Dynamics

Our goal is to obtain a bound on the expected value of the norm of the approximation error $\tilde{Q}_k$. Towards this end, we define $D_k := \tilde{Q}_k - z_k$. Then $D_0 = \tilde{Q}_0$, since $z_0 = 0$. The dynamics of $D$ are given by

$$D_{k+1}(i,u) = (1 - \varepsilon)D_k(i,u) + \varepsilon \left( (T\tilde{Q}_k)(i,u) - (TQ)(i,u) \right),$$

which yields

$$|D_{k+1}(i,u)| \leq (1 - \varepsilon)|D_k(i,u)| + \varepsilon \|T\tilde{Q}_k - TQ\|_\infty,$$

where the last step follows from the fact that the operator $T$ is a contraction mapping. Recall that $\tilde{Q}_k = \tilde{Q}_k - Q$; thus,

$$|D_{k+1}(i,u)| \leq (1 - \varepsilon)|D_k(i,u)| + \varepsilon \|\tilde{Q}_k - Q\|_\infty. \hspace{1cm} (7)$$

and

$$\|D_{k+1}\|_\infty \leq (1 - \varepsilon)\|D_k\|_\infty + \alpha \varepsilon \|\tilde{Q}_k\|_\infty.$$ \hspace{1cm} (8)

giving us

$$\|D_{k+1}\|_\infty \leq (1 - \varepsilon)\|D_k\|_\infty + \alpha \varepsilon \|\tilde{Q}_k\|_\infty^k = \|\tilde{Q}_0\|_\infty.$$

Iterating from $\|D_0\|_\infty$ gives us

$$\|D_k\|_\infty \leq (1 - \varepsilon(1 - \alpha))^k\|D_0\|_\infty + \alpha \varepsilon \sum_{l=0}^{k-1} (1 - \varepsilon(1 - \alpha))^l \|z_{k-1-l}\|_\infty.$$

Since $\|z_k\|_\infty^2 \leq \|z_k\|_2^2$, we have from (6)

$$E(\|z_k\|_2^2) \leq \frac{\alpha^2 \varepsilon W_{\max}^2}{2 - \varepsilon}.$$

Since $\varepsilon(1 - \alpha) < 1$, we have

$$E(\|D_k\|_\infty) \leq (1 - \varepsilon(1 - \alpha))^k \tilde{Q}_{\max} + \alpha \sqrt{\frac{\alpha^2 \varepsilon W_{\max}^2}{2 - \varepsilon}} \frac{1}{1 - \alpha} \hspace{1cm} \frac{\varepsilon W_{\max}^2}{(2 - \varepsilon)^2},$$

7
where we have used the fact that $\|D_0\|_\infty = \|\tilde{Q}_0\|_\infty \leq \tilde{Q}_{\max}$.

This leads us to the following theorem, which is our main result for synchronous $Q$-learning.

**Theorem 2.1.** The error at time step $k$ is bounded as follows:

$$E(\|\tilde{Q}_k\|_\infty) \leq (1 - \epsilon(1 - \alpha))^k \tilde{Q}_{\max} + \frac{\alpha}{1 - \alpha} \sqrt{\frac{\epsilon W_{\max}^2}{2 - \epsilon}}. \quad (9)$$

Thus, given any $x > 0$,

$$P(\|\tilde{Q}_k\|_\infty \geq x) \leq \frac{1}{x} \left( (1 - \epsilon(1 - \alpha))^k \tilde{Q}_{\max} + \frac{\alpha}{1 - \alpha} \sqrt{\frac{\epsilon W_{\max}^2}{2 - \epsilon}} \right).$$

**Proof.** The proof follows from the discussion prior to the theorem and directly from the relation

$$E(\|\tilde{Q}_k\|_\infty) \leq E(\|D_k\|_\infty) + E(\|z_k\|_\infty),$$

and the Markov inequality. $\square$

3. Asynchronous $Q$-learning

Consider the asynchronous $Q$-learning algorithm given by

$$\hat{Q}_{k+1}(i, u) = \begin{cases} (1 - \epsilon)\hat{Q}_k(i, u) + \epsilon [r(i, u) + \alpha \max_v \hat{Q}_k(x_{k+1}, v)], & \text{if } (i, u) = (x_k, u_k) \\ \hat{Q}_k(i, u), & \text{if } (i, u) \neq (x_k, u_k). \end{cases}$$

The above dynamics can be interpreted as an online implementation of $Q$-learning where one can observe the reward for a given state-action pair only when the system reaches that particular state, in which case the corresponding $Q$-value is updated. Otherwise, the reward is not observed, and therefore, the $Q$-value for this (state, action) cannot be updated. Defining

$$w_k(i, u, \hat{Q}_k, x_{k+1}) := \left[ \max_v \hat{Q}_k(x_{k+1}, v) - \sum_j p_{ij}(u) \max_v \hat{Q}_k(j, v) \right], \quad (10)$$

the asynchronous $Q$-learning algorithm can be rewritten as

$$\hat{Q}_{k+1}(i, u) = \begin{cases} (1 - \epsilon)\hat{Q}_k(i, u) + \epsilon \left[ T\hat{Q}_k(i, u) + w_k(x_k, u_k, \hat{Q}_k, x_{k+1}) \right], & \text{if } (i, u) = (x_k, u_k) \\ \hat{Q}_k(i, u), & \text{if } (i, u) \neq (x_k, u_k). \end{cases}$$
Note that, defining 
\[
\varepsilon_k(i,u) = \begin{cases} 
\varepsilon, & \text{if } (x_k,u_k) = (i,u), \\
0, & \text{otherwise},
\end{cases}
\]
allows us to write the Q-learning algorithm more succinctly as
\[
\hat{Q}_{k+1}(i,u) = (1 - \varepsilon_k(i,u))\hat{Q}_k(i,u) + \varepsilon_k(i,u)[(T\hat{Q}_k)(i,u) + w_k(x_k,u_k,\hat{Q}_k,x_{k+1})].
\] (11)

To complete the specification of the Q-learning algorithm, we have to specify how the control action \( u \) is chosen in each state. Many choices have been suggested in the literature: these choices tradeoff the need to explore all state-action pairs with the need to use the “optimal” control suggested by the current estimates of the Q-values. In our analysis, we assume that each state is visited sufficiently often. Before we make our assumption precise, we first define the past history up to time \( k \) as
\[
\mathcal{I}_k = \{(x_j,u_j), j \leq k\}.
\]
Divide time into contiguous time slots called frames: the zeroth frame is time slot 0. The \( m \)th frame (for \( m \geq 1 \)) consists of time slots \((m-1)L+1,(m-1)L+2,\ldots,mL\), for an appropriately chosen \( L \) defined below. Let \( K_{i,u}^{(m)} \) denote the number of times that \((i,u)\) is visited in the \( m \)th frame, and define \( K_{\min}^{(m)} := \min_{i,u} K_{i,u}^{(m)} \).

Our persistence of excitation assumption can now be stated as follows.

**Assumption 1.** There exists a positive integer \( L \) such that
\[
E(K_{\min}^{(m)} | \mathcal{I}_{(m-1)L}) \geq 1
\]
for all \( m \geq 1 \) and all possible histories \( \mathcal{I}_{(m-1)L} \), where \( m = 1,2,\ldots \). \( \square \)

Assumption 1 states that, on average, each (state, action) pair is visited at least once in every frame. This condition is stronger than the condition in [14, 13] where it is only assumed that infinitely many updates to each (state, action) pair.

Now we are ready to analyze the convergence of asynchronous Q-learning. As before, we decompose the system into two parts: the first is driven by the noise sequence \( w \) and is given by
\[
z_{k+1}(i,u) = (1 - \varepsilon_k(i,u))z_k(i,u) + \alpha \varepsilon_k(i,u)w_k(i,u,\hat{Q}_k,x_{k+1}); \quad z_0(i,u) = 0. \] (12)
Then, we study the error term centered around \( z_k(i,u) \):
\[
D_k(i,u) := \hat{Q}_k(i,u) - z_k(i,u), \] (13)
where we recall that the error $\tilde{Q}_k(i,u)$ is defined to be $\tilde{Q}_k(i,u) = \hat{Q}_k(i,u) - Q(i,u)$. The dynamics of $D$ are given by

$$D_{k+1}(i,u) = (1 - \varepsilon_k(i,u)) D_k(i,u) + \varepsilon_k(i,u) \left( (T \hat{Q}_k)(i,u) - (T Q)(i,u) \right). \quad (14)$$

Before we study the dynamics of the $z_k$ and $D_k$, we first present a number of bounds which will be useful later.

### 3.1. Useful Bounds

As in the synchronous case, the $Q$-values also remain bounded (as a function of the initial estimate) in the asynchronous case [6].

**Fact 3.** There exists a finite $\hat{Q}_{max}$ such that

$$\| \hat{Q}_k \|_\infty \leq \hat{Q}_{max}, \quad \forall k.$$  

The following fact is an immediate consequence.

**Fact 4.** The following bounds hold:

1. $\| \hat{Q}_k \|_\infty \leq \hat{Q}_{max} := \hat{Q}_{max} + \| Q \|_\infty$.
2. $\| w_k \|_\infty \leq W_{max} := 2 \hat{Q}_{max}$, for all $k$.
3. $\| z_k \|_\infty \leq W_{max}$, for all $k$.
4. $\| D_k \|_\infty \leq D_{max} := \hat{Q}_{max} + \| Q \|_\infty + W_{max}$, for all $k$.
5. Define $\Delta \hat{Q}_k = \hat{Q}_{k+1} - \hat{Q}_k$. Then, for all $k$, $\| \Delta \hat{Q} \|_\infty \leq \varepsilon \Delta \hat{Q}_{max}$, where $\Delta \hat{Q}_{max} = \hat{Q}_{max} + W_{max} + r_{max} + \alpha \| Q \|_\infty$.
6. Define $\Delta D_k = D_{k+1} - D_k$. Then, for all $k$, $\| \Delta D_k \|_\infty \leq \varepsilon \Delta D_{max}$, where $\Delta D_{max} = D_{max} + \alpha \hat{Q}_{max}$.
7. Define $\Delta z_k = z_{k+1} - z_k$. Then, for all $k$, $\| \Delta z_k \|_\infty \leq \varepsilon \Delta z_{max}$, where $\Delta z_{max} = (1 + \alpha)W_{max}$.

**Proof.** The first two bounds follow from the definitions of $\hat{Q}_k$ and $w_k$. To show that bound (3) holds, note that since

$$z_{k+1}(i,u) = (1 - \varepsilon_k(i,u)) z_k(i,u) + \varepsilon_k(i,u) w_k(i,u, \hat{Q}_k, x_{k+1}),$$

we have

$$z_{s_l+1}(i,u) = (1 - \varepsilon) z_{s_l}(i,u) + \varepsilon w_{s_l},$$

where $s_l$ is the time slot during which $(i,u)$ is visited for the $l$th time. Thus $z_{s_l}(i,u) \leq W_{max}$, and therefore, $z_k(i,u) \leq W_{max}, \forall k$. Bound (4) follows from the definition of $D_k$ and bound (3).
Next, we consider bound (5). Note that
\[ \Delta \tilde{Q}_k(i,u) = \hat{Q}_{k+1}(i,u) - \hat{Q}_k(i,u) \]
\[ = -\varepsilon_k(i,u) \hat{Q}_k(i,u) + \varepsilon_k(i,u) [(T \hat{Q}_k)(i,u) + w_k] \]
Therefore,
\[ |\Delta \tilde{Q}_k(i,u)| \leq \varepsilon \hat{Q}_{\text{max}} + \varepsilon W_{\text{max}} + \varepsilon |(T \hat{Q}_k)(i,u)|. \]
Recall that
\[ (T \hat{Q}_k)(i,u) = r(i,u) + \alpha \sum_j p_{ij} \left( \max_{v \in \text{of}} \hat{Q}(j,v) \right), \]
which implies
\[ |(T \hat{Q}_k)(i,u)| \leq r_{\text{max}} + \alpha \| \hat{Q} \|_{\infty}, \]
yielding the desired result.

Finally,
\[ D_{k+1}(i,u) - D_k(i,u) = -\varepsilon D_k(i,u) + \varepsilon [(T \hat{Q})(i,u) - (T Q)(i,u)] \]
\[ \leq \varepsilon D_{\text{max}} + \varepsilon \| T \hat{Q} - T Q \|_{\infty} \]
\[ \leq \varepsilon D_{\text{max}} + \varepsilon \alpha \| \hat{Q} \| \]
\[ \leq \varepsilon (D_{\text{max}} + \alpha \hat{Q}_{\text{max}}). \]
This proves bound (6). The last bound follows in a similar fashion.

3.2. A Noise-Driven System

We are now ready to study the dynamics of (12). Let \( \tau_l(i,u) \) be the time instant of the \( l \)th visit (after time 0) to \( (i,u) \), with \( \tau_0(i,u) = 0 \). Then (12) can be rewritten as
\[ z_{\tau_l+1}(i,u) = (1 - \varepsilon) z_{\tau_l}(i,u) + \alpha \varepsilon w_{\tau_l}(i,u, \hat{Q}_k, x_{k+1}); \quad z_{\tau_0}(i,u) = 0, \]
where we have dropped the arguments for \( \tau \) for notational convenience. Further simplifying our notation, we define \( \hat{z}_l = z_{\tau_l(i,u)}(i,u) \) and \( \hat{w}_l = w_{\tau_l(i,u)}(i,u, \hat{Q}_k, x_{k+1}) \), giving us
\[ \hat{z}_{l+1} = (1 - \varepsilon) \hat{z}_l + \alpha \varepsilon \hat{w}_l; \quad \hat{z}_0 = 0. \quad (15) \]

Squaring both sides of the expression in (15) and taking expectations gives us
\[ E(\hat{z}_{l+1}^2) = (1 - \varepsilon)^2 E(\hat{z}_l^2) + \alpha^2 \varepsilon^2 E(\hat{w}_l^2) + 2\alpha \varepsilon (1 - \varepsilon) E(\hat{z}_l \hat{w}_l). \]

We note the following:
1. Since $\|w_k\|_\infty \leq W_{\text{max}}$, we have $\hat{w}_l^2 \leq W_{\text{max}}^2$.
2. Further,

$$E(\hat{z}_l \hat{w}_l | \tau_l(i,u) = k) = E\left[E(\hat{z}_l \hat{w}_l | \tau_l(i,u) = k, \{Q_m\}_{m=0}^{k}, K_{k-1})| \tau_l(i,u) = k\right]$$

$$= E\left[z_k(i,u)E(w_k(i,u,\hat{Q}_l, x_k) | \tau_l(i,u) = k, \{Q_m\}_{m=0}^{k}, K_{k-1})| \tau_l(i,u) = k\right]$$

$$= 0.$$

Thus,

$$E(\hat{z}_l \hat{w}_l) = \sum_{k=1}^{\infty} E(\hat{z}_l \hat{w}_l | \tau_l(i,u) = k) P(\tau_l(i,u) = k) = 0.$$

Therefore, we have

$$E(\hat{z}_{l+1}^2) \leq E(\hat{z}_l^2)(1 - \epsilon)^2 + \epsilon^2 W_{\text{max}}^2, \quad E(\hat{z}_0^2) = 0$$

which implies

$$E(\hat{z}_l^2) \leq \epsilon^2 W_{\text{max}}^2 \sum_{j=0}^{l-1} (1 - \epsilon)^2 j$$

$$\leq \epsilon^2 W_{\text{max}}^2 \sum_{j=0}^{\infty} (1 - \epsilon)^2 j$$

$$= \frac{\epsilon^2}{1 - (1 - \epsilon)^2} W_{\text{max}}^2$$

$$= \frac{\epsilon}{2 - \epsilon} W_{\text{max}}^2.$$

We summarize the above result in the following lemma.

**Lemma 3.1.**

$$E(z_k^2(i,u)) \leq \frac{\epsilon}{2 - \epsilon} W_{\text{max}}^2, \text{ for all } k \geq 0. \quad (16)$$

**Proof.** Since $z_k(i,u)$ changes only when state $(i,u)$ is visited, the upper bound on $E(z_0^2(i,u))$ will also hold for $E(z_k^2(i,u))$ for all $k$. \qed

The following bound for $E(\|z_k\|_\infty)$ is an immediate consequence of Lemma 3.1.
Lemma 3.2.

\[ E(\|z_k\|_\infty) \leq \sqrt{|F \times A|} \frac{\epsilon}{2-\epsilon} W_{\text{max}}^2, \text{ for all } k \geq 0. \]  

(17)

**Proof.** By definition,

\[ E(\|z_k\|_\infty) = E\left( \max_{i,u} |z_k(i,u)| \right). \]

Clearly,

\[ E\left( \max_{i,u} |z_k(i,u)| \right) = E\left( \sqrt{\max_{i,u} |z_k(i,u)|^2} \right) \leq \sqrt{E\left( \sum_{i,u} |z_k(i,u)|^2 \right)} \leq \sqrt{|F \times A|} \frac{\epsilon}{2-\epsilon} W_{\text{max}}^2. \]

3.3. Error Dynamics

We now study the dynamics of the error centered around \( z \) given in (14). From (14), it follows that

\[ |D_{k+1}(i,u)| \leq (1 - \epsilon_k(i,u)) |D_k(i,u)| + \epsilon_k(i,u) (T \hat{Q}_k)(i,u) - (TQ)(i,u) | \leq (1 - \epsilon_k(i,u)) |D_k(i,u)| + \epsilon_k(i,u) \|T \hat{Q}_k - (TQ)\|_\infty \]

and therefore

\[ |D_{k+1}| \leq (1 - \epsilon_k(i,u)) |D_k(i,u)| + \alpha \epsilon_k(i,u) \|\hat{Q}_k - Q\|_\infty, \forall k \geq 0, \]  

(18)

with

\[ |D_0(i,u)| \leq \|\hat{Q}_0\|_\infty + \|Q\|_\infty. \]

Consider the \((m+1)^{th}\) frame, and let \( t_l(i,u) \) be the time of the \( l^{th} \) visit during this frame to \((i,u)\). Then, from (18), we have

\[ |D_{t_l+1(i,u)}(i,u)| \leq (1 - \epsilon) |D_{t_l(i,u)}(i,u)| + \alpha \epsilon \|\hat{Q}_{t_l}\|_\infty, \]  

(19)

with \( D_{t_0(i,u)}(i,u) = D_{mL}(i,u) \) by definition.
Theorem 3.3. If $\varepsilon < \frac{1}{L}$ and Assumption 1 holds, then

$$ E \left( \| D_{mL} \|_\infty \right) \leq (1 - \varepsilon(1 - \alpha))^m \tilde{Q}_{\text{max}} + \tilde{D}_{\text{max}}(L, \varepsilon), $$

(20)

where

$$ \tilde{D}_{\text{max}} := \frac{\alpha L}{(1 - \alpha)} \sqrt{\frac{\varepsilon}{2 - \varepsilon}} \frac{W^2_{\text{max}} |\mathcal{S} \times \mathcal{A}|}{2} + \frac{L(L - 1)}{2(1 - \alpha)} \varepsilon \left( \Delta_{D_{\text{max}}} + \alpha \Delta_{\tilde{Q}_{\text{max}}} \right). $$

Proof. See the appendix.

The right-hand side in (20) can be made arbitrarily small by choosing $\varepsilon$ small, and then choosing $m$ large. Further, for any $\theta > 0$, if the right-hand side of (20) is less than or equal to $\theta$ for some $m > 1$, then $\theta$ continues to be an upper bound to the right-hand side for all $n \geq m$. Further,

$$ E \left( \| D_k \|_\infty \right) \leq E \left( \| D_{mL} \|_\infty \right) + L \varepsilon \Delta_{D_{\text{max}}}, \forall k = mL + 1, \ldots, (m + 1)L. $$

Recalling that

$$ \| \tilde{Q}_k \|_\infty \leq \| D_k \|_\infty + \| z_k \|_\infty, $$

(21)

we have

$$ E(\| \tilde{Q}_k \|_\infty) \leq E(\| D_k \|_\infty) + \sqrt{\frac{\varepsilon}{2 - \varepsilon} \frac{W^2_{\text{max}} |\mathcal{S} \times \mathcal{A}|}{2}}. $$

Thus, we have the following convergence result, in addition to the above bound on the first moment of the error.

Theorem 3.4. Under Assumption 1, given $\theta > 0$, there exists an $\varepsilon_0$ such that $\forall \varepsilon \leq \varepsilon_0$, there is a sufficiently large $K_\varepsilon > 0$ for which the following statement holds:

$$ E(\| \tilde{Q}_k \|_\infty) \leq \theta, \quad \forall k \geq K_\varepsilon. $$

4. Conclusions

We have provided error bounds for synchronous and asynchronous $Q$-learning applied to controlled finite-state, discrete-time Markov chains. By considering the finite-state case, the problem of bounding the $Q$-value estimates becomes quite simple. However, even in this case, explicit error bounds were not established previously. Our contribution here is to derive such error bounds using the following assumption regarding the choice of the control action sequence: there exists
an $L$ such that every state is visited at least once on average in every interval of length $L$.

It is important to note that the maximum value for $\varepsilon$ is inversely proportional to the size of the state-action space. This may require $\varepsilon$ to be very small and thus, the progress of the $Q$-learning algorithm towards the vicinity of the correct $Q$-values may be slow. Therefore, one may want to first aggregate the state-action pairs into a small set before using the algorithm (see [4] for a discussion). Additionally, one may start with large values of the step-size $\varepsilon$ and slowly let it decrease. An appropriately modified version of our results will still apply to this model if the sequence of step-sizes is lower bounded. A lower-bounded sequence of step-sizes can be useful in practice because, if the step-size decreases to zero, then the algorithm cannot track the $Q$-values when the system parameters vary slowly with time.

5. Appendix: Proof of Theorem 3.3

Let $K_{i,u}^{(m+1)}$ denote the number of times $(i,u)$ is visited in the $(m + 1)^{th}$ frame. Then, from inequality (19), we have

$$|D_{k_{i,u}^{(m+1)}}(i,u) - D_{0}(i,u)| \leq -\varepsilon \sum_{l=0}^{k_{i,u}^{(m+1)}-1} |D_{l}(i,u)| + \alpha \varepsilon \sum_{l=0}^{k_{i,u}^{(m+1)}-1} \|\tilde{Q}_l\|_\infty$$

$$\leq -\varepsilon \sum_{l=0}^{k_{i,u}^{(m+1)}-1} (|D_{mL}(i,u)| - l\varepsilon \Delta_{D_{max}})$$

$$+ \alpha \varepsilon \sum_{l=0}^{k_{i,u}^{(m+1)}-1} (\|\tilde{Q}_{mL}\|_\infty + l\varepsilon \Delta_{\tilde{Q}_{max}})$$

For convenience, we will use $\kappa$ to denote $K_{i,u}^{(m+1)}$. Thus,

$$|D_{k_{i,u}^{(m+1)}}(i,u)| \leq (1 - \kappa \varepsilon)|D_{mL}(i,u)| + \alpha \varepsilon \kappa \|\tilde{Q}_{mL}\|_\infty + \varepsilon^2 \Delta_{D_{max}} \kappa (\kappa - 1)$$

$$+ \alpha \varepsilon^2 \Delta_{\tilde{Q}_{max}} \frac{\kappa (\kappa - 1)}{2}.$$
frame after time instant $t_\kappa$. Thus,

$$|D_{(m+1)L}(i,u)| \leq (1 - \kappa \varepsilon)\|D_{mL}\|_\infty + \alpha \varepsilon \kappa \|\tilde{Q}_{mL}\|_\infty + \varepsilon^2 \Delta_{D_{\max}} \frac{\kappa(\kappa - 1)}{2} + \alpha \varepsilon^2 \Delta_{\tilde{Q}_{\max}} \frac{\kappa(\kappa - 1)}{2}.$$  

Using the fact $\tilde{Q}_{mL} = D_{mL} + z_{mL}$, we have

$$|D_{(m+1)L}(i,u)| \leq (1 - \kappa \varepsilon(1 - \alpha))\|D_{mL}\|_\infty + \alpha \varepsilon \kappa \|z_{mL}\|_\infty + \varepsilon^2 \Delta_{D_{\max}} \frac{\kappa(\kappa - 1)}{2} + \alpha \varepsilon^2 \Delta_{\tilde{Q}_{\max}} \frac{\kappa(\kappa - 1)}{2}.$$  

Next, using the fact $\kappa < L$ and Assumption 1, we have

$$E\left(\|D_{(m+1)L}\|_\infty\right) \leq (1 - \varepsilon(1 - \alpha))E\left(\|D_{mL}\|_\infty\right) + \alpha \varepsilon L E\left(\|z_{mL}\|_\infty\right) + \varepsilon^2 \frac{L(L - 1)}{2} (\Delta_{D_{\max}} + \alpha \Delta_{\tilde{Q}_{\max}}).$$  

Using the bound on $\|z\|_\infty$ gives the desired result.

**Acknowledgments**

Research supported by NSF Grant CMMI 1100257 and AFOSR MURI Grant FA 955010-10573.

**References**


